

Modelos Aditivos Generalizados (GAM): Uma visão prática

Eduardo Camargo
INPE/DPI

29/09/2015

- O modelo de regressão simples – uma breve recordação

$$Y_i = \beta_0 + \beta_1 X_i + \xi_i$$

em que:

Y : é denotada de variável dependente ou resposta

X : variável independente

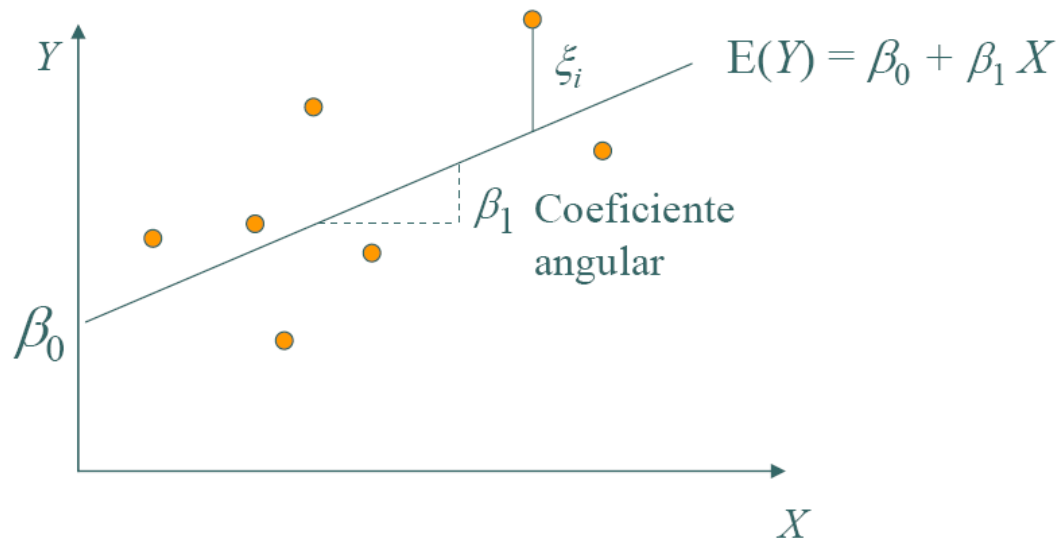
β_0 : intercepto

β_1 : inclinação

ξ_i : erro aleatório, $\xi_i \sim N(0, \sigma^2)$, $\xi_i = Y_i - \beta_0 - \beta_1 X_i$,

- O modelo de regressão simples – uma breve recordação

$$Y_i = \beta_0 + \beta_1 X_i + \xi_i$$



■ O modelo de regressão simples – uma breve recordação

$$Y_i = \beta_0 + \beta_1 X_i + \xi_i$$

- Em geral não se conhece os parâmetros β_0 e β_1 .
- Eles podem ser estimados através de dados obtidos por amostras.
- O método utilizado na estimação de β_0 e β_1 é o método dos mínimos quadrados, o qual considera os desvios dos Y_i de seu valor esperado:

$$\xi_i = Y_i - \beta_0 - \beta_1 X_i$$

- O método dos mínimos quadrados requer que consideremos a soma dos n desvios quadrados, denotado por Q :

$$Q = \sum_{i=1}^n [Y_i - \beta_0 - \beta_1 X_i]^2$$

■ O modelo de regressão simples – uma breve recordação

$$Y_i = \beta_0 + \beta_1 X_i + \xi_i$$

- De acordo com o método dos mínimos quadrados, os estimadores de β_0 e β_1 são aqueles, denotados por b_0 e b_1 , que tornam mínimo o valor de Q .

- Derivando:
$$\frac{\partial Q}{\partial \beta_0} = -2 \sum_{i=1}^n [Y_i - \beta_0 - \beta_1 X_i] \quad \frac{\partial Q}{\partial \beta_1} = -2 \sum_{i=1}^n [Y_i - \beta_0 - \beta_1 X_i] X_i$$

- Igualando as equações a zero obtém-se os valores b_0 e b_1 que minimizam Q :

$$b_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} \quad \Rightarrow \quad \begin{aligned} \hat{Y} &= b_0 + b_1 X \\ e_i &= Y_i - \hat{Y}_i \quad (\text{resíduo}) \end{aligned}$$
$$b_0 = \bar{Y} - b_1 \bar{X}$$

■ Síntese - GAM

- *Modelo de Regressão Linear Múltiplo* $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \dots$

- *Modelo Linear Generalizado (MLG)*. Ingredientes básicos:

1. K valores independentes $Y_1, \dots, Y_K$, de uma variável resposta que segue uma distribuição da família exponencial, com valor esperado $E(Y_i) = \mu_i$;
2. Uma função de ligação, denotada por $g(\mu_i)$, tal que:

$$g(\mu_i) = \mathbf{X}\boldsymbol{\beta}$$

$\mathbf{X}$: vetor das variáveis explicativas.

