Métodos estatísticos aplicados em saúde pública

Wagner Hugo Bonat*

23 de outubro de 2007

1 Introdução

A degradação do meio ambiente e os problemas sócio-culturais no Brasil afetam o cenário epidemiológico urbano brasileiro, levando-o a ser destaque na mídia nacional e internacional, decorrente de epidemias de dengue, leptospirose, a recorrência de tuberculose, entre outras.

Diante dessa realidade constatou-se que é de fundamental importância criar métodos capazes de detectar precocemente o número de casos que caracterizam surtos epidêmicos, modelar e identificar fatores de risco e de proteção nas situações endêmicas e epidêmicas.

Nesta perspectiva foi elaborado o "Projeto SAUDAVEL"(Sistema de Apoio Unificado para Detecção e Acompanhamento em Vigilância) , o qual pretende contribuir para aumentar a capacidade do setor de saúde no controle de doenças transmissíveis, demonstrando ser necessário desenvolver novos instrumentos para a prática da vigilância epidemiológica, incorporando aspectos ambientais, identificadores de risco e métodos automáticos e semi-automáticos, que permitam a detecção de surtos e seu acompanhamento no espaço e no tempo.

2 Objetivos

O objetivo principal deste trabalho é desenvolver um procedimento para a visualização espaço-temporal dos dados do experimento de coleta de ovos do mosquito *Aedes aegypti* que está sendo desenvolvido pelo projeto SAUDAVEL na cidade de Recife-PE. Para isto, optou-se por uma resolução de

*Graduando em estatística - UFPR

análise por bairros, visando interpolar uma superfície com base nas amostras coletadas. Para atender a este propósito buscou-se na literatura, métodos para a interpolação de superfícies, dois destes métodos são descritos a saber Regressão Local e Superfície de Tendência.

Além disto, também buscou-se demonstrar a utilização de softwares desenvolvidos pelos grupos integrantes do projeto SAUDAVEL, dando enfâse ao **aRT-API R-TerraLib**[1][8] desenvolvido pelo LEG - Laboratório de Estatística e Geoinformação da Universidade Federal do Paraná.

3 Procedimentos Metodológicos

Nesta seção serão descritos o experimento, métodos de coleta de dados, e a metodologia estatística utilizada.

3.1 Área de estudo, Instrumentos e Técnicas de Campo

O experimento está sendo desenvolvido na cidade de Recife/PE onde foram criteriosamente instaladas 564 armadilhas para o mosquito *Aedes aegypti* cuja a femêa é o principal vetor da Dengue. Estas armadilhas foram monitoradas durante o período de 03/2004 a 12/2006, cerca de um quarto das armadilhas são monitoradas a cada 7 dias, assim em um ciclo de 28 dias todas as armadilhas são monitoradas, o experimento foi realizado em 5 dos 94 bairros da cidade de Recife.

A rede de armadilhas foi instalada de modo a cobrir toda a extensão do bairro, caracterizando bem o tipo de delineamento utilizado para a coleta de dados, durante o período do experimento foram realizadas 17.668 coletas, com as quais foram contados ao todo 13.628.909 ovos do mosquito *Aedes aegypti*.

De forma geral, os dados estão disponíveis na forma de amostras pontuais do processo e para utilizá-las de forma efetiva necessita-se de um procedimento de interpolação, para gerar uma superfície que represente o fenômeno em toda a área. Na literatura de estatística espacial este tipo de procedimento é comumente chamado de análise de superfície.

Uma parte importante em projetos desta magnitude, é a forma de armazenamento, tratamento e visualização dos dados, já que um experimento como este gera uma grande quantidade de dados que não são facilmente manipuláveis, requerendo-se para isto ferramentas específicas.

