

Estatística Espacial (MI418) / Geoestatística (ME907)

Guilherme Ludwig

2019-02-05

Fenômenos que podem ser modelados com processos pontuais

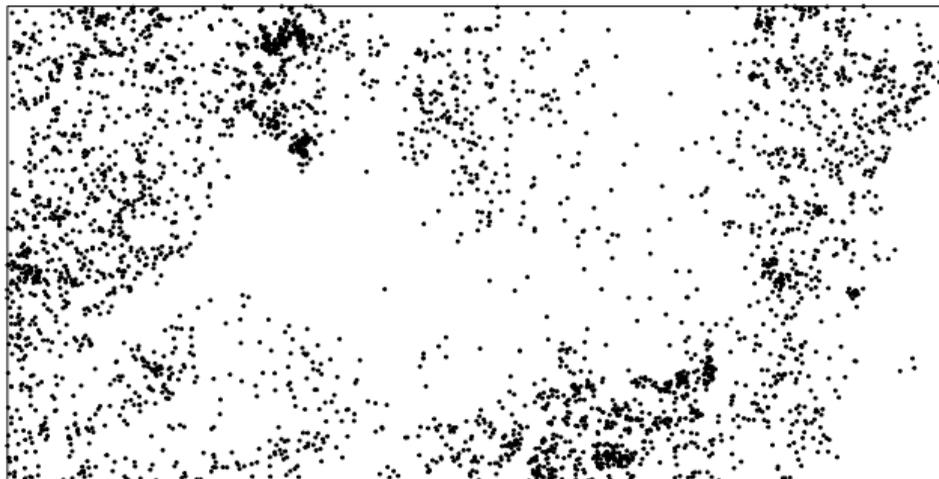
Processo de Poisson homogêneo, máxima verossimilhança

Processo de Poisson não-homogêneo: Regressão

Processo de Poisson não-homogêneo: Inferência não paramétrica

Exemplo: árvores tropicais

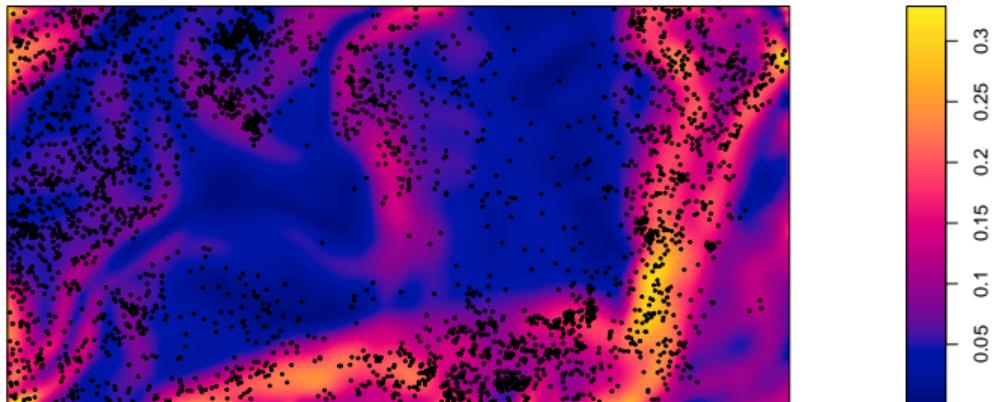
Beilschmiedia pendula



Árvores do tipo *Beilschmiedia pendula*, na ilha de Barro Colorado (Panamá). São 3605 árvores em uma janela de $1000 \times 500m^2$. Fonte: spatstat::bei.