$\boldsymbol{\beta}$: representa o vetor de parâmetros a serem estimados.

$$g(\mu_i) = E(Y_i)$$

■ Síntese - GAM

Modelo Aditivo Generalizado - GAM

- É uma extensão do *MLG*, em que o termo $\Sigma \mathbf{X}\boldsymbol{\beta}$ é substituído por $\Sigma f(\mathbf{X})$, assim:

$$g(\mu_i) = \Sigma f(\mathbf{X})$$

- $f(\mathbf{X})$ é uma função não paramétrica (i.e. cuja forma não é especificada)
- $f(\mathbf{X})$ é estimada através de curvas de alisamento (ex: *splines*).
- A curva alisada permite descrever a forma e revelar possíveis não linearidades nas relações estudadas, uma vez que não apresenta a estrutura rígida de uma função paramétrica, como nos *MLG*'s.

■ Síntese – GAM: exemplos em R

Pacote MGCV

```
require(mgcv)
```

Modelo sem suavização em *vexp* (variável explicativa) e sem efeito espacial

```
fit <- gam(vresp ~ vexp, family=binomial, data=gam.data)
```

Modelo com suavização em *vexp* e sem efeito espacial

```
fit <- gam( vresp ~ s(vexp), family=binomial, data=gam.data)
```

Modelo sem suavização em *vexp* e com efeito espacial *s(x,y)*

```
fit <- gam(vresp ~ vexp + s(x,y), family=binomial, data=gam.data)
```

Modelo com suavização em *vexp* e com efeito espacial *s(x,y)*

```
fit <- gam (vresp ~ s(vexp) + s(x,y), family=binomial, data=gam.data)
```

■ SPGAM um modelo alternativo ao GAM

- $g(u_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \dots + k(s)$
- Uma abordagem semiparamétrica.
- $k(s)$ é uma função de *Kernel*, em que s é um vetor de coordenadas da variável resposta.
- Exemplo:

Distribuição espacial do risco: modelagem da mortalidade infantil em Porto Alegre, Rio Grande do Sul, Brasil (*Silvia E. Shimakura, Marília Sá Carvalho, Denise R. G. C. Aerts, Rui Flores*)

$$g(u_i) = \log \left\{ \frac{p(s, x)}{1 - p(s, x)} \right\} = \beta_0 + \beta_1 \text{sexo} + \beta_2 \text{peso} + \beta_3 \text{idade} + \beta_4 \text{inst} + \beta_5 \text{ges} + \beta_6 \text{grav} + \beta_7 \text{parto} + k(s).$$

Aplicação de Modelos Aditivos Generalizados (GAM)

10

■ Projeto



<http://www.dpi.inpe.br/eureqa>

■ Processo FAPESP No. 2006/53922-9

■ Objetivo Principal:

Investigar a correlação entre resistência bacteriana e fatores de risco populacionais (em particular uso populacional de antimicrobianos).

Coordenador Geral

Prof. Dr. Antônio Miguel Vieira Monteiro¹

Coordenadores Setoriais

Prof. Dr. Antônio Carlos C. Pignatari³

Dr. Caio Márcio Figueiredo Mendes²

Dr. Carlos Roberto Velga Kiffer⁶

Dr. Eduardo Celso Gerbi Camargo¹

Pesquisadores

Dra. Soraya Andrade³

Dra. Corina da Costa Freitas¹

Dra. Virginia Ragoni de Moraes Correia¹

Dra. Elisabete Carla Moraes¹

Dra. Jussara de Oliveira Ortiz¹

Dra. Silvia Emiko Shimakura⁷

Dr. Paulo Justiniano Ribeiro Jr.⁷

Dr. Getúlio Batista⁴

Msc. Dra. Bianca Lucarevski⁴

Msc. Paula Celia Mariko Koga²

Msc. Gabriel Pereira¹

Amilton Mouro⁵

Tiago dos Santos Agostinho⁴

¹ Instituto Nacional de Pesquisas Espaciais - INPE

² Fleury Medicina e Saúde

³ UNIFESP / Escola Paulista de Medicina (EPM) / Laboratório Especial em Microbiologia Clínica (LEMC)

