Marilia Sá Carvalho - DEMQS
Oswaldo G. Cruz - PROCC

Estudos Ecológicos

Análise de dados temporais e espaciais

Fev/2000

Estrutura do curso

Semana		2ª	3ª	4ª	5ª	6ª
1	M		S-Plus	Séries Temporais - Análise Exploratória	Modelagem ARIMA	Modelos Hierárquicos
	T	Estudos Ecológicos				
2	M	Padrão de Pontos	Geo-estatística	Áreas	Tempo-Espaço * Modelos Bayesianos	Discussão dos trabalhos
	T					

- Aulas teóricas
- Aulas práticas usando S-Plus (em micros e estações RISC)
- Aulas demonstrativas: SIG/pacote estatísticoWinBugs

Estudos Ecológicos - conceitos

- *"Um estudo ecológico ou agregado focaliza a comparação de grupos, ao invés de indivíduos. A razão subjacente para este foco é que dados a nível individual da distribuição conjunta de duas (ou talvez todas) variáveis estão faltando internamente nos grupos; neste sentido um estudo ecológico é um desenho incompleto".*

(Morgenstern, cap. Ecologic Studies - in Rothmans, Modern Epidemiology, 2ª Ed., 1998)

Estudos Ecológicos - conceitos

- *"... estudar saúde no contexto ambiental. O objetivo é ambicioso: entender como o contexto afeta a saúde de pessoas e grupo através de seleção, distribuição, interação, adaptação, e outras respostas. Medidas de atributos do indivíduo não podem dar conta destes processos* [...] *Sem medir estes contextos, nem padrão de mortalidade e morbidade, nem o espalhamento epidêmico, nem a transmissão sexual podem ser explicados"*

(Susser, Am.J.Public Health, 1994;84:825-835)

Estudos Ecológicos - conceitos

- *"Textos de Epidemiologia fazem uma avaliação consistente sobre estudos ecológicos: eles são tentativas cruas de estimar correlações em nível individual. [...] Examinar esta questão de uma perspectiva diferente - como um problema geral de validade - mostrará que a falácia ecológica, conforme freqüentemente usada, encoraja três noções interrelacionadas e falaciosas: (1) que modelos em nível individual são mais perfeitamente especificados que os de nível ecológico, (2)que correlações ecológicas são sempre substitutos para correlações de nível individual, e (3) que variáveis de nível de grupo não causam doença."*
  (Schwartz, Am.J.Public Health, 1994;84:819-824)

Estudos Ecológicos - conceitos

- *"A Epidemiologia é freqüentemente definida em termos do estudo da determinação da distribuição da doença; mas não se deve esquecer que quanto mais espalhada é uma causa particular, menos ela contribui para explicar a distribuição da doença."*
- *"...dois tipo de perguntas etiológicas. A primeira busca as causas dos casos, e a segunda as causas da incidência."*
  (Rose, G. Int.J.Epidemiol., 1985;14:32-38)

Estudos Ecológicos - conceitos

- *"Aplicada à etiologia, a visão centralizada no indivíduo leva ao uso do risco-relativo como a representação básica da força etiológica: ou seja, o risco em indivíduos expostos realtivo aos não-expostos.* [...] *Esta pode ser geralmente a melhor medida de força etiológica, mas não é medida de* [...] *importância em saúde pública."* (Rose, G. Int.J.Epidemiol., 1985;14:32-38)

Estudos Ecológicos - conceitos

- *"É rara a doença cuja taxa de incidência não tenha variado largamente, seja ao longo do tempo ou entre populações* [...] *Isto significa que as causas da incidência, desconhecidas que sejam, não são inevitáveis.* [...] *Mas identificar o agente causal pelos métodos tradicionais de caso-controle e coorte não terá sucesso se não houver sufucientes diferenças na exposição dentro da população* [...] *Nestas circunstâncias tudo os que os métodos tradicionais fazem é encontrr marcadores de susceptibilidade individual. A chave deve ser buscada nas diferenças entre populações ou em mudanças nas populações ao longo do tempo."* (Rose, G. Int.J.Epidemiol., 1985;14:32-38)

Estudos Ecológicos - conceitos

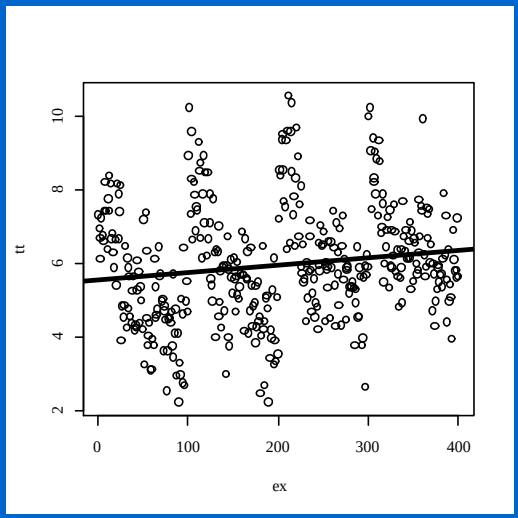
- *“ ... torna-se aparente que muitas das explicações convencionais dos determinantes da saúde - porque algumas pessoas são saudáveis e outras não - são, na melhor das hipóteses seriamente incompletas, se não simplesmente erradas. É assim, infelizmente, porque as sociedades modernas dedicam uma parte muito grande de sua riqueza, esforço e atenção tentando manter ou melhorar a saúde dos indivíduos que compõem suaa populações. Estes esforços maciços são primeiramente canalizados para os sistemas de assistência à saúde, presumivelmente refletindo uma crença que receber uma boa assistência é o mais importante determinante de saúde.”* (Evans,R.G.”Why are some people healthy and others not”)

As árvores

ou

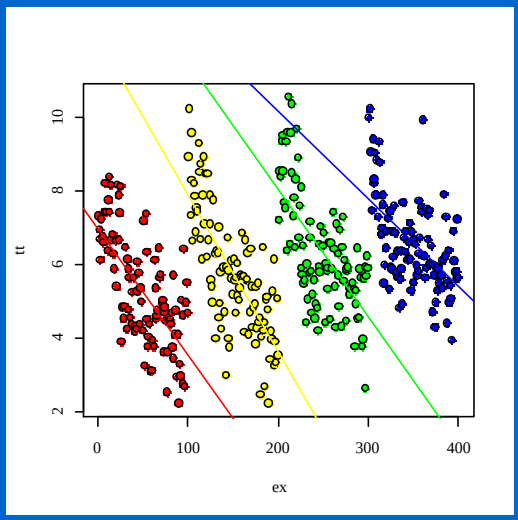
a floresta?