Neste experimento cada armadilha contém uma lâmina na qual a fêmea do mosquito coloca os ovos, essas lâminas são recolhidas e a contagem dos ovos

[1] http://www.leg.ufpr.br/aRT/

é feita em laboratório especializado. Os dados são inseridos em um banco de dados através de uma interface *Web* que foi desenvolvida pelo INPE-Instituto Nacional de Pesquisas Espaciais e CPqAM (Centro de pesquisas Argeu Magalhães), esta interface foi projetada para evitar formas complexas de entrada de dados e fornecer algumas medidas espaço-temporais rápidas.

Os serviços de saúde locais e o laboratório de Entomologia são os coordenadores operacionais e logísticos, e responsáveis pela realização do experimento [13].

O banco de dados do Recife SAUDAVEL, está implementado em **TerraLib** tecnologia de código aberto *light-DBMS*, e *MySql* Database Server[2], como um repositório e sistema gerenciador de dados espaço-temporais, baseado no modelo espaço-temporal da **TerraLib** [10]. O banco de dados original do SAUDAVEL Recife fica no INPE e um backup é feito pelo LEG-Laboratório de Estatística e Geoinformação da UFPR.

O LEG tem particular importância para o experimento de Recife, já que ele é o responsável por coordenar e implementar os modelos estatísticos. O LEG também é responsável por desenvolver tecnologias de integração entre a biblioteca **TerraLib** e um ambiente de computação e modelagem estatística, o **projeto R** [12]. Neste sentido vem sendo desenvolvido o pacote **aRT-API R-TerraLib** [8]

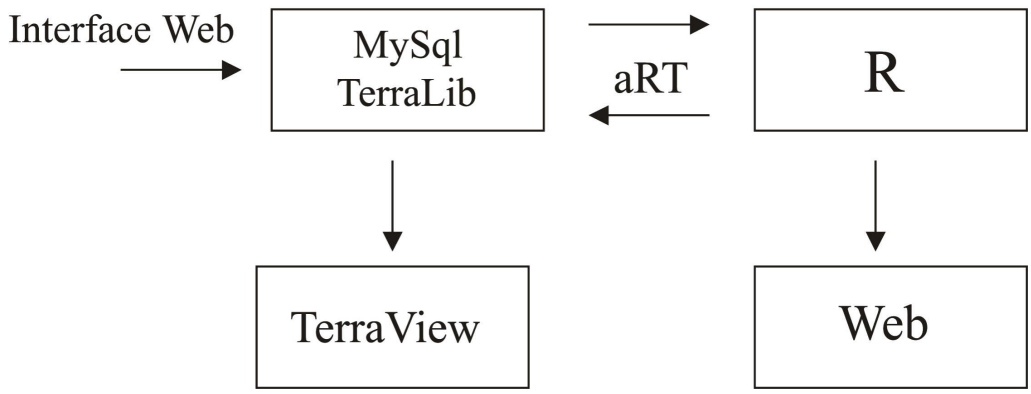

Figura 1: Formato geral de análise

A figura 1 ilustra o formato geral da análise. Após a equipe de campo coletar as lâminas e estas serem analisadas pelo laboratório especializado, os dados entram no banco através da interface *Web* e ficam armanenados no banco de dados geográficos, construído sobre a plataforma **TerraLib** com o *MySql* Server Database. Isto feito, pode-se acessar o banco via **aRT**, assim

[2]www.mysql.com

todas as análises estatísticas são realizadas em um ambiente próprio no caso o **R**. Após as análises serem concluídas tem-se através do **aRT**, a opção de retornar os dados para o banco geográfico, podendo ser acessado por um visualizador de SIG(Sistema de Informação Geográfica) como é o caso do **TerraView**, ou então gerar uma página *Web* para a visualização pública dos resultados.

De maneira ampla pode-se resumir o procedimento para a visualização espaço-temporal do fenômeno após a coleta e entrada dos dados no banco, pelos seguintes passos:

1. Ler o banco geográfico através do **aRT**.
2. Estimar uma superfície para cada semana do experimento através dos métodos descritos na seção (3.2).
3. Interpolar 7 imagens entre cada superfície estimada.
4. Exportar as imagens geradas em formato **jpeg**[3], para um diretório do Linux.
5. Gerar um arquivo do tipo **.avi**[4] para a visualização das imagens através do software *mencoder* contido no *Mplayer*[5] para Linux.
6. Disponibilizar o arquivo **.avi** em uma página na *Web* para visualização.