⁴ Universidade de Taubaté - UNITAU

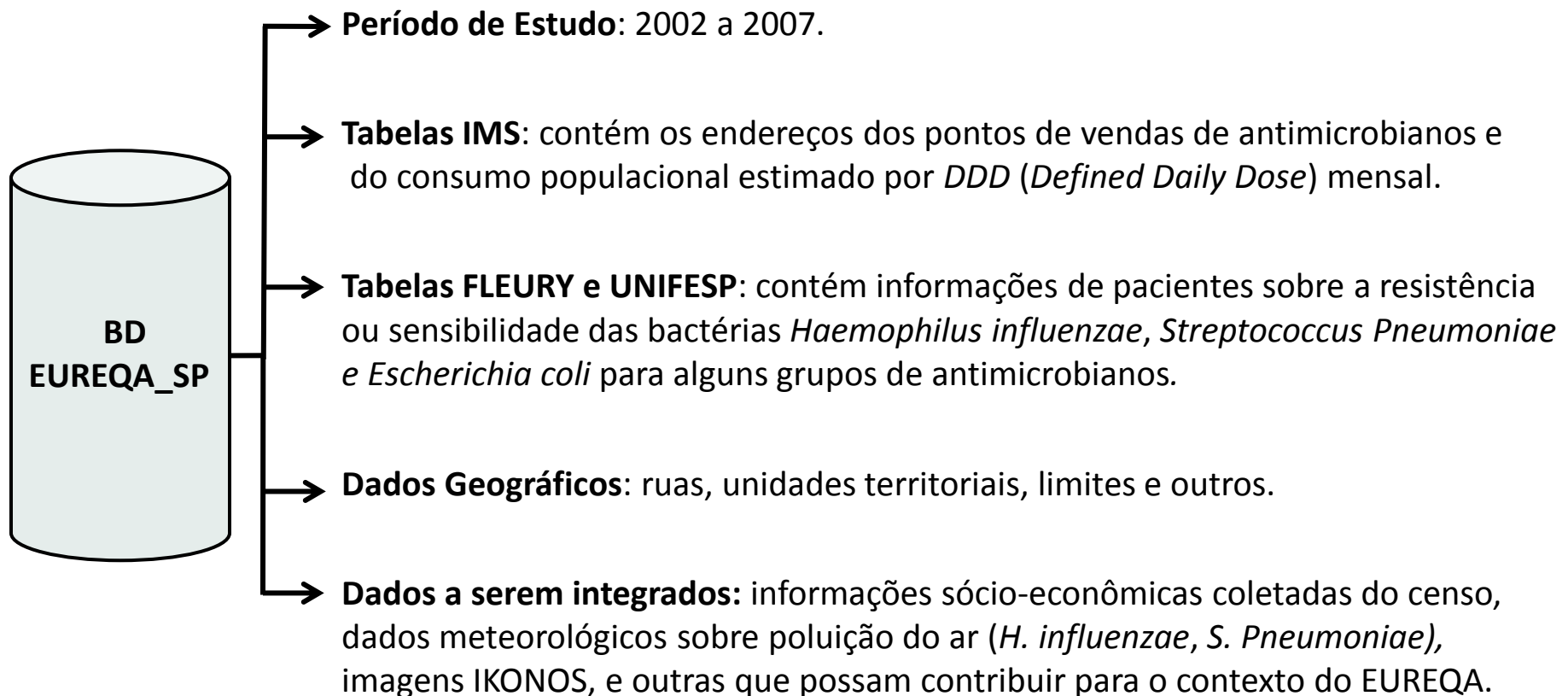
⁵ Hospital Israelita Albert Einstein

⁶ Gestão do Conhecimento Científico - GC₂

⁷ Laboratório de Estatística e Geoinformação - LEG, Universidade Federal do Paraná

29/09/2015

■ 1ª Fase: Construção do Banco de Dados EUREQA





■ 3ª Fase: Realização de um Primeiro Estudo

- Investigar se há associação do risco de resistência bacteriana, oriunda de *E. coli*, com o consumo populacional de ciprofloxacina expresso pela *DDDD*.
- População: mulheres maior que 16 anos.
- Área de estudo: cidade de São Paulo, 2002.
- Covariável desconhecida (*DDDD*): densidade de uso de antimicrobiano populacional.