Exemplo: árvores tropicais

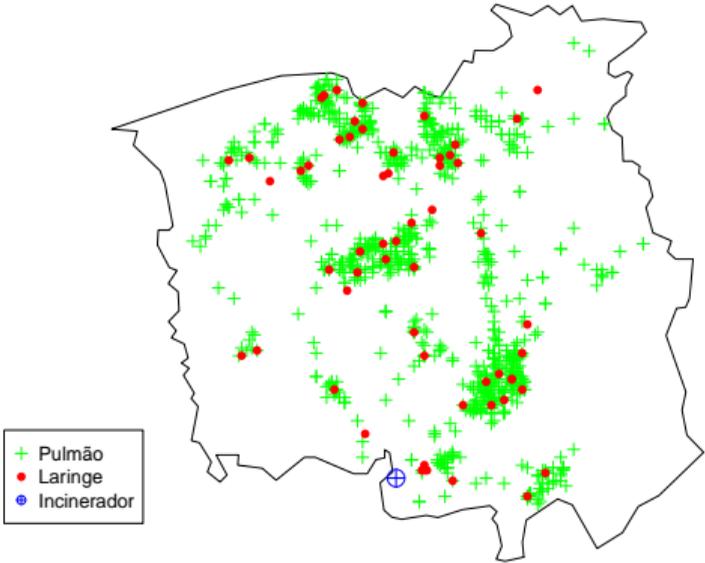
Beilschmiedia pendula + Slope gradient



Sobreposição das árvores à covariável correspondente à inclinação do terreno (efeito ambiental).

Chorley and South Ribble, UK

Câncer em Chorley-Ribble



Revisão conceitual

- ▶ Se o processo de Poisson é homogêneo, as estimativas têm fórmula fechada.
- ▶ Caso o processo seja não-homogêneo, mas a intensidade é função de covariáveis que eu conheço no domínio (ou pelo menos na janela inteira), também é possível fazer regressão.
- ▶ Caso contrário, o melhor que podemos fazer é encontrar um estimador não-paramétrico para a intensidade.
- ▶ Exemplos e parte do código usa o pacote `spatstat` (Baddeley et al., 2005).

Função de intensidade

Defina a variável aleatória $N(\mathbf{X} \cap B) = n(\mathbf{X}_B)$, e a medida de intensidade μ por

$$\mu(B) = \mathbb{E}(N(\mathbf{X} \cap B)).$$

A *função de intensidade* ρ é dada por

$$\mu(B) = \int_B \rho(\mathbf{s}) d\mathbf{s},$$

e se $\rho(\mathbf{s}) = \rho$ para todo \mathbf{s} , o processo é chamado de *homogêneo*.

Processo de Poisson

Seja μ uma medida difusa e localmente finita, tal que

$$\mu(B) = \int_B \rho(\mathbf{s}) d\mathbf{s}.$$

Um processo de Poisson no \mathbb{R}^2 pode ser caracterizado como

1. Para qualquer $B \subset \mathbb{R}^2$ com $\mu(B) < \infty$, $N(B) \sim \text{Poisson}(\mu(B))$.
2. Se $B_1, B_2 \subset \mathbb{R}^2$ são disjuntos, então $N(B_1), N(B_2)$ são independentes.

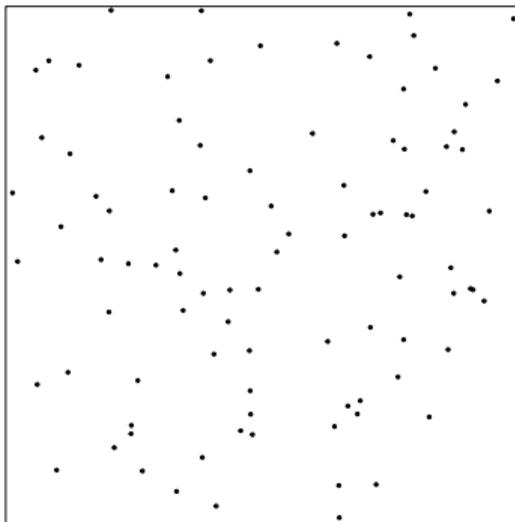
Equivalente:

2'. Seja f uma função de densidade, $f = \rho / \int \rho$, x_1, \dots, x_n condicionado a $N = n$ tem densidade conjunta

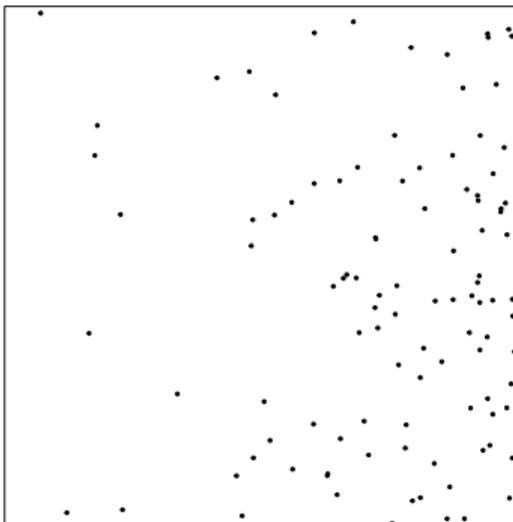
$$f_n(x_1, \dots, x_n | N = n) = \prod_{i=1}^n f(x_i)$$

