Henrique Silva Dallazuanna

Análise estatística da dispersão espaço temporal da Dengue na cidade de Recife - PE

Curitiba
2007

Henrique Silva Dallazuanna

Análise estatística da dispersão espaço-temporal da Dengue na cidade de Recife - PE

Relatório de análise estatística, apresentado à disciplina de Laboratório I, do curso de Bacharelado em Estatística, do Setor de Ciências Exatas da Universidade Federal do Paraná.

Orientador: Prof. Ricardo Sander Ehlers

Curitiba
2007

Sumário

Análise estatística da disperção espaço-temporal da Dengue na cidade de Recife - PE

Resumo

Diante dos recentes avanços de epidemias nas grandes cidades brasileiras, dentre as quais destacamos a Dengue, devido à proliferação de mosquitos da espécie *Aedes aegypti*, tornou-se necessário um entendimento e consequentemente um controle maior da evolução e do comportamento desta espécie. Neste sentido o projeto SAUDAVEL[1] está desenvolvendo um experimento na cidade de Recife/PE, que consiste na coleta de ovos deste mosquito, a fim de construir um ambiente para o monitoramento e vigilância, buscando contribuir para um maior controle de surtos epidêmicos. Este trabalho propõe ferramentas estatísticas para estimação de superfícies onde o fenômeno em estudo não foi observado.

Palavras-chave: Modelos Aditivos Generalizados, Análise Exploratória de Dados, R.

[1]Sistema de Apoio Unificado para Detecção e acompanhamento em Vigilância Epidemiológica

1 Introdução

O crescimento constante na ocorrência de Dengue tem causado vários problemas para a população e para os governos, não apenas no Brasil, mas em todo mundo, principalmente nos países de clima tropical. A dengue é transmitida pela fêmea do mosquito *Aedes aegypti*, pois o macho se alimenta apenas da seiva das plantas. Um desses mosquitos durante toda sua vida (em torno de 45 dias), pode contaminar até 300 pessoas. Diante disto houve a necessidade de se criar medidas para controle do avanço da doença.

Uma dessas medidas foi o controle da proliferação do mosquito na cidade de Recife, estudo esse que teve a iniciativa do projeto SAUDAVEL[2], o qual é uma subivisão do Instituto Nacional de Pesquisas Espaciais (INPE), que busca, através de parcerias com Instituições de ensino nacionais e governos, implementar métodos automatizados para controle de epidemias.

No estado do Paraná, a Universidade Federal do Paraná é representada neste projeto através do Laboratório de Estatística e Geoinformação[3]

2 Objetivos

O objetivo do trabalho é analisar os dados relativos ao número de ovos do mosquito *Aedes aegypti* coletados em armadilhas instaladas em cinco bairros da cidade de Recife. A análise estatística desenvolvida será utilizada para estimar a superfície não observada nos bairros em que haviam armadilhas, de forma que sirva de base para prever possíveis descontroles no número de ovos, e assim desenvolver planos de ação para controlar tais avanços.

Deseja-se que tal análise possa ser feita de forma automatizada, devido

[2]http://saudavel.dpi.inpe.br/

[3]http://leg.est.ufpr.br

ao grande número de informações disponíveis e a necessidade da obtenção da informação, para que possa ser possível o controle da epidemia mencionada. Para tal usamos o software R [R Development Core Team, 2007].

3 Procedimentos Metodológicos

Nesta seção será feito uma descrição do estudo, dos dados e da metodologia utilizada na análise.

3.1 Área de estudo, Instrumentos e Técnicas de Campo

O estudo teve início no ano de 2004 com a instalação de 564 armadilhas dispostas uniformemente em 5 bairros da cidade de Recife:

1. Brasília Teimosa
2. Dois Irmãos
3. Morro da Conceição
4. Casa Forte/Parnamirim
5. Engenho do Meio

com o objetivo de que a fêmea do mosquito *Aedes aegypti* fizesse o depósito dos ovos. As armadilhas foram então dividas em quatro grupos.