As Árvores



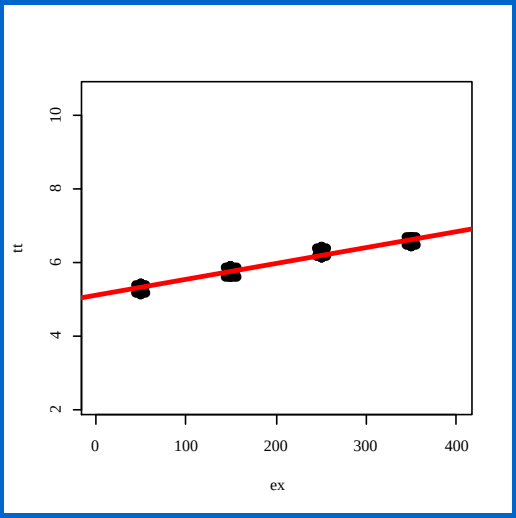
- Suponha os dados abaixo, onde a variável "X" representa um efeito de exposição e a variável "Y" um taxa.
- Ao fazermos uma regressão obtemos uma correlação de apenas 0,1469 entre as duas variáveis.

Os Bosques



- Ao estratificarmos os dados evidencia-se uma estrutura
- Ao fazer uma regressão em cada grupo obtém-se
- Vermelho r = -0,6760
- Amarelo r = -0,7469
- Verde r = - 0,6503
- Azul r = -0,5487

As Florestas



- Tirando-se a média para cada grupo iremos obter
- Quatro pontos sob os quais faremos uma regressão
- O coeficiente de correlação obtido e r = 0,9938

Medidas - conceitos

- Medidas agregadas - sumários de distribuição de observações colhidas a nível individual, usualmente proporções, médias, ou percentis da distribuição. (Ex: renda média do chefe da família; % de chefes com renda abaixo de um salário mínimo; mediana etária de homens; idade onde 95% das crianças entram na escola)
- Medidas ambientais - características físicas do meio onde vivem ou trabalham os indivíduos. Observar que para cada medida ambiental existe um análogo no nível individual (medidas de exposição ou dose) que varia entre os indivíduos do grupo (Ex: poluição do ar, intensidade de UV)
- Medidas globais - não existe análogo individual (densidade populacional; existência de leis, acesso ao serviço de saúde, etc.)

Problemas práticos

- Numerador:
  - subregistro
  - duplicidade de registros
  - georreferenciamento:
    - não localização
    - informação incorreta
    - preenchimento inadequado
  - mudança na classificação ao longo do tempo
- Denominador:
  - espaçamento do censo
  - migração
  - mudança de fronteiras (!!!!)

Problemas práticos

- Exposição:
  - pode ocorrer em diversos lugares
  - dificilmente mensurável com precisão
  - uso de “proxy”
  - diferentes áreas para medida de exposição e de efeito, e áreas não compatíveis
  - Informações mais detalhadas (PNAD, amostra do censo) não extrapoláveis para populações pequenas
- Análise:
  - migração
  - multicolinearidade

Fonte: Walter, S.D.Ecological Studies - discussion. In Int. Conf. on the Analysis and Interpretation of Disease Clusters and Ecological Studies, Londres, 16-17 de dezembro, 1999.

Séries Temporais

Análise exploratória

O que é

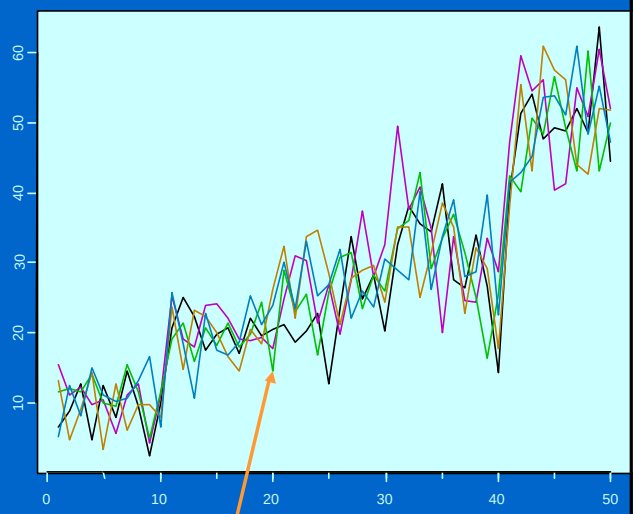
- Conjunto de observações ordenadas no tempo
- Classificação:
  - discretas:
    - a unidade de tempo é discreta, usualmente intervalos iguais (dia, semanas epidemiológicas); o mês não tem o mesmo tamanho)
    - Ex. mortalidade, notificações
  - contínuas:
    - a informação é obtida por amostragem (discretizando em intervalos iguais) ou acumulada por período
    - Ex. temperatura, pluviosidade, partículas em suspensão.

Processo estocástico

- Um processo estocástico pode ser pensado de duas formas:
  - um conjunto de possíveis trajetórias que poderiam ser observadas;
  - um conjunto de variáveis aleatórias uma para cada tempo *t*
- Cada valor observado de uma trajetória é um dos possíveis valores que poderiam ter sido observados, de acordo com a distribuição de probabilidades da respectiva variável aleatória.
- Série temporal é o conjunto de observações disponíveis para análise - uma parte de uma trajetória entra muitas que poderiam ter sido observadas

Exemplo

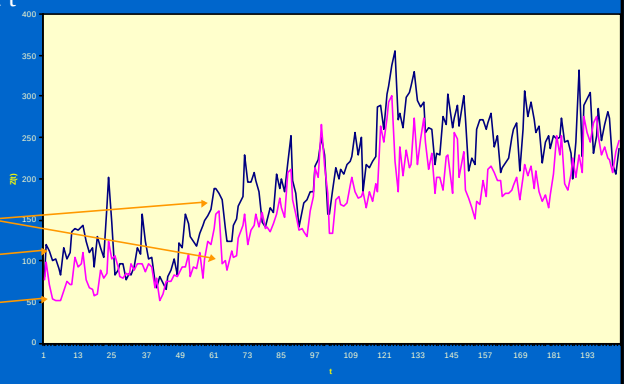
- Série com a mesma estrutura: cada série é uma possível realização do mesmo processo.