Processo de Poisson simulado na caixa $[0, 1]^2$

PP Poisson homogêneo, $\rho = 100$



Não-homogêneo, $\rho(\mathbf{s}) = \exp(2 + 4s_1)$



Estatísticas

- ▶ Para um processo pontual no \mathbb{R}^d ,

$$\mathbb{E}(N(B)) = \mu(B).$$

No caso de um processo de Poisson homogêneo, $\mu(B) = \rho|B|$.

- ▶ O *momento fatorial de segunda ordem* para $C \subset \mathcal{D} \times \mathcal{D}$ é

$$\alpha^{(2)}(C) = \mathbb{E} \left(\sum_{\xi, \eta \in X: \xi \neq \eta} \mathbf{1}_{\{(\xi, \eta) \in C\}} \right).$$

- ▶ Em particular, note

$$\mathbb{E}(N(B_1)N(B_2)) = \alpha^{(2)}(B_1 \times B_2) + \mu(B_1 \cap B_2).$$

Estatísticas

- ▶ Se $\rho^{(2)}$ existir, reescreva

$$\alpha^{(2)}(C) = \int \int \mathbf{1}(\{(\xi, \eta) \in C\}) \rho^{(2)}(\xi, \eta) d\xi d\eta,$$

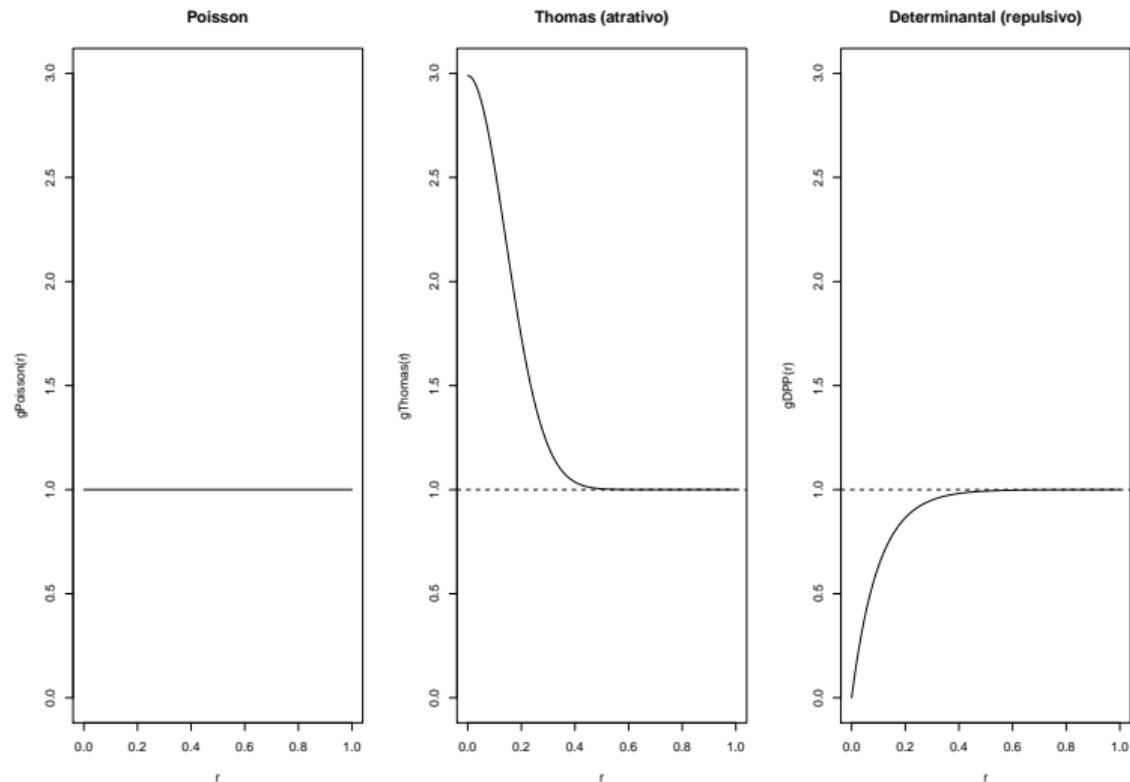
com $\rho^{(2)} > 0$. Então $\rho^{(2)}$ é a *densidade de segunda ordem*. No caso do processo de Poisson, $\rho^{(2)}(\xi, \eta) = \rho(\xi)\rho(\eta)$.