A cada uma semana (7 dias), um grupo de armadilhas (25%) em cada bairro é coletado para que a contagem do número de ovos seja realizada em laboratório e a informação seja encaminhada para análise. Vale destacar, que

a cada observação, a armadilha começa novamente do 0. A última atualização no banco de dados ocorreu em 15/05/2007.

Os dados são repassados para um Sistema Gerenciador de Banco de Dados[4] através de uma ferramenta de Sistema de Informação Geográfica[5] (No nosso caso estamos utilizando o TerraLib). Estes dados podendo ser acessadas remotamente desde que o usuário tenha permissão.

Utilizando o pacote estatístico `R` como ferramenta, dispomos de várias formas para leitura do banco de dados, dentre as quais podemos destacar a utilização dos pacotes:

- `RODBC` [Lapsley & Ripley, 2007]
- `RMySQL` [James & DebRoy, 2007]
- e o que usamos neste trabalho `aRT` [Andrade *et al.*, 2007]

A figura 1 abaixo ilustra a distribuição das armadilhas pertencentes ao grupo 1 no bairro Brasília Teimosa na sua segunda observação (25-05-2004), que será o bairro e a data em estudo neste relatório. Nesta figura os números representam os códigos das armadilhas e os círculos tem diâmetro proporcional ao número de ovos.

Pode-se notar que as armadilhas pertencentes à esse grupo se concetram na parte central do bairro, com exceção da armadilha 102.

[4]SGBD
[5]SIG

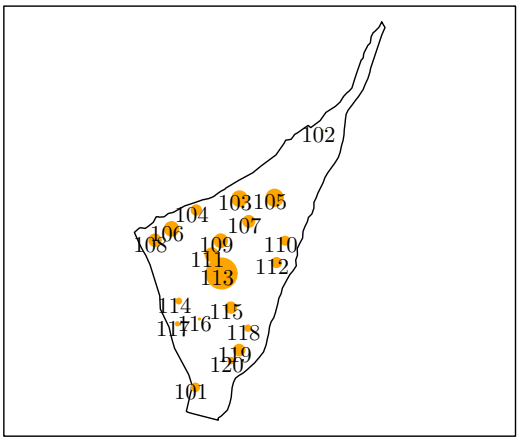

Figura 1: Distribuição espacial das armadilhas. Quanto maior o tamanho do círculo, maior a contagem de ovos realizada naquela armadilha. O número abaixo de cada armadilha representa o código da armadilha.

4 Metodologia Estatística

Para nosso objetivo, que é a estimação na superfície não observada, faremos uso dos Modelos Aditivos Generalizados, devido à sua estrutura oferecer uma flexibilidade na especificação da relação entre a variável resposta e as covariáveis.

4.1 Modelo Aditivo Generalizado

Um modelo aditivo generalizado[Hastie & Tibshirani, 1986] é um modelo linear generalizado, com um ou mais preditores lineares envolvendo uma soma de funções suaves das covariáveis. O modelo tem a seguinte estrutura:

$$g(\mu_i) = X_i^*\theta + f_i(x_1 i) + f_2(x_2 i) + f_3(x_3 i, x_4 i) + ...,$$

onde $\mu_i \equiv E(Y_i)$ e $Y_i \sim$ alguma distribuição da família exponencial.

Y_i é a variável resposta, X_i^* é uma linha da matriz do modelo, para a parte estritamente paramétrica, θ é o vetor de parâmetros correspondentes, e as f_j são as funções suaves das covariáveis x_k g é uma função de ligação conhecida.