Trajetória ou série temporal ou função amostral

Notação e nomenclatura

- Utilizando o exemplo óbitos por causa por local:
  - Z(t) - óbitos no instante t



- Processo estocástico: o conjunto de todas as possíveis realizações;
- $Z^{(n)}(t)$-cada trajetória
- $Z^{(1)}(6)$= 87
- $Z^{(2)}(6)$= 52
- valor da medida de cada série no instante $t=6$

• A série pode ser:
  – multidimensional (Ex. homicídios UFs do Sudeste)
  – multivariada (Ex. homicídios e acidentes)

Objetivos: análise de séries temporais

Objetivo	Exemplo
Descrição: verificar existência de tendência, sazonalidade, ciclos. Histogramas, boxplots, são ferramentas da análise exploratória descritiva	Identificar tendência da AIDS; sazonalidade da dengue visando estabelecer melhor período de intervenção.
Estabelecimento de causalidade: estudo da relação de causa-efeito	Vacina X sarampo; Mortalidade por DIC X melhor assistência
Classificação: identificação de padrões	A série de leishmaniose tegumentar é “igual” à visceral?
Controle: sistemas dinâmicos, caracterizados por uma entrada *X(t)*, uma série de saída *Z(t)* e uma função de transferência *v(t)*	Modelar a resposta a medidas de controle de epidemia

Independência

- Os métodos usuais de análise de dados têm como pressuposto básico a independência dos eventos (casos). Ou seja, a ocorrência de um caso de doença em uma dada pessoa é independente da ocorrência em outra pessoa.
- Na análise da incidência de doenças (ou qq outro indicador ecológico) ao longo do tempo isso não é verdade: a incidência em um determinado dia/mês ou ano em geral é correlacionada com a ocorrência no dia/mês/ano anterior.
- Esta correlação é expressa em uma função denominada função de autocorrelação.

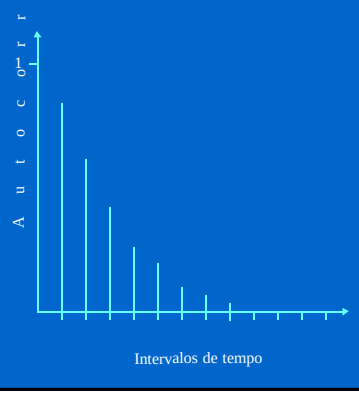
Dependência

- Classifica-se a dependência quanto à:
  - Sem dependência serial: série puramente aleatória ou ruído branco
  - Memória longa: a dependência desaparece lentamente (os valores de pontos no passado influenciam momentos muito adiante no tempo - ex, doenças com grande latência)
  - Memória curta: dependência desaparece rapidamente (doenças de alta infecciosidade e “explosivas” - gripe, por ex.)

Função de autocorrelação

- Para cada intervalo de tempo fixo *j*, pode-se calcular a correlação amostral entre os valores de Z_t e $Z_{(t+j)}$
- Para cada intervalo de tempo fixo *j*, pode-se calcular a correlação amostral entre os valores de Z_t e $Z_{(t+j)}$. O gráfico destes valores para cada *j* denomina-se correlograma.
- O correlograma é uma das principais ferramentas de análise exploratória e modelagem das séries temporais, pois indica em que medida cada valor em um dado instante de tempo *t* se relaciona com os valores em *t+1*, *t+2*,... *t+j*

$$r_j = \frac{\sum_{t=1}^{N-1}(Z_t - \bar{Z})(Z_{t+j} - \bar{Z})}{\sum_{t=1}^{N}(Z_t - \bar{Z})^2}$$



Tratamento dos dados

- Intervalo amostral - somente se detecta fenômenos cuja periodicidade é maior que o intervalo amostral (sazonalidade com dados anuais não é detectável)
- Estacionariedade:
  - 1ª ordem - média constante ao longo de todo o período
  - 2ª ordem - variância constante ao longo de todo o período
- Transformações - visam estabilizar a série
  - diferenças sucessivas $\Delta Z(t) = Z(t) - Z(t-1)$
  - estabilizar variância (log) $\Delta \log Z_t = \log Z_t - \log Z_{t-1}$
- observações perdidas ou irregulares - interpolação, etc.
- outliers - exclusão, tratamento
- registros curtos - CUIDADO!

Componentes

- A série pode ser descrita como sendo a soma dos componentes: tendência, sazonalidade, ciclicidade e termo aleatório.

$$Z_t = T_t + S_t + C_t + a_t \text{ , t} = 1,2,\ldots,\text{N}$$

- Se a sazonalidade varia em conjunto com a tendência (aumenta de amplitude quando aumenta a tendência), o modelo melhor é multiplicativo, que pode ser transformado em aditivo usando log.

$$Z_t = T_t \cdot S_t \cdot a_t$$
$$\log(Z_t) = \log(T_t) + \log(S_t) + \log(a_t)$$

- Removendo as componentes T e S, supõe-se que sobra?
  - Ruído branco;
  - cada a_t é "determinado" pelo $a_{(t-1)}$ - modelo AR
  - a variância de a_t é "determinada" por $a_{(t-1)}$ - modelo MA

Tendência e sazonalidade

- Estimar a tendência ou a sazonalidade:
  - ajustar polinômio, exponencial ou reta (paramétrico);
  - suavizar (filtros - não paramétricos);
  - diferenciar.
- Diferenças:
  - pode-se diferenciar tantas vezes quanto necessário até estabilizar (não + que duas diferenças)
  - para sazonalidade usa-se diferenciar com período igual ao da sazonalidade

$$\Delta Z(t) = Z(t) - Z(t-s) \text{ , s é o período da sazonalidade}$$

  - não permite previsão da tendência ou sazonalidade, as retira

Alisamentos

- Médias móveis:
  - o tamanho da janela é empírico
  - perde-se *k* observações
  - para estudar sazonalidade usa-de janela de ordem *s* (o período da sazonalidade)
  - se k = N/2, então a previsão será igual a média aritmética dos valores observados, o que é o valor esperado para séries inteiramente aleatórias
  - não pode ser usado para previsão se a série é não estacionária

$$Z_t^* = \frac{1}{2k+1} \sum_{j=-k}^{k} Z_{t+j} \text{ , ordem } 2k$$

Séries Temporais

Modelo Box & Jenkins
(ARIMA)