- ▶ Se ρ e $\rho^{(2)}$ existirem, a *função de correlação* é

$$g(\xi, \eta) = \frac{\rho^{(2)}(\xi, \eta)}{\rho(\xi)\rho(\eta)}.$$

- ▶ No caso de Poisson, $g(\xi, \eta) = 1$. Processos onde $g(\xi, \eta) > 1$ implicam que pares de pontos são mais prováveis de serem observados em ξ, η do que o caso de Poisson.
- ▶ Processos estacionários têm $g(\xi, \eta) = g(\eta - \xi)$.

Estatísticas



Note: a função de correlação pode não ter forma fechada.

Complete Spatial Randomness

Um dos primeiros testes χ^2 foi implementado para testar homogeneidade espacial

```
library(spatstat)
set.seed(1)
X <- rpoispp(100, win = owin(c(0,1), c(0,1)))
f <- function(x,y) exp(2+4*x)
Y <- rpoispp(f, win = owin(c(0,1), c(0,1)))
Test.A <- quadrat.test(X, 3, 3); Test.A$p.value
```

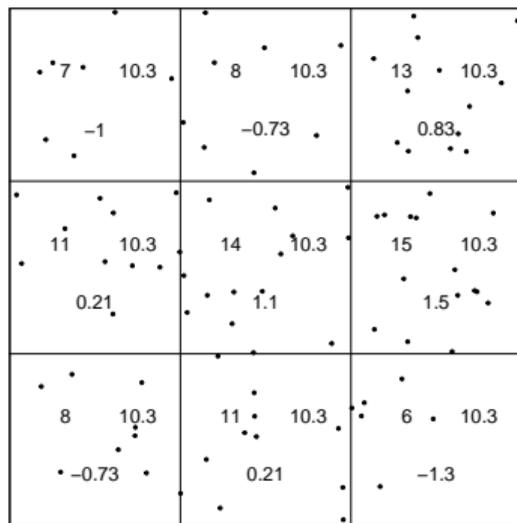
```
## [1] 0.841937
```

```
Test.B <- quadrat.test(Y, 3, 3); Test.B$p.value
```

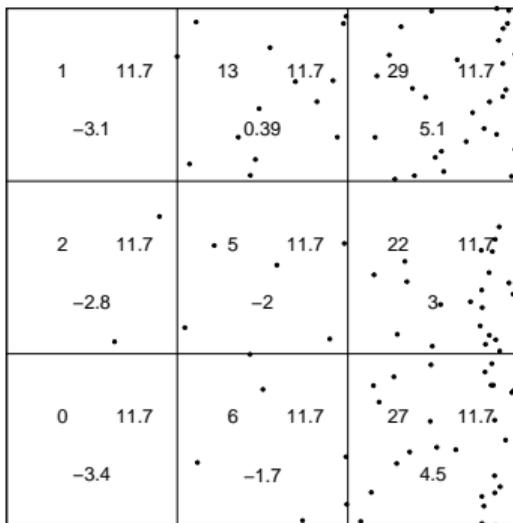
```
## [1] 5.306599e-16
```

Complete Spatial Randomness

Teste quando PPP homogêneo, $\rho = 100$



Teste quando PPPNH, $\rho(\mathbf{u}) = \exp(2 + 4u_1)$



Poisson: máxima verossimilhança

A verossimilhança de um processo corresponde à distribuição condicional (em $N(X) = n$) do processo (essas medidas são chamadas *medidas de Janossy*, a Profa. Nancy tem um conjunto de notas sobre elas:

http://www.ime.unicamp.br/~nancy/Cursos/mi626/mi626_2012.html.)