3.2 Metodologia Estatística

Para gerar uma superfície que aproxime o fenômeno em estudo de forma realista, é necessário modelar sua estrutura de covariância espacial. De uma forma ampla pode-se considerar três tipos de abordagem para este propósito, conforme [11].

1. Modelos determinísticos de efeitos locais; cada ponto da superfície é estimado com base apenas na interpolação dos valores das amostras mais próximas. A suposição implícita é que predominam apenas os efeitos puramente locais.
2. Modelos determinísticos de efeitos globais; a suposição implícita, nesta classe de interpoladores, é de que, para a caracterização do fenômeno, predomina a variação em larga escala, este é o caso dos interpoladores por superfície de tendência.

[3]Joint Photographic Experts Group
[4]Audio Video Interleave
[5]www.mplayerhq.hu

3. Modelos estatísticos de efeitos globais e locais (Krigagem); cada ponto da superfície é estimado apenas tendo como fundamento a interpolação dos valores das amostras mais próximas, utilizando um estimador estatístico.

A primeira será explicada aqui levando em consideração uma forma mais ampla de considerar os efeitos espaciais de forma local, representada pela Regressão Polinomial Local considerada neste trabalho mais eficiente que os estimadores de *kernel* tradicionais, a segunda será considerada pelos interpoladores de superfície de tendência e a terceira não será considerada neste trabalho, para maiores informações ver [9].

3.2.1 Regressão Local

Regressão Local *(Loess)* é um método não paramétrico que estima curvas e superfícies através de suavização *(smoothing)*. Este método ganhou popularidade a partir da década de 70 com o desenvolvimento de computadores e a publicação dos estudos independentes de [14], [15] e [7]. Sendo que [15] desenvolveu o software Lowess, que foi implementado em diversos pacotes estatísticos. As idéias básicas do método podem ser observadas ao considerar-se o mais simples dos modelos de regressão, onde a variável dependente, y, e a independente, x, são relacionadas por:

$$y_i = g(x_i) + \epsilon_i$$

onde ϵ_i denota o termo de erro independente e identicamente distribuído com distribuição normal, média zero e variância constante.

Ao contrário dos métodos paramétricos que estimam a função globalmente, regressão local estima a função "g" na vizinhança de cada ponto de interesse $x = x_0$. Uma forma simples de estimar uma função localmente é considerar a média ponderada das observações que estão na vizinhança do ponto de interesse, x_0. Duas escolhas devem ser feitas para realizar esta estimativa. Primeiro, deve ser escolhido o tamanho da vizinhança, h, do ponto $x = x_0$ e , segundo, deve ser escolhida uma função K que pondera o conjunto de pontos vizinhos a x_0. A função K é denominada de núcleo ($Kernel$), enquanto que h é denominada de banda ou parâmetro de suavização. Com este procedimento, a equação para a média local ponderada por K é dada por:

$$\widehat{g}(x_0) = \frac{\sum_{i=1}^{n} K_h(x_i - x_0) y_i}{\sum_{i=1}^{n} K_h(x_i - x_0)}$$

Este estimador de núcleo foi proposto inicialmente por [5] e [6]. Existem sérias limitações com a estimativa de uma constante localmente, como

por exemplo, viés nas regiões de fronteira e no interior se a variável independente não for uniforme e se a função de regressão tiver curvatura. Uma maneira de resolver este problema é através de regressão local linear ponderada, proposta inicialmente por [14] e [15]. Ao estimar uma linha reta localmente ao invés de uma constante, o problema de viés de primeira ordem é eliminado, desta forma, regressão local linear resolve um problema de mínimos quadrados ponderados a cada ponto de interesse, x_0, conforme:

$$\min_{\alpha\beta} \sum_{i=1}^{n} K_h(x_i - x_0)[y_i - \alpha - \beta(x_i - x_0)]^2 \tag{1}$$