$$DDDD = \frac{DDD}{(\text{Pop}/1000) * 30}$$

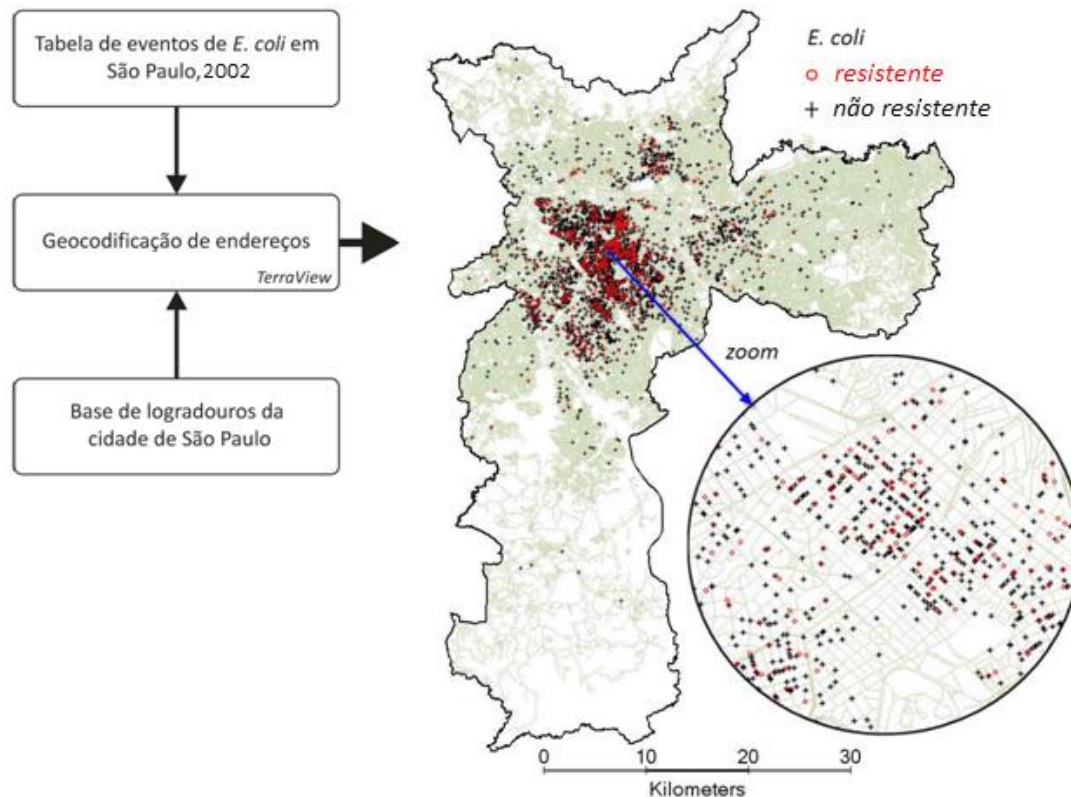
“Defined Daily Dose” (OMS)
informação conhecida
adquirida da IMS

?

Aplicação de Modelos Aditivos Generalizados (GAM)

14

- Preparação dos dados para o GAM
- Espacialização dos eventos, *resistentes e não resistentes*, decorrente de *E. coli*



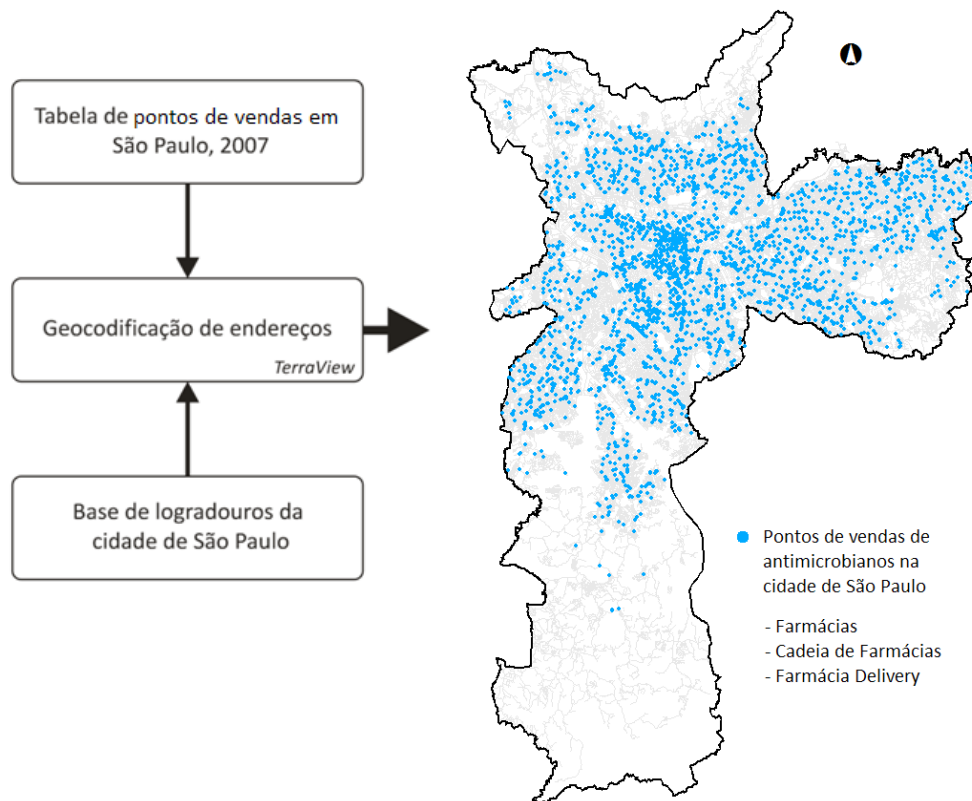
29/09/2015

Aplicação de Modelos Aditivos Generalizados (GAM)