4.1.1 Representação da Função Suave

A representação da função suave do modelo pode ser feita usando métodos comuns de estimação, tais como mínimos quadrados e máxima verossimilhança. Para tal é necessário que se escolha uma base, ou seja, definir o espaço de funções do qual f é um elemento. Para ilustrar, considere que f seja um função polinomial de quarta ordem.Uma base para este espaço é $b_1(x) = 1, b_2(x) = x, b_3(x) = x^2, b_4(x) = x^3$ e $b_5(x) = x^4$. A figura 2 ilustra a representação de uma função de base usando uma base polinomial. Os 5 primeiros paíneis ilustram as 5 funções de base $b_j(x)$ para uma base polinomial de ordem 4. As funções são então multiplicadas por uma valor real, β_j, e somadas para resultar na curva final $f(x)$

Bases polinomiais tendem a ser muito úteis em situações onde o foco de análise são as propriedades de f na vizinhança de um único ponto, porém quando o interesse é em toda a extensão do domínio da função as bases

polinomiais apresentam alguns problemas.Uma base *spline* tem melhor perfomance perante uma vasta quantidade de situações e pode ser mostrado que elas apresentam boas propriedade teóricas[Wood, 2006].

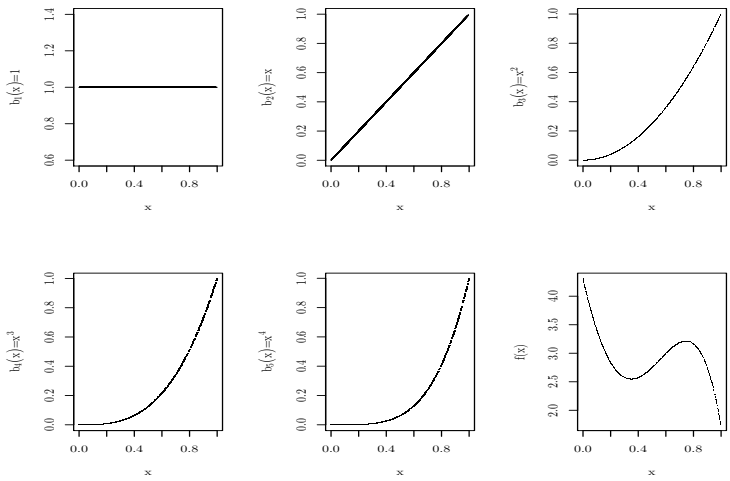

Figura 2: Representação de uma Base polinomial de quarta ordem

4.1.2 Escolhendo o parâmetro de suavização

Uma necessidade para o ajuste do modelo é escolher o melhor parâmetro de suavização λ, para os dados. Se o λ for muito pequeno o modelo alisará muito os dados, caso seja muito grande, o modelo não alisará os dados de forma suficiente, em ambos os casos significa que a *spline* estimada $\hat{f}$ não é uma boa aproximação para a função real f.A melhor escolha para λ é aquela que mais aproxima $\hat{f}$ da real função f.Um critério adequado para escolher λ seria escolher o valor de λ que minimiza a seguinte expressão:

$$M = \frac{1}{n}\sum_{i=1}^{n}(\hat{f}_i - f_i)^2$$

onde a notação $\hat{f}_i \equiv \hat{f}(x_i)$ e $f_1 \equiv f(x_i)$ foi adotada por conveniência.

Note que f é desconhecida , então M não pode ser usada diretamente, mas é possível derivar uma estimativa da $E(M) + \sigma^2$ que é o erro quadrático médio esperado ao prever uma nova variável.Define-se $\hat{f}^{[-i]}$ como o modelo ajustado com todos os dados exceto y_i, e define-se o escore ordinário da validação cruzada como:

$$v_0 = \frac{1}{n} \sum_{i=1}^{n} (f_i^{[-i]} - y_i)^2 \tag{1}$$

Este processo de escolha de λ que minimiza v_0 é conhecida como Validação Cruzada Ordinária.Este método é uma aproximação razoável , já que ele minimiza o erro quadrático de predição , se os modelos forem julgados apenas por sua capacidade preditiva, modelos mais complexos sempre serão escolhidos ao invés de modelos mais simples. Computacionalmente o processo de calcular v_0 para todos os dados e ajustar o modelo para cada um dos n componentes do conjunto de dados, entretanto pode-se decompor (1), chegando assim ao escore da Validação Cruzada Generalizada.