Regressão local linear será igual ao estimador de Nadaraya-Watson expresso pela equação (1) se o termo $\beta(x_i - x_0)$ for removido. Neste caso, uma constante será estimada localmente. Apesar de regressão local linear ser utilizado como técnica padrão por muitos autores [1], não há razões para não utilizar polinômios de ordem mais alta, mesmo porque a regressão local pode apresentar viés quando a função a ser estimada possui uma forte curvatura. Nestes casos uma polinomial de grau d pode ser estimada através da seguinte função:

$$\min_{\alpha\beta_j, j=1,\ldots,d} \sum_{i=1}^{n} K_h(x_i - x_0)[y_i - \alpha - \sum_{i=1}^{d} \beta(x_j - x_0)^j]^2$$

Portanto para modelar-se determinado processo por regressão local, deve-se de forma geral fazer três escolhas: a função núcleo, o parâmetro de suavização e o grau da polinomial. Existe ainda uma outra escolha que deve ser feita, que diz respeito a distribuição assumida para os termos de erro, no presente trabalho assume-se que os erros seguem uma distribuição gaussiana. Para uma discussão de regressão local com a consideração de outras distribuições do erro ver [4].

Parâmetro de suavização (banda). O parâmetro de suavização (span ou bandwidth), h, controla o tamanho da vizinhança no entorno de x_0 no qual a função núcleo será aplicada. O parâmetro de suavização possui papel determinante na variabilidade e no viés da estimativa. Se o h escolhido for pequeno, a estimativa terá um viés reduzido, mas uma variabilidade elevada. Por outro lado, se o h escolhido for grande, a estimativa terá um viés elevado mas pequena variabilidade.O objetivo é produzir uma estimativa que seja a mais suave possível sem distorcer a relação de dependência entre as variáveis em análise [2]. [3] discute e compara diferentes procedimentos para a escolha da banda, os quais são classificados em dois grupos. O primeiro, constituído pelos métodos clássicos, são baseados em extensões dos procedimentos já

utilizados em regressão paramétrica, tais como validação cruzada (cross validation), critério de informação de Akaike e Cp de Mallows, que consistem basicamente em empregar alguma medida de aderência, como por exemplo, minimizar a média da integral do erro ao quadrado ou uma simplificação desta. Os métodos do segundo grupo são baseados em anexo (plug-ins). Estes consistem em escrever a função inicialmente estimada, $\widehat{g}$, como uma função g desconhecida e aproximada por uma expansão de Taylor ou outra expansão assintótica. Uma estimativa de g é então "anexada"(plugged-in) para derivar uma estimativa da tendenciosidade e uma estimativa da aderência, tal como, o erro quadrado médio integrado *(mean integrated squared error)*. Segundo [4] os métodos clássicos apresentam melhores resultados em termos práticos, bem como se ajustam a uma grande variedade de casos.

Grau do polinômio local. Esta escolha também afeta a relação entre variância e viés, quanto maior o grau da polinomial menor será o viés e maior a variância para um mesmo parâmetro de suavização. De modo geral , o aumento da variância que decorre da utilização de polinomiais de ordem mais elevada pode ser compensado empregando-se um parâmetro de suavização maior. A utilização de polinomiais de baixa ordem é suficiente para produzir estimativas de ótima qualidade, normalmente são utilizados polinomiais com graus variando de zero a três. A escolha do grau da polinomial é, em sua maior parte, determinada pelos objetivos do trabalho e pelos dados, na prática a escolha do grau da polinomial pode ser realizada pela inspeção visual do gráfico com os dados originais e a estimativa de regressão local. De forma geral, a presença de "picos"ou "vales"nos dados são um indicativo de que d deve ser igual a dois ou três, enquanto que a presença de um padrão único indicam que d deve ser igual a um.