O que é

- Método de modelagem de séries temporais tratando simultaneamente tendência, sazonalidade, ciclicidade e estrutura de dependência serial.
- A dependência serial é influência que um dado evento no tempo recebe de pontos anteriores.
- O processo de modelagem é feito em um ciclo iterativo de 3 estágios (repetido até ...):
  - identificação - análise exploratória, baseada em gráficos (dos dados brutos, autocorrelação, autocorrelação parcial), buscando identificar o tipo de modelo + adequado
  - estimação - estimativa de termos e parâmetros e seleção do "melhor modelo"
  - diagnóstico - critérios de ajuste, parcimônia

Alguns processos estocásticos

- Processo aleatório:
  - seqüência de variáveis aleatórias (a_t)que são mutuamente independentes e identicamente distribuídas. Possui média e variância constantes e os coeficientes de correlação são iguais a:

$$r_k = \begin{cases} 1 \text{ , se } k = 0 \\ 0 \text{ , se } k = \pm 1, \pm 2, \ldots \end{cases}$$

logo $_t$ é estacionário

- Passeio aleatório (*random walk*):
  - Denomina-se passeio aleatório quando a variável aleatória Z_t é igual à Z_{t-1}mais um erro aleatório $Z_t = Z_{t-1} + a_t$
  - quando $t = 0 \rightarrow Z_1 = a_1$ logo, $Z_t = \sum_{i=1}^{t} a_i$

Modelo AutoRegressivo - AR(p)

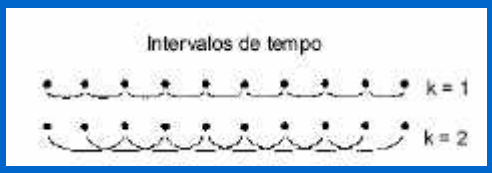
- Supondo que a variável aleatória Z_t é linearmente correlacionada com seus próprios valores defasados, este é um modelo autoregressivo geral de ordem p

$$Z_t = C + f_1 Z_{t-1} + f_2 Z_{t-2} + \ldots + f_p Z_{t-p} + a_t, \quad t = 1,2,..,p$$

- O objetivo é estimar:
  - a constante C - média do processo ou intercepto
  - a ordem p do modelo - até onde vai a dependência
  - os parâmetros f de cada termo - peso de cada ponto passado na determinação do ponto i
- Para estimar os parâmetros f de um AR a estacionariedade de 1ª e 2ª ordens é fundamental!!!

Função de Autocorrelação - ACF

- Para cada intervalo de tempo fixo k, pode-se calcular a correlação amostral entre os valores de Z_t e $Z_{(t+k)}$



- Para cada intervalo de tempo fixo k, pode-se calcular a correlação amostral entre os valores de Z_t e $Z_{(t+k)}$. O gráfico destes valores para cada k denomina-se correlograma.

$$r_k = \frac{\sum_{t=1}^{N-1} (Z_t - \bar{Z})(Z_{t+k} - \bar{Z})}{\sum_{t=1}^{N} (Z_t - \bar{Z})^2}$$

Autocorrelação Parcial - PACF

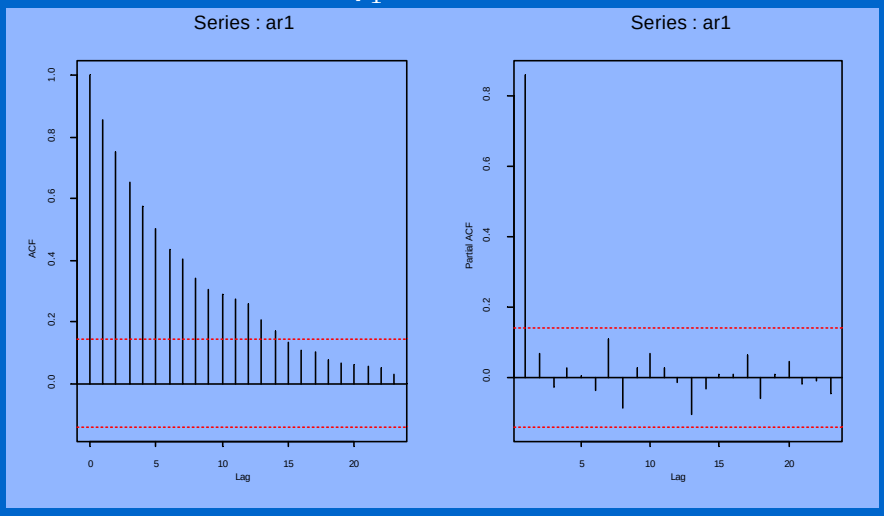
- A correlação medida diretamente em *t*-1, *t*-2 até *t-p* é a função de autocorrelação.
- Outra função que pode ser calculada é a função de autocorrelação parcial, onde o cálculo da autocorrelação entre os pontos é feito excluindo o efeito dos pontos intermediários.
- No lag = 1, a ACF e a PACF são iguais.
- Na PACF somente existe correlação até o lag igual a ordem do modelo - modelo de ordem 3 somente apresenta valores de PACF até o 3º lag.

Condições de estacionariedade

- Uma série é estacionária quando suas propriedade não variam ao longo do tempo. Em um processo AR, a estacionariedade se reflete na estimação dos parâmetros:
  - AR de ordem 1: $|f_1| < 1$
  - AR de ordem 2: $|f_2| < 1$, $f_2 + f_1 < 1$, $f_2 - f_1 < 1$

Exemplos

- AR de ordem 1, com $f_1 = 0{,}8$



Modelo de Médias Móveis - MA(q)

- Independente do processo autoregressivo, cada elemento da série pode também ser afetado pelo erro passado - processo "Médias Móveis"
- Neste caso, o valor de Z correlaciona-se aos valores do componente aleatório de pontos anteriores (usa-se a denominação *choque aleatório*).

$$Z_t = C - q_1 a_{t-1} - q_2 a_{t-2} - \ldots - q_p a_{t-q} + a_t$$

Por convenção os termos em a são escritos com sinais negativos

- Cada observação é a soma de um componente aleatório *a* e uma combinação dos componentes aleatórios anteriores.