$$v_g = \frac{n \sum_{i=1}^{n} (y_i - \hat{f}_i)^2}{[tr(I - A)]^2}$$

O escore de validação cruzada ordinária(GCV) tem vantagens computacionais sobre o escore de validação cruzada ordinária(OCV), e é um dos critérios para escolha do melhor modelo, bem como o Critério de Informação de Akaike (AIC) e o coeficiente de correlação múltipla R^2 ajustado.

5 Resultados

Como a coleta é feita semanalmente, precisamos fazer a estimação da superfície para todas as datas coletadas, o que gera em torno de 160 datas para cada um dos 5 bairros observados. Sendo assim, este relatório irá conter apenas um modelo ilustrativo para o bairro Brasília Teimosa e para a data 25-05-2004. O restante da modelagem é feita de maneira automatizada pelo pacote estatístico `R`. Houve a necessidade de se criar funções próprias para este fim.

Para visualização da superfície gerada para cada uma das semanas do experimento, foi gerado um vídeo mostrando a evolução espaço-temporal do fenômeno em toda a área observada, seguindo o procedimento descrito na página `http://leg.est.ufpr.br/doku.php/projetos:saudavel`. Todos arquivos gerados estão disponíveis na página
`http://www.leg.ufpr.br/doku.php/pessoais:henriqued:lab`

Foram ajustados 3 modelos para cada um dos bairros:

1. $log(E(Z_i)) = f_1(x_i) + f_2(y_i)$
2. $log(E(Z_i)) = f_1(x_i) + f_2(y_i) + log(Z_i(t-1))$
3. $log(E(Z_i)) = f_1(x_i) + f_2(y_i) + f_3(log(Z_i(t-1)))$

Onde $Z_i \sim N(\mu, \sigma^2)$ é o número de ovos observados na data descrita acima, x_i, y_i são as coordenadas das armadilhas observadas, f_i são funções suaves das covariáveis e $Z_i(t-1)$ é a variável resposta defasada em um período (27-04-2004).

Conforme descrito na seção 3.1 o `R` dispõe de alguns pacotes para realizar a modelagem por Modelos Aditivos Generalizados. Neste trabalho foi usado o pacote `mgcv`[Wood, 2007].

O resultado do modelo é descrito na Tabela 1 abaixo

Modelo	GCV	AIC	Deviance	R^2_{adj}
1	0.33189	31.1485	85.8%	0.760
2	0.35745	30.93682	87.3%	0.764
3	0.17912	11.48958	96.2%	0.909

Tabela 1: Validação Cruzada Ordinária Generalizada, Critério de Informação de Akaike, Deviance Explicada e Coeficiente de Correlação R^2 ajustado, que leva em consideração a parcimonia do modelo

Conforme descrito na metodologia acima, o modelo escolhido será aquele que tenha o menor GCV, a maior *Deviance*, o maior R^2_{adj} e o menor AIC. Neste caso, o modelo 3.

Precisamos verificar o pressuposto de homocedasticidade do modelo, já que normalidade não é assumido, apenas se requerido, conforme descrito por Wood (2006).

O gráfico da figura 3 nos fornece indicação de que a variância é aproximadamente constante.