Em particular, a verossimilhança de uma configuração de pontos de um processo de Poisson é dada por

$$\ell(\rho) = \sum_{i=1}^n \log \rho(x_i) - \int_W \rho(\mathbf{u}) d\mathbf{u},$$

que no caso homogêneo torna-se

$$\ell(\rho) = n \log(\rho) + \rho |W|,$$

e conseqüentemente $\hat{\rho} = n/|W|$.

Poisson: máxima verossimilhança

A verossimilhança de uma configuração de pontos de um processo de Poisson é dada por

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^n \log \rho(x_i; \boldsymbol{\beta}) - \int_W \rho(\mathbf{u}; \boldsymbol{\beta}) d\mathbf{u}$$

Uma suposição comum para processos de Poisson não-homogêneos é o modelo log-linear da intensidade

$$\log(\rho(\mathbf{u})) = \beta_0 + \sum_{j=1}^p \beta_j z_j(\mathbf{u}).$$

Note que as covariáveis z_j precisam ser conhecidas para todo ponto na região de interesse W , e não somente nos pontos x_1, \dots, x_N .

Regressão para Processos de Poisson:

Berman and Turner (1992) encontram a solução aproximada de máxima verossimilhança:

- ▶ Segmentam a caixa em M células, de área a . Um ponto *dummy* x_i , $i = n + 1, \dots, n + M$ é colocado no centro de cada célula.
- ▶ Seja (x_i, Δ_i) , $i = 1, \dots, n + M$ um par com a localização (x_i) e Δ_i a marca do ponto *dummy* ($\Delta_i = 0$) ou real ($\Delta_i = 1$).
- ▶ Seja n_i o número de pontos na i -ésima célula, $w_i = n_i^{-1} a$
- ▶ A log-verossimilhança é aproximadamente

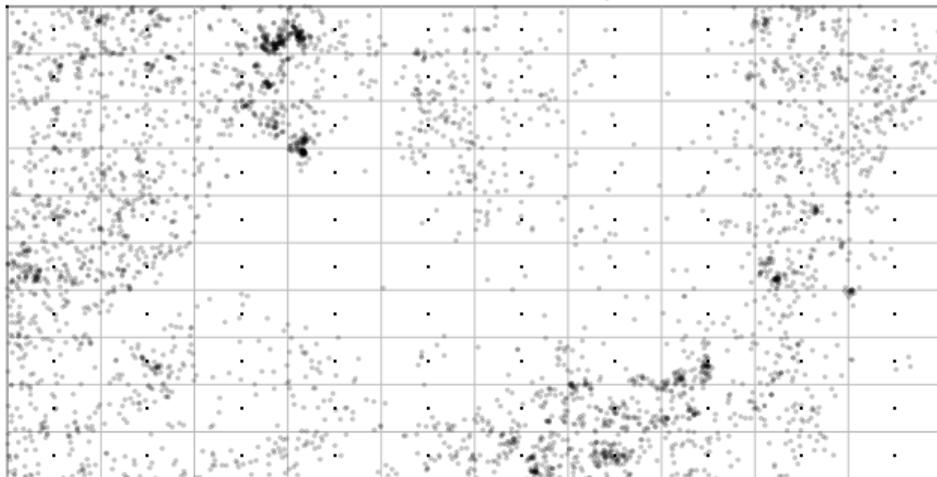
$$\ell(\beta) = \sum_{i=1}^{n+M} w_i \left(w_i^{-1} \Delta_i \log \rho(x_i; \beta) - \rho(x_i; \beta) \right)$$

- ▶ Esta verossimilhança coincide com o MLG Poisson ponderado (note que $w_i^{-1} \Delta_i$ pode não ser inteiro).

Algoritmo esta implementado em `spatstat::ppm`.