Invertibilidade

- Através de substituições sucessivas pode-se transformar um AR de ordem p em um MA de ordem infinita: MA(∞)

$$Z_t = f_1 Z_{t-1} + a_t \qquad (1)$$
$$Z_{t-1} = f_1 Z_{t-2} + a_{t-1} \qquad (2)$$
$$Z_{t-2} = f_1 Z_{t-3} + a_{t-2} \qquad (3)$$
$$\vdots$$

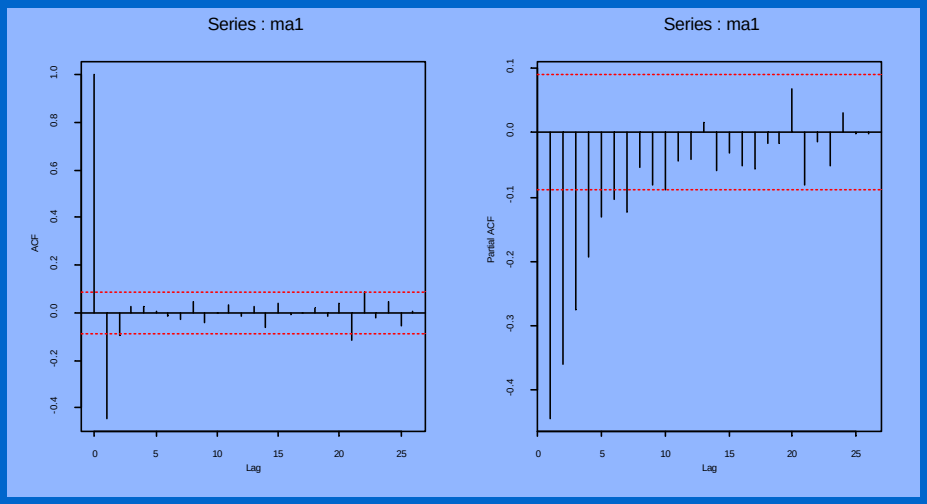
- Substituindo (2) em (1) e depois (3) em (1) e assim sucessivamente, teremos:
- $Z_t = a_t + q_1 a_{t-1} + q_2^2 a_{t-2} + \ldots$ ➔ MA(∞)

Condições de invertibilidade

- No modelo MA não há restrição sobre os f_q para que o processo seja estacionário, mas é necessário garantir a invertibilidade.
- Existe uma dualidade entre processos de médias móveis e autoregressivo, onde a equação de MA pode ser reescrita na forma AR (de ordem infinita). Para isso algumas condições devem ser satisfeitas:
- MA(1) $|q_1| < 1$ MA(2) $|q_2| < 1$, $q_2 + q_1 < 1$, $q_2 - q_1 < 1$

Exemplo MA

- MA de ordem 1, $q = 0{,}8$



Modelo ARMA(p,q)

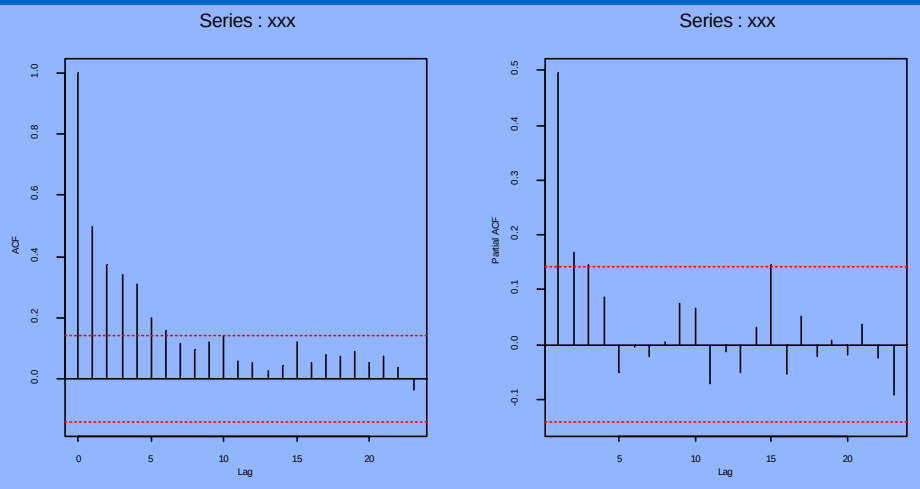
- A importância de um modelo ARMA está no fato de eu uma série estacionária pode ser descrito por um modelo ARMA que envolve menos parâmetros que um MA ou AR puro.

$$Z_t = f_1 Z_{t-1} + f_2 Z_{t-2} + \ldots + f_p Z_{t-p} + a_t - q_1 a_{t-1} - q_2 a_{t-2} - \ldots - q_p a_{t-q}$$

- Cada observação é definida por combinação linear de observações anteriores e combinação de componentes aleatórios anteriores.
- Neste modelo misto, as duas condições - estacionariedade e invertibilidade - são necessárias

Exemplo

- ARMA(1,1), parâmetros: $f = 0{,}8$ $q = 0{,}4$



Modelo ARIMA(p,d,q)

- Para série não estacionária é necessário utilizar o modelo ARIMA - *AutoRegressive Integrated Moving Average*.
- Neste modelo se utiliza o método de diferenças para obter a estacionariedade da série:

$$W_t = \nabla Z_t = (1-B)Z_t = Z_t - Z_{t-1}$$

operador de deslocamento (*backshift*)

- O modelo então passa a ser:

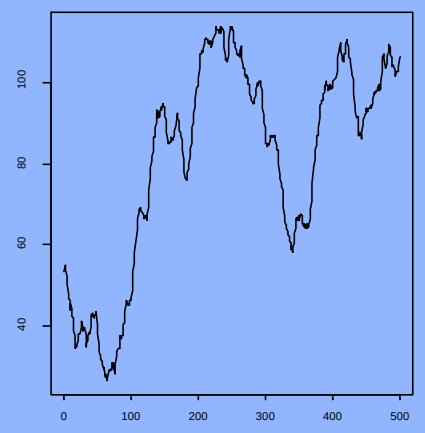
$$W_t = f_1 W_{t-1} + \ldots + f_p W_{t-p} + a_t - q_1 a_{t-1} - \ldots - q_q a_{t-q}$$

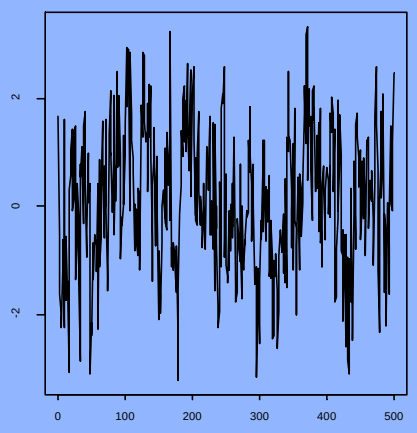
$$f(B)W_t = q(B)a_t$$

$$f(B)(1-B)^d Z_t = q(B)a_t$$

Exemplo

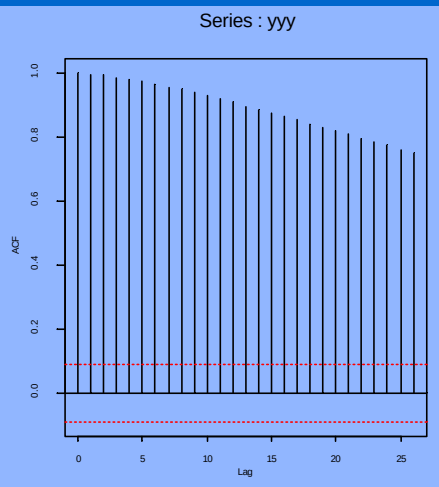
- Série não estacionária antes e após diferenciação - d(1)

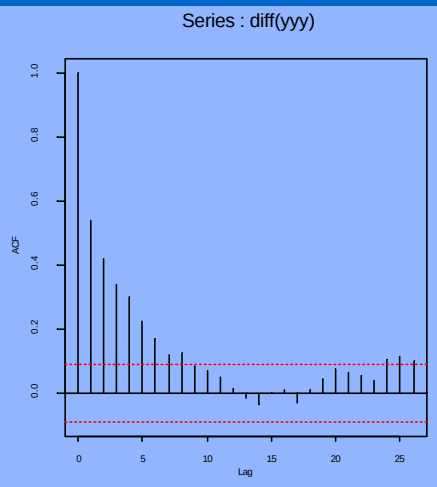




Exemplo

- ACF antes e após diferenciação





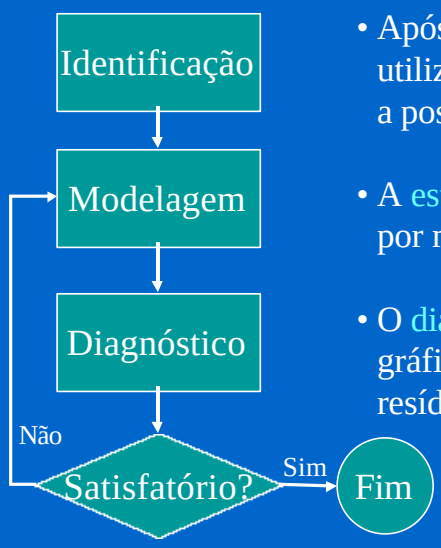
Modelos sazonais - SARIMA

- Em epidemiologia é comum haver sazonalidade. O seja, considerando medidas mensais, pode-se esperar que a série dependa também dos termos Z_{t-12} e talvez Z_{t-24}: SARIMA(p,d,q)x(P,D,Q).

$$f(B)\Phi(B^s)\nabla_s^D\nabla^d Z_t = C + q(B)\Theta(B^s)a_t$$

AR(p) *backshift* AR(P) - sazonal *backshift* sazonal diferenciação sazonal diferenciação tendência Z_t	=	Média do processo MA(q) *backshift* MA(Q) - sazonal *backshift* sazonal erro aleatório

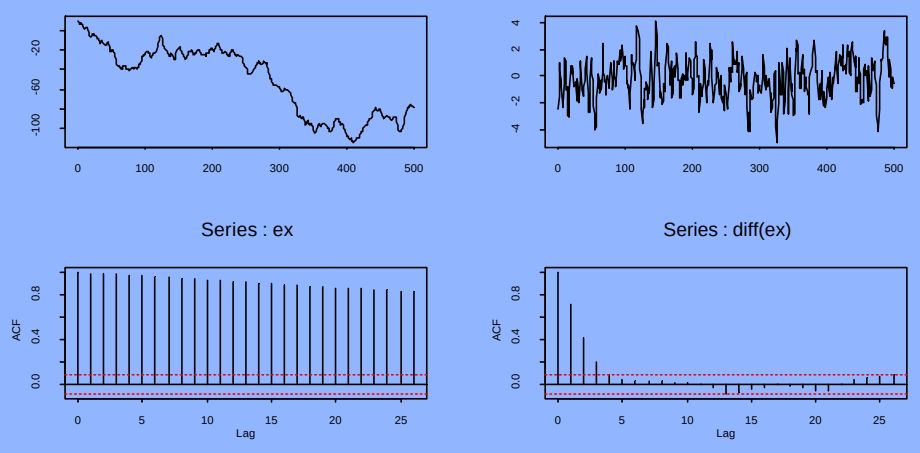
Processo de Modelagem



- Após remover tendência (se houver), utiliza-se ACF e PACF para identificar a possível ordem do modelo.
- A estimativa dos parâmetros é feita por máxima verossimilhança
- O diagnóstico é feito através dos gráficos de resíduos, ACF dos resíduos, e outros indicadores de ajuste
  - O indicador de ajuste mais usado é o AIC (*Akaike Information Criteria*) que colabora na escolha do melhor modelo