15

■ Preparação dos dados para o GAM

➤ Espacialização dos pontos de vendas de antimicrobianos na cidade de São Paulo



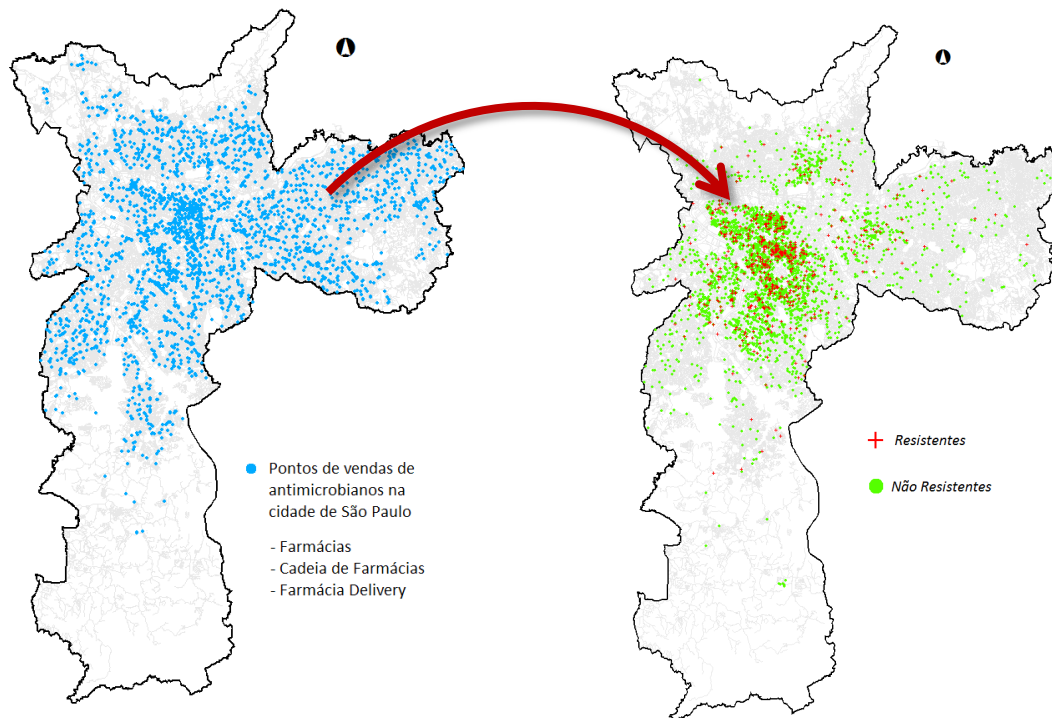
29/09/2015

Aplicação de Modelos Aditivos Generalizados (GAM)

16

■ Preparação dos dados para o GAM

Como calcular *DDDD*? Como associar valores de *DDDD* aos eventos *resistentes* e *não resistentes*?



$$DDDD = \frac{DDD}{(Pop/1000)*30}$$

↓
?

29/09/2015

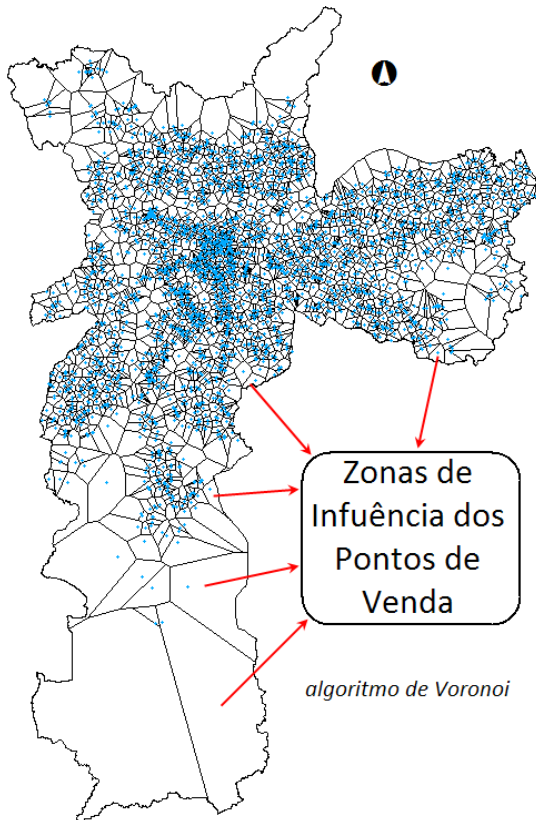
Aplicação de Modelos Aditivos Generalizados (GAM)