Quadratura

Quadrature Scheme, 10×10 grid



Regressão da Beilschmiedia pendula

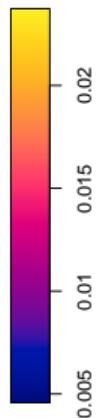
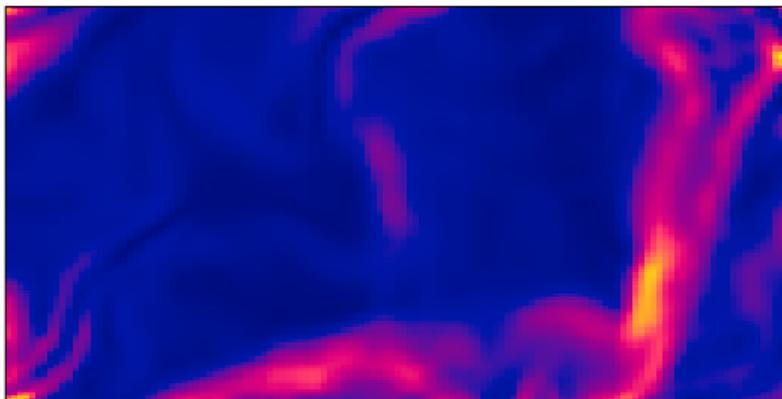
O *default* na quadratura da função ppm é mais refinado (32×32).

```
library(spatstat)
slope <- bei.extra$grad
model <- ppm(bei ~ slope)
coef(summary(model))
```

##	Estimate	S.E.	CI95.lo	CI95.hi	Ztest	Zval
## (Intercept)	-5.390553	0.03001716	-5.449385	-5.331720	***	-179.58236
## slope	5.022021	0.24540264	4.541041	5.503002	***	20.46441

Regressão da Beilschmiedia pendula

```
plot(predict(model), main = "")
```



Resíduos

Há algumas alternativas possíveis para estudar resíduos, veja Baddeley et al. (2015). Por exemplo,

- ▶ Obter estimativa não-paramétrica $\tilde{\rho}$ e comparar com o modelo paramétrico (`residuals.ppp()` e `plot.msr()` fazem isso).
- ▶ Usar $\alpha/\hat{\rho}$ como *thinning* probabilístico independente: o processo resultante será Poisson com intensidade α (pode falhar com caixas muito grandes ou intensidades muito pequenas, escolha um α que funcione).

Vou mostrar o resíduo do segundo tipo; depois introduzimos métodos não-paramétricos e voltaremos no exemplo. . .

Resíduos

```
npoints(bei); area(bei)
```

```
## [1] 3604
```

```
## [1] 5e+05
```

```
locations <- cbind(bei$x, bei$y)
```

```
p <- predict(model, locations = bei)
```

```
set.seed(1)
```

```
# Higher delete probability when intensity is large
```

```
delete <- !(runif(npoints(bei)) > 0.001/p)
```

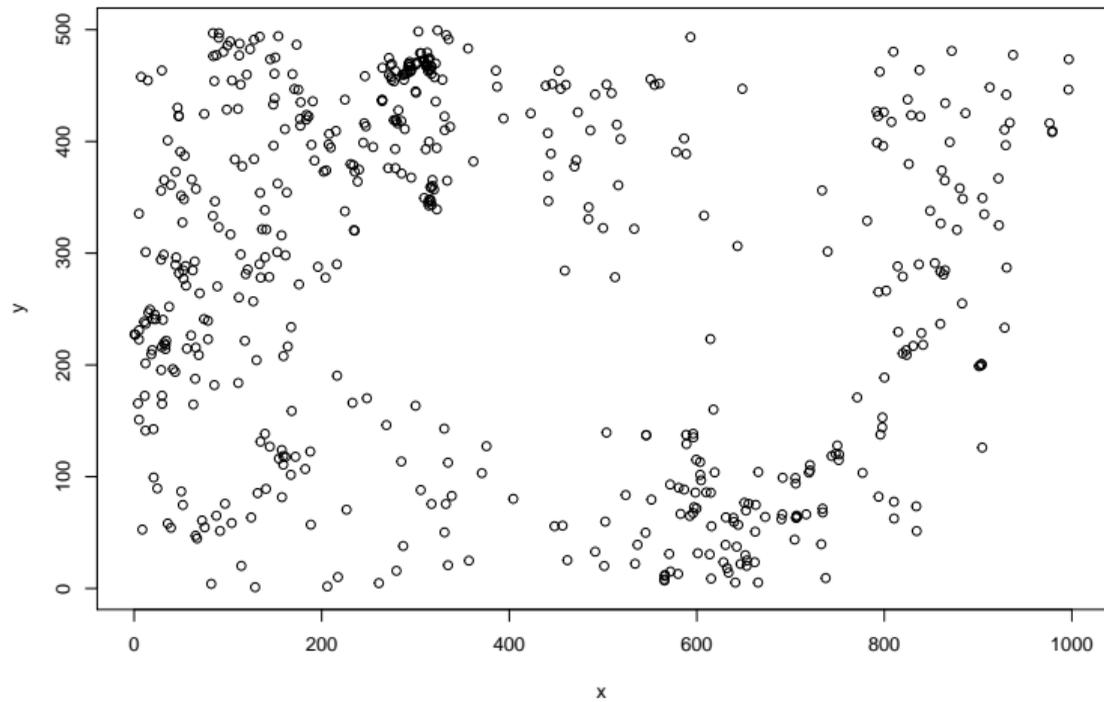
```
table(delete)
```

```
## delete
```

```
## FALSE TRUE
```

```
## 3071 533
```