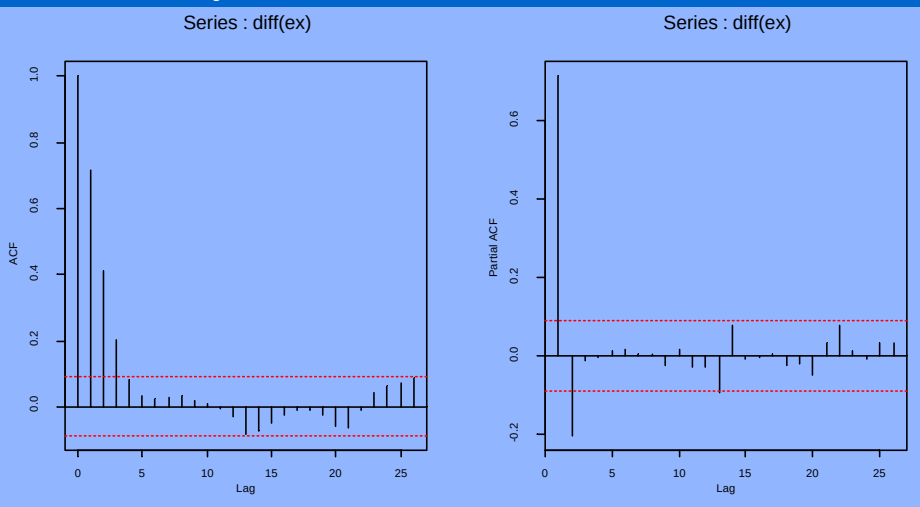
Exemplo 1 - tendência

- Tendência - identificação e remoção



Exemplo 1 - identificação

- Identificação da ordem do modelo

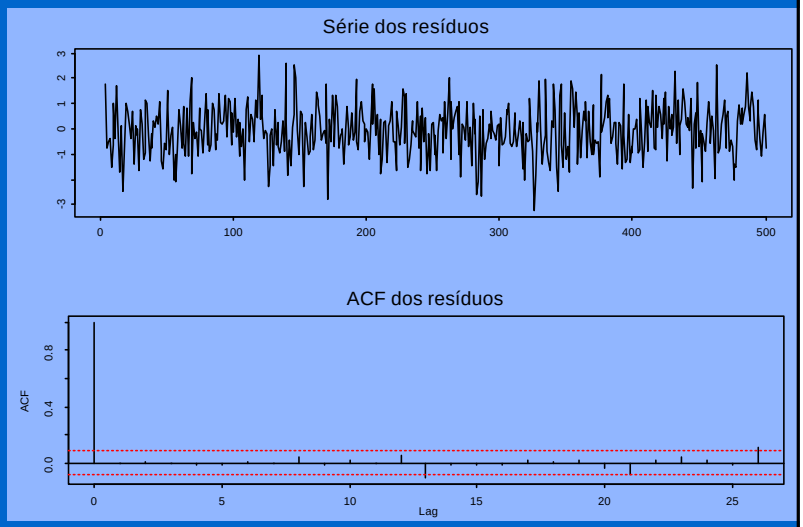


Exemplo 1 - modelagem 1

- Modelo de ARIMA(2,1,0) - (máxima verossimilhança)

f_1 = 0,87
f_2 = -0,20

AIC = 1410.81
loglike = 1406.81



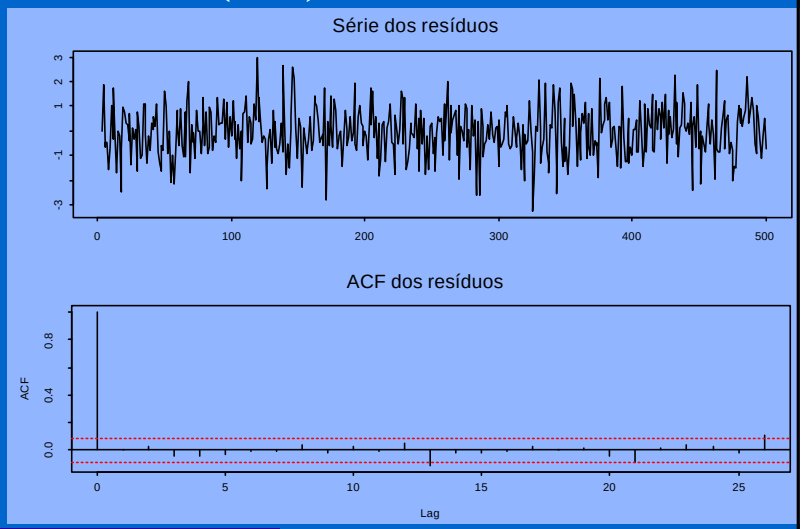
Exemplo 1 - modelagem 2

- Modelo de ARIMA(1,1,1)