A função de *kernel*. Esta função é responsável por ponderar as observações na vizinhança de cada ponto de interesse, x_0. Segundo [2] e [4] esta função deve ser contínua, simétrica, com maior peso em torno de x_0 e decrescente a medida em que x se afasta de x_0. Dentre as escolhas possíveis, destaca-se aqui a função gaussiana que será utilizada no trabalho.

Para analisar esta função vamos considerar uma variável transformada u_i definida por:

$$u_i = \frac{(x_i - x_0)}{h_i}$$

Então a função de peso K é obtida em função da variável u, isto é,

$$K(u) = K[\frac{(x_i - x_0)}{h_i}]$$

A função gaussiana é centrada em x_0, a banda h é o desvio padrão da amostra, assim valores que estão situados a mais de dois desvios (2h) receberão um peso negligenciável, pois a área da curva normal além dos dois desvios é muito pequena. A expressão para a função de peso gaussiana é apresentada a seguir:

$$K(u) = \frac{1}{\sqrt{2\pi}} \exp^{u^2/2}, \quad \text{se} \quad -\infty < u < +\infty$$

Muitas outras funções de *Kernel* ou pesos, podem ser utilizadas, ela vai depender dos objetivos do trabalho.

3.2.2 Superfícies de Tendência

As superfícies de tendência são interpoladores globais, a superfície é estimada por um ajuste polinomial aos dados, por um processo de regressão múltipla entre os valores do atributo e as coordenadas geográficas das observações, estes interpoladores buscam modelar a variação espacial em larga escala. A saída é uma função polinomial na qual o valor da variável é expresso em função das coordenadas da superfície, expressas em duas ou três dimensões. Exemplos incluem equações lineares do tipo:

$$z = \alpha_1 + \alpha_2 x + \alpha_3 y$$

e equações quadráticas como:

$$w = \alpha_1 + \alpha_2 x + \alpha_3 y + \alpha_4 x^2 + \alpha_5 y^2$$

A suposição implícita nos interpoladores por superfície de tendência é de que, para a caracterização do fenômeno em estudo, predomina a variação em larga escala e que a variabilidade local não é relevante, neste caso o processo é não estacionário.

4 Resultados

O primeiro método utilizado para gerar a superfície foi a Regressão Local, como descrito na metodologia para a aplicação deste método necessita-se fazer três escolhas: o parâmetro de suavização *(span ou bandwidth*), grau do polinômio local e a função de suavização ou *Kernel*.

O *span* foi definido como 0.75 como é *default* em vários pacotes estatísticos, para o grau do polinômio local foi definido como 2 e a função de suavização utilizada foi a Gaussiana.

As superfícies interpoladas foram suaves e aproximam o fenômeno em estudo, um cuidado especial também foi tomado decorrente dos dados serem provenientes de uma contagem, esses foram reexpresso pelo seu logaritmo, sendo que onde as contagens eram iguais a zero foi adicionado uma unidade.

As superfícies foram geradas semanalmente, para cada um dos 5 bairros que compõe o experimento em Recife, gerando de 128 a 136 imagens dependendo do número de observações feitas em cada bairro, essas imagens foram animadas utilizando uma interpolação pixel a pixel, sendo que entre duas superfície estimadas foram interpoladas 7 uma para cada dia da semana.