Resíduos



Resíduos

```
r <- ppp(x = bei$x[delete], y = bei$y[delete],  
        window = bei>window)  
quadrat.test(r) # Regression misses structure
```

```
##  
## Chi-squared test of CSR using quadrat counts  
## Pearson X2 statistic  
##  
## data:  r  
## X2 = 419.39, df = 24, p-value < 2.2e-16  
## alternative hypothesis: two.sided  
##  
## Quadrats: 5 by 5 grid of tiles
```

Obtendo estimativas não-paramétricas

Uma alternativa não-paramétrica é considerar um modelo log-linear em que

$$\log(\rho(\mathbf{s})) = \sum_{i=1}^n \theta_i \varphi_i(\mathbf{s}),$$

onde φ_i são funções conhecidas (splines, Kernels) e θ_i são constantes.

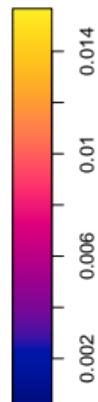
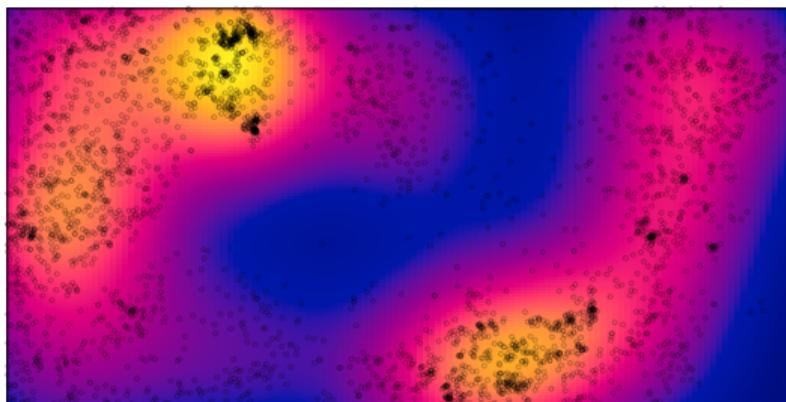
A função `density.ppp` usa Kernels, e emprega vários tipos de *edge correction*. Isto é,

$$\rho(\mathbf{s}) = \frac{1}{e_\sigma} \sum_{i=1}^n K_\sigma(x_i - \mathbf{s}),$$

em que $e_\sigma = \int_{\mathbf{s} \in W} K_\sigma(\mathbf{s}) d\mathbf{s}$ é a correção de viés de bordas.

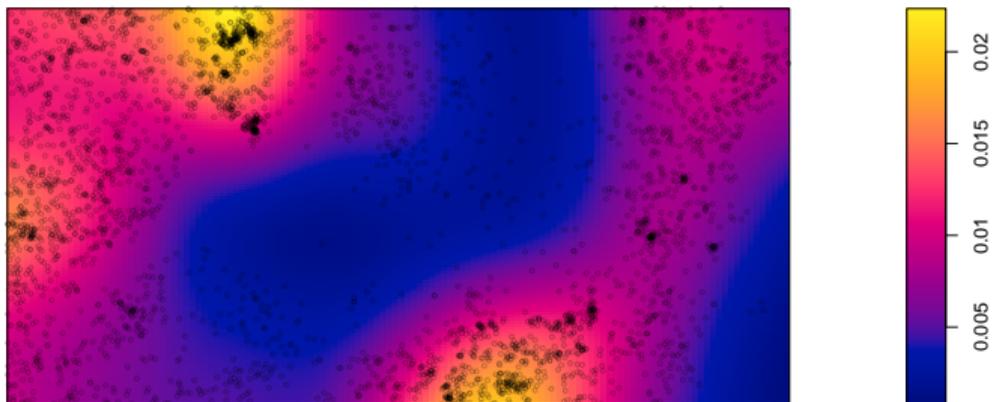
Estimativa Kernel ($\sigma = 70$, sem edge correction)