17

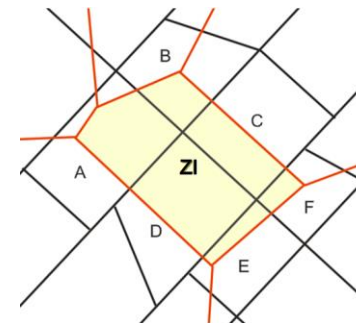
■ Preparação dos dados para o GAM

Solução adotada para o cálculo da *DDDD*

Parte I: construção de Zonas de Influência (ZI)



Parte II: Interseção dos mapas Zonas de Influência com Unidades Territoriais (Fundação SEADE - SP)



Unidades Territoriais: A, B, C, D, E, F
ZI: Zona de Influência

$$Pop_{ZI} = Pop_A \left(\frac{\text{Área}_{ZI_A}}{\text{Área}_A} \right) + Pop_B \left(\frac{\text{Área}_{ZI_B}}{\text{Área}_B} \right) + Pop_C \left(\frac{\text{Área}_{ZI_C}}{\text{Área}_C} \right) + \\ + Pop_D \left(\frac{\text{Área}_{ZI_D}}{\text{Área}_D} \right) + Pop_E \left(\frac{\text{Área}_{ZI_E}}{\text{Área}_E} \right) + Pop_F \left(\frac{\text{Área}_{ZI_F}}{\text{Área}_F} \right)$$

$$DDDD_{ZI} = \frac{DDD}{\frac{Pop}{1000} * 30}$$

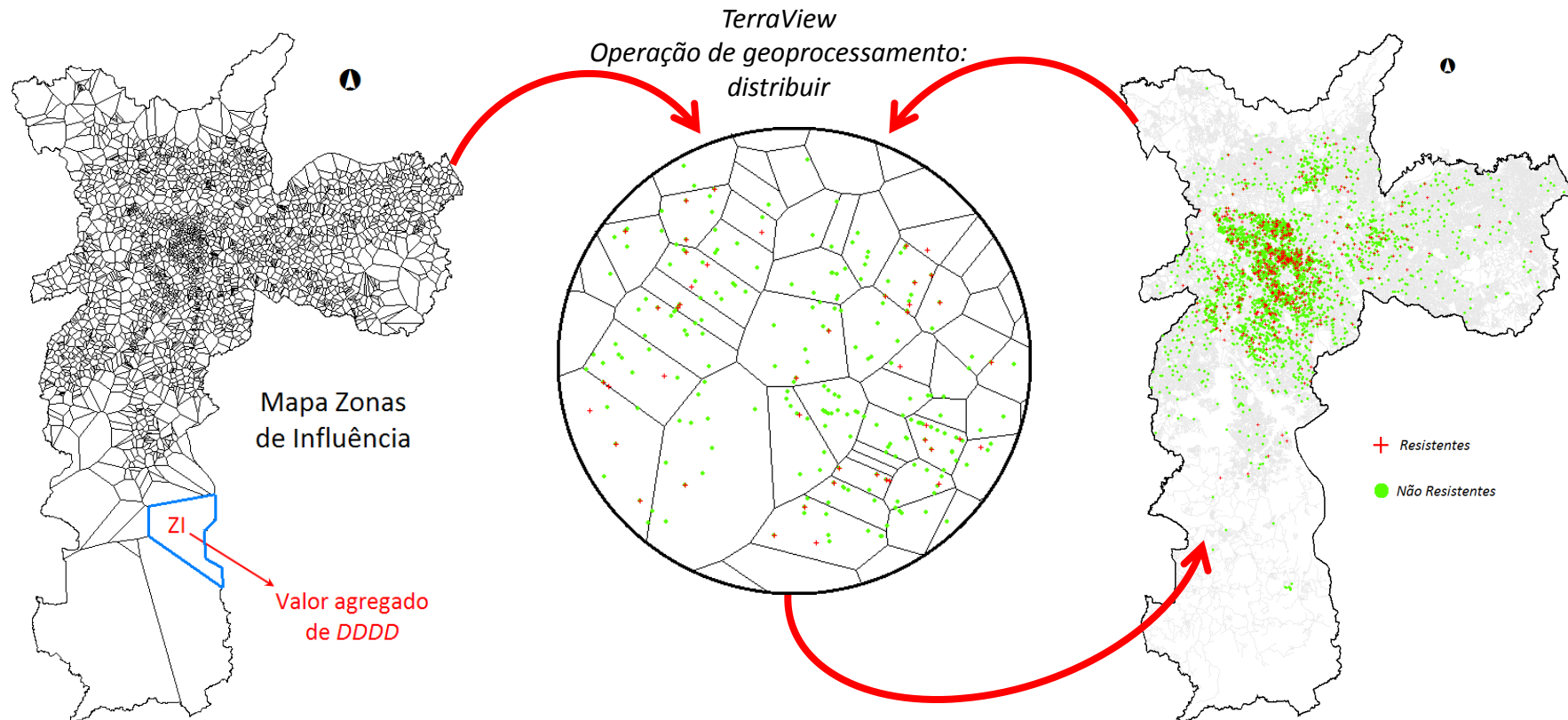
29/09/2015

Aplicação de Modelos Aditivos Generalizados (GAM)