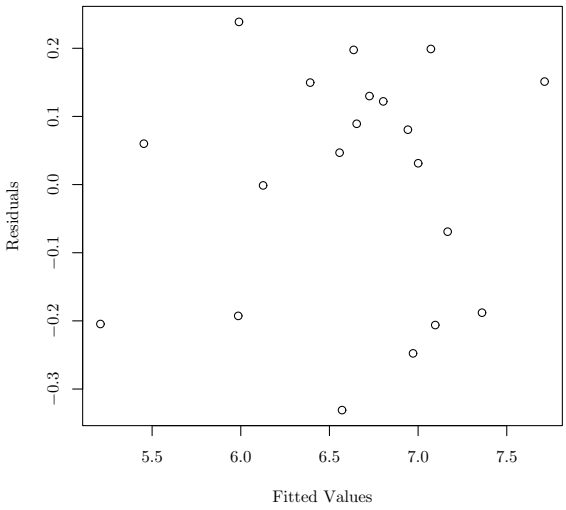

Figura 3: Valores Ajustdos *versus* Resíduos

Por fim, podemos fazer a interpolação para a área não observada. A maneira utilizada para interpolar a superfície foi, gerar uma nuvem de pontos sobre a extensão do bairro, que pode ser feito usando a função `expand.grid` do `R`. Após isso precisamos apenas predizer os dados gerados, usando o modelo ajustado, o que resultou no gráfico da figura 4.

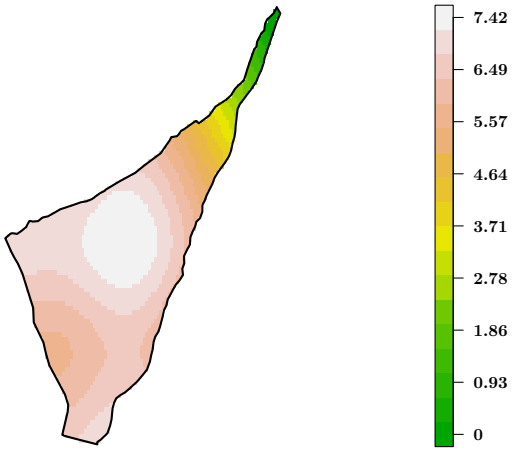

Figura 4: Superfície interpolada para o bairro Brasília Teimosa para o dia de 25 de maio de 2004. Como descrito anteriormente, os dados estão em escala logarítimica

6 Conclusão

Por se tratar de uma metodologia de análise recente, a análise e visualização de dados espaço-temporais ainda não nos permite avaliar comparações com outros métodos, como ocorre em outras áreas da estatística, porém como podemos perceber pela aplicação acima, o Modelo Aditivo Generalizado foi satisfatório, tendo em vista que através da figura 3 é possível entender a proliferação do mosquito em todo o bairro.

Devemos levar em consideração o baixo número de observações para todas as semanas observadas, em torno de 20 a 25 armadilhas observadas, o que

dificulta o ajuste de um modelo de tal complexidade como o GAM, já que este estima uma quantidade considerável de parâmetros.

Fica como sugestão para trabalhos futuros do projeto SAUDAVEL, o ajuste de um modelo que leve em consideração a estrutura de covariancia espaço-temporal presente nos dados.

7 Referências

Referências

[Andrade *et al.*, 2007] Andrade, Neto Pedro R., A., Carrero Marcos, S., Silva Eduardo, & J., Ribeiro Jr Paulo R. 2007. *aRT: R-terralib api*. R package version 1.4-1.

[Hastie & Tibshirani, 1986] Hastie, T., & Tibshirani, R. 1986. Generalized additive models. *Statistical science*, **1**, 297–318.

[James & DebRoy, 2007] James, David A., & DebRoy, Saikat. 2007. *Rmysql: R interface to the mysql database*. R package version 0.6-0.

[Lapsley & Ripley, 2007] Lapsley, Michael, & Ripley, B. D. 2007. *Rodbc: Odbc database access*. R package version 1.2-2.

[R Development Core Team, 2007] R Development Core Team. 2007. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

[Wood, 2006] Wood, Simon. 2006. *Generalized additive models (texts in statistical science)*. Chapman & Hall/CRC.

[Wood, 2007] Wood, Simon. 2007. *Gams with gcv smoothness estimation and gamms by reml/pql*. R package version 1.3-23.