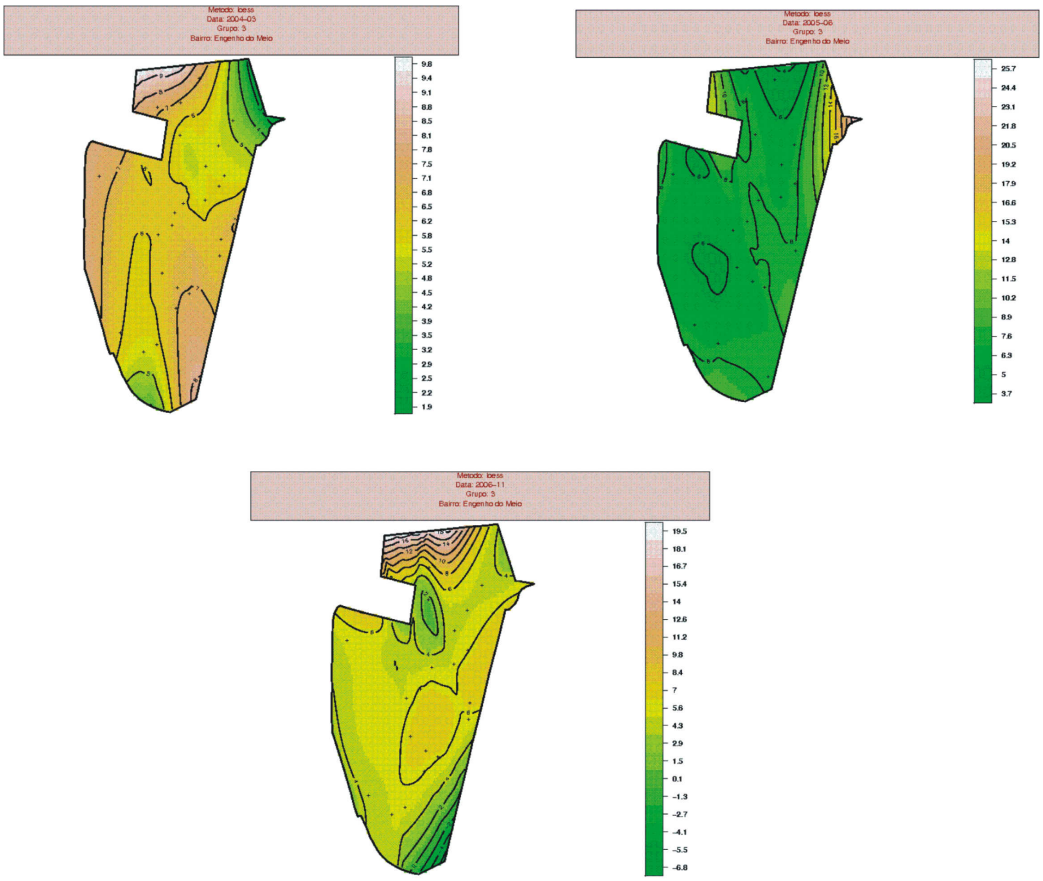

Figura 2: Superfícies estimadas pelo método Loess

O segundo método utilizado busca mostrar efeitos globais de tendências

no espaço, a superfície foi interpolada utilizando uma regressão polinomial de segunda ordem, onde a contagem de ovos do mosquito foi a variável resposta e as coordenadas de cada armadilha as regressoras do modelo.

As superfícies para este método mostram-se menos suaves e variações abruptas ocorrem em determinados pontos no tempo, somente um procedimento visual é capaz de verificar tais variações.