18

■ Preparação dos dados para o GAM

Associando valores de *DDDD* aos eventos *resistentes* e *não resistentes*

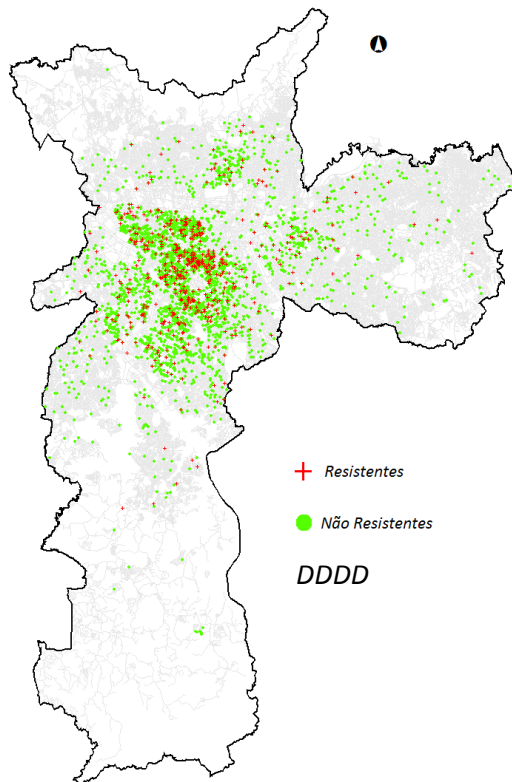


29/09/2015

Aplicação de Modelos Aditivos Generalizados (GAM)

19

■ Preparação dos dados para o GAM



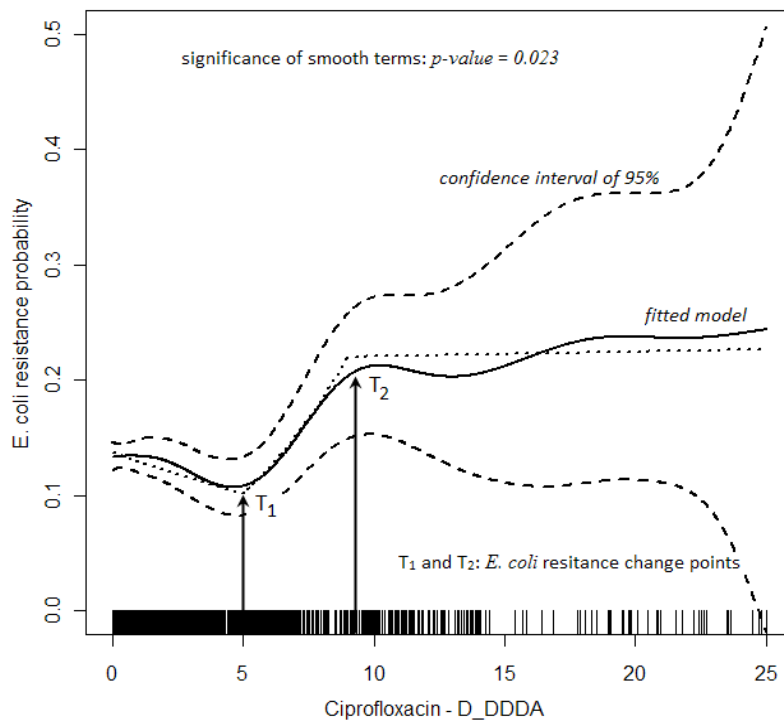
COORD_X	COORD_Y	Y	DDDD
330800.361	7395872.440	0	0.342
326544.146	7386268.087	1	0.202
332509.240	7390334.753	1	0.192
329254.299	7389932.687	0	0.416
326670.859	7379104.028	0	0.565
328809.895	7386776.251	0	0.887
329077.650	7386911.517	1	0.887
353109.185	7395543.311	0	0.389
330423.156	7392924.632	0	0.324
332164.392	7388986.068	0	1.212
332358.686	7388984.142	1	1.212
334261.560	7390542.337	0	0.366
326341.197	7393667.842	0	0.567
333549.674	7390623.169	1	0.524
331057.184	7392260.967	0	1.015
331076.514	7392342.553	0	1.015
331292.362	7392105.803	0	1.015
328075.540	7382259.943	1	0.525
325996.190	7387793.954	0	0.480
326128.331	7387648.820	0	0.480
325731.385	7387640.003	0	0.480
325652.046	7387592.818	1	0.480
330605.768	7395941.163	0	0.906
327428.067	7395592.610	1	0.525
333659.078	7392205.810	0	0.197
328254.306	7390217.238	0	0.377
339534.373	7394879.061	1	0.714
334233.900	7389921.206	0	0.845