Beilschmiedia pendula (densidade)



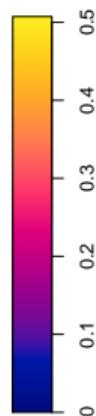
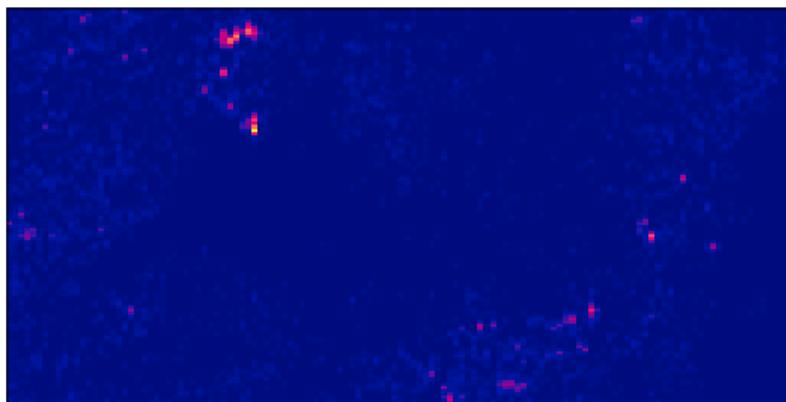
Estimativa Kernel ($\sigma = 70$, *default* edge correction)

Beilschmiedia pendula (densidade)



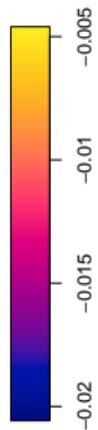
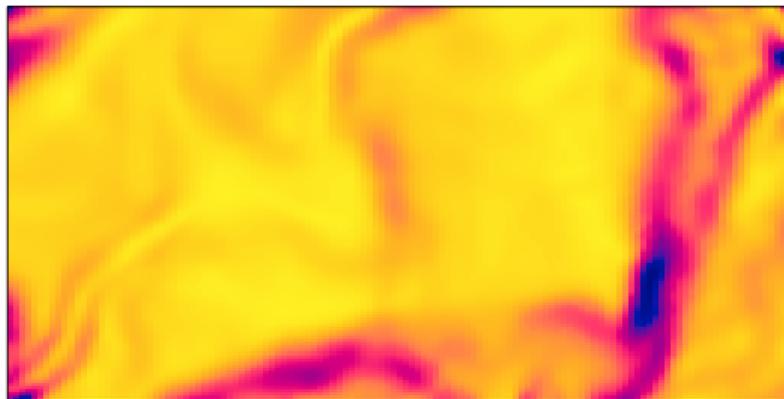
Estimativa Kernel (σ por LOO-CV, *default* edge correction, ruim)

Beilschmiedia pendula (densidade)



Resíduos de regressão $\tilde{\rho} - \hat{\rho}$

residuals(model)



Estimação não-paramétrica:

Câncer de laringe em Chorley-Ribble



Veja Barr and Schoenberg (2010), Turner (2017).

Referências I

- Baddeley, A., Rubak, E., and Turner, R. (2015). *Spatial point patterns: methodology and applications with R*. CRC Press.
- Baddeley, A., Turner, R., et al. (2005). Spatstat: an r package for analyzing spatial point patterns. *Journal of statistical software*, 12(6):1–42.
- Barr, C. D. and Schoenberg, F. P. (2010). On the voronoi estimator for the intensity of an inhomogeneous planar poisson process. *Biometrika*, 97(4):977–984.
- Berman, M. and Turner, R. (1992). Approximating point process likelihoods with GLIM. *Applied Statistics*, pages 31–38.
- Turner, R. (2017). *deldir: Delaunay Triangulation and Dirichlet (Voronoi) Tessellation*. R package version 0.1-14.