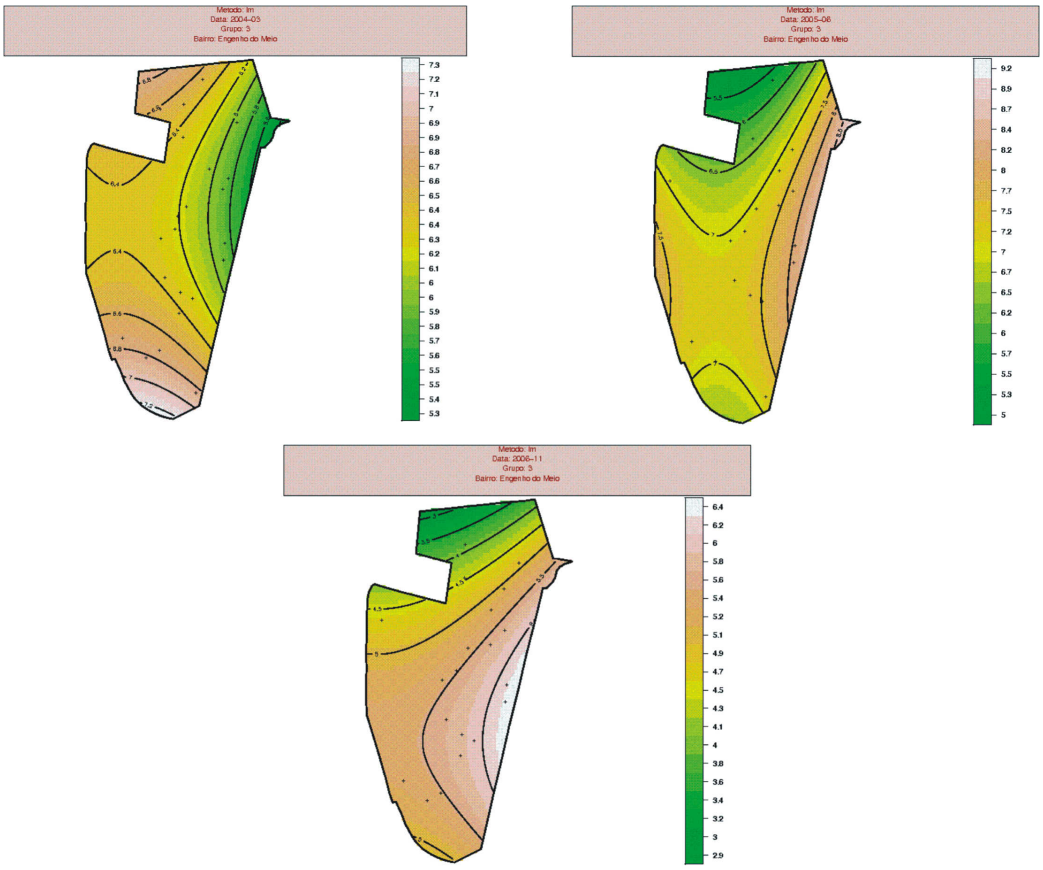

Figura 3: Superfícies estimadas pelo método superfície de tendência

Aqui foram apresentadas apenas três imagens, a inicial, intermediária e a final, para o bairro Engenho do Meio.

http://www.leg.ufpr.br/doku.php/projetos:saudavel são apresentadas as animações contendo todas as imagens para cada um dos cinco bairros da cidade de Recife onde foi realizado o experimento. Além disso também são apresentadas todas as funções que foram desenvolvidas para efetuar as análises.

5 Conclusão

A visualização de dados espaço-temporal é algo recente na estatística e portanto, pouco se tem de ferramentas exploratória para dados desta natureza. O intuito de uma análise exploratória é dar subsídios para o estatístico buscar modelos que se adequem aos dados, dando a ele a visualização de como o fenômeno se desenvolveu tanto no espaço como no tempo.

Para o caso do experimento de contagens de ovos do mosquito *Aedes aegypti* na cidade de Recife/PE, o procedimento mostrou-se satisfatório, gerando uma visualização espaço-temporal do experimento, evidenciando picos e tendências espaciais e mostrando a evolução do fenômenos em toda a área, dando a possibilidade de ver os dados, coisa que não é nada fácil em um experimento desta magnitude.

Com este trabalho também é possível verificar as potencialidades da integração entre os Sistemas de Informações Geográficas (SIG's) e ambientes estatísticos como **R**, onde através do **aRT** o ambiente estatístico ganha as potencialidades do SIG agregada ao potencial de análise estatística próprios destes ambientes, não antes disponível em um único ambiente integrado.

Além disso esta análise preliminar busca dar subsídios para um modelo espaço-temporal, o qual pretende contribuir para aumentar a capacidade do setor de saúde no controle de doenças transmissíveis como a Dengue.