29/09/2015

Aplicação de Modelos Aditivos Generalizados (GAM)

20

■ Pacote SPGAM (R) com modelo segmentado:



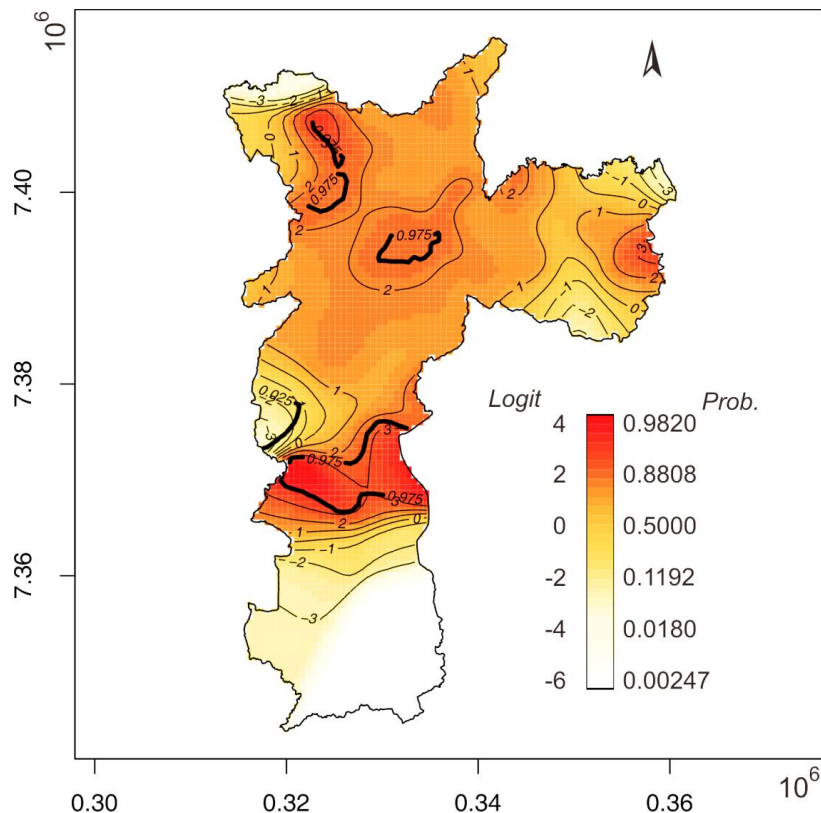
$$\log \left\{ \frac{p(s, x)}{1 - p(s, x)} \right\} = \beta_0 + \beta_1 D_DDDA + \beta_2 I_1 + \beta_3 I_2 + g(s)$$

- β_0 : regression model intercept;
- β_1, β_2 and β_3 : covariates effects;
- D_DDDA : antimicrobial usage mean density;
- $I_1 = 1$ if $T_1 < D_DDDA < T_2$ and $I_1 = 0$ otherwise;
- $I_2 = 1$ if $D_DDDA > T_2$; and $I_2 = 0$ otherwise;
- $T_1 = 5$ and $T_2 = 9$ E. coli resistance change points.

Aplicação de Modelos Aditivos Generalizados (GAM)

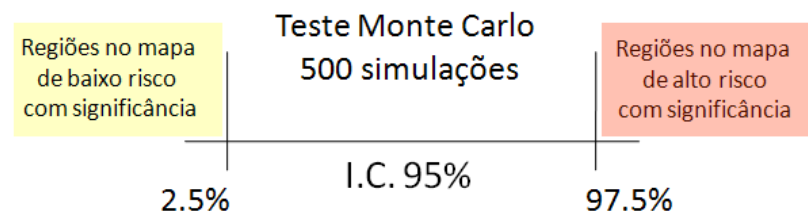
21

Estimação do mapa de risco



$$\log \left\{ \frac{p(s, x)}{1 - p(s, x)} \right\} = \beta_0 + \beta_1 D_DDDA + \beta_2 I_1 + \beta_3 I_2 + g(s)$$

Os contornos de tolerância e o teste global da hipótese nula de risco não variável espacialmente são obtidos via



p-valor do teste global de risco constante na região

0.001996008

29/09/2015