Uma limitação do procedimento proposto é a escala dos dados já que dados em diferentes pontos no tempo apresentaram valores bastante diferentes, sendo dificil a geração de uma única animação com escalas variadas do processo, além disso, a presença de pontos discrepantes tendem a deixar a escala de visualização distorcida escondendo variações em pequenas escalas.

Deixa-se aqui como futuras agendas de pesquisa para o projeto SAUDAVEL Recife, incorporar nas superfícies estimadas outras possíveis covariáveis como condições climáticas, e defasagens temporais da variável resposta. Além de métodos mais flexíveis que não assumam formas lineares como os estimadores de superfície de tendência e não sejam dependentes de um raio (*span*) de influência como a Regressão Local.

Referências

[1] BOWMAN ADRIAN ; AZZALINI ADELCHI. Applied smoothing techniques for data analysis: The kernel approach with s-plus illustrations. *Oxford: Oxford University Press*, 1997.

[2] CLEVELAND W. S. ; LOADER C. Smoothing by local regression: Principles and methods. *Physica-Verlag, p. 10-49*, 1996a.

[3] LOADER CLIVE. Old faithful erupts: Bandwidth selection reviewed. *Working paper, ATT Bell Laboratory*, 1995.

[4] LOADER CLIVE. Local regression and likelihood. *Springer-Verlag*, 1999.

[5] NADARAYA E.A. On estimating regression. *Theory of Probability and ilts Applications, v.9, p. 141-142*, 1964.

[6] WATSON G.S. Smooth regression analysis. *Sankhya, Series. A, v.26, p. 359-372*, 1964.

[7] STONE C. J. Optimal rates of convergence for nonparametric estimators. *The annals of Statistics, v.8 p.1348-1360*, 1980.

[8] RIBEIRO PEDRO DE ANDRADE NETO ; MARCOS AURÉLIO CARRERO ; THIAGO EUGÊNIO BEZERRA DE MELLO ; PAULO JUSTINIANO RIBEIRO JUNIOR. Integration of geographic information systems and statistical computing: the terralib/r case. *V Simpósio Brasileiro de Geoinformática*, 2004.

[9] DIGLE PETER J.; RIBEIRO PAULO JUSTINIANO. *Model-based Geostatistics*. Springer Verlag, Brasília, 2007.

[10] SILVEIRA JOSÉ CONSTANTINO ; WAYNER VIEIRA DE SOUZA ; LEDA NARCISA RÉGIS ; MARIA ALICE V. DE M. SANTOS ; TIAGO MARIA LAPA; JOSÉ LUIZ PORTUGAL ; THANISSE SILVA BRAGA ; ANTONIO MIGUEL VIEIRA MONTEIRO. Recife em "pedaços": Geotecnologias para a detecção e acompanhamento em vigilância epidemiológica. *VI Congresso Brasileiro de Epidemiologia*, 2004.

[11] SUZANA DRUCK ; MARILIA SÁ CARVALHO ; GILBERTO CÂMARA ; ANTÔNIO MIGUEL VIEIRA MONTEIRO. *Análise Espacial de Dados Geográficos*. Embrapa, Brasília,DF, 2004.

[12] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2007. ISBN 3-900051-07-0.

[13] MONTEIRO ANTONIO M. VIEIRA ; MARILIA SÁ CARVALHO ; RENATO ASSUNÇÃO ; WAYNER VIEIRA ; PAULO JUSTINIANO RIBEIRO; CLODOVEU DAVIS Jr ; LEDA REGIS. Saudavel: Bridging the gap between research and services in public health operational programs by multi - institutional networking development and use of spatial information technology innovative tools. 2006.

[14] C. J. STONE. Consistent nonparametric regression, with discussion. *The annals of Statistics, 5 p.549-645*, 1977.

[15] CLEVELAND WILLIAM. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association, v.74, p. 829-836*, 1979.