f_1 = 0,61
q_2 = -0.25

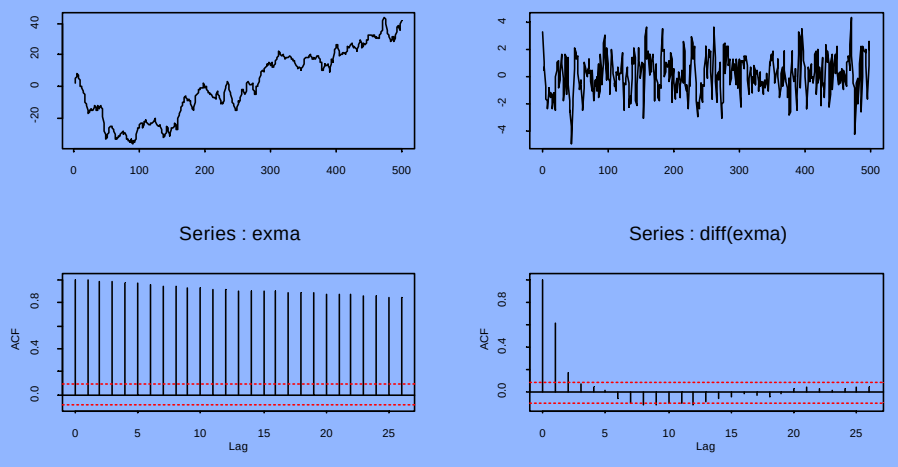
AIC = 1415.00
loglike = 1411.00

f_1 = 0,55
q_2 = -0.3



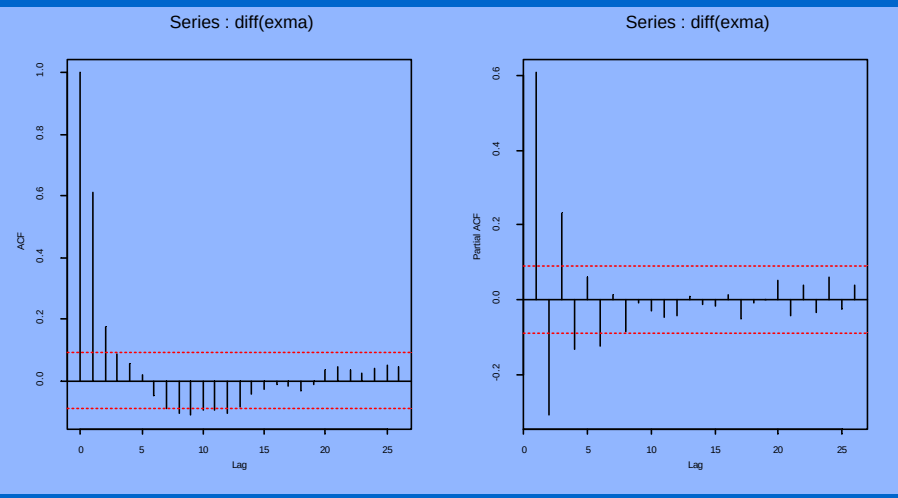
Exemplo 2 - tendência

- Tendência



Exemplo 2 - identificação

- Identificação da ordem do modelo

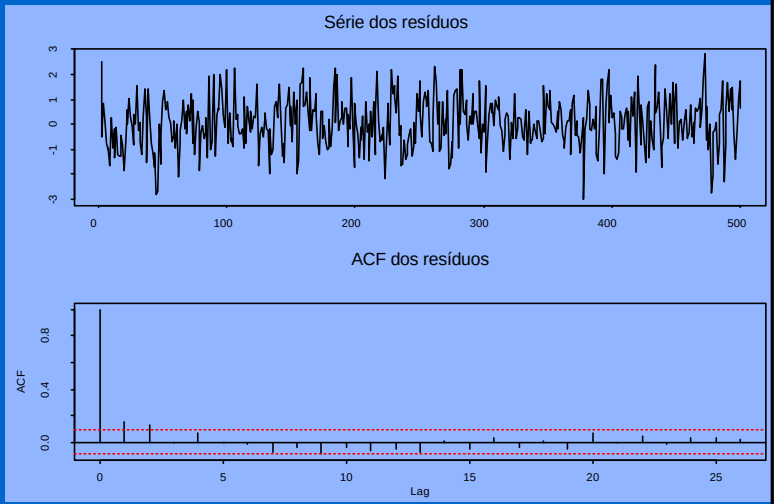


Exemplo 2 - modelagem

- Modelo de ARIMA(0,1,1)

q_1 = -0.80

AIC = 1419.63
loglike = 1417.63



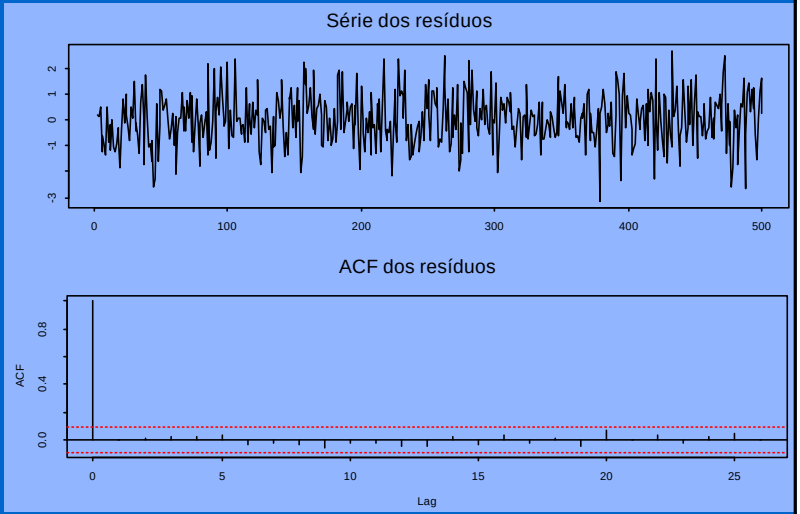
Exemplo 2 - modelagem 2

- Modelo de ARIMA(1,1,1)

f_1 = 0,25
q_2 = -0.70

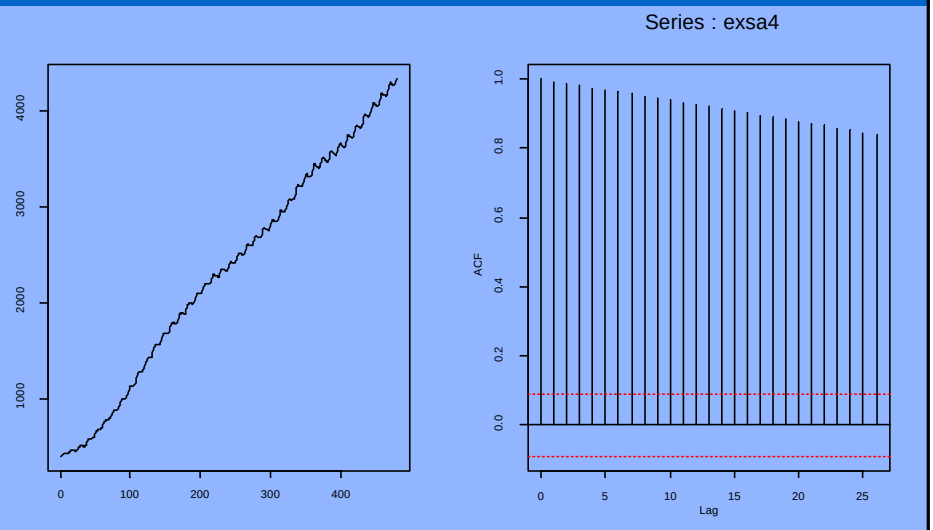
AIC = 1391.80
loglike = 1387.80

f_1 = 0,3
q_2 = -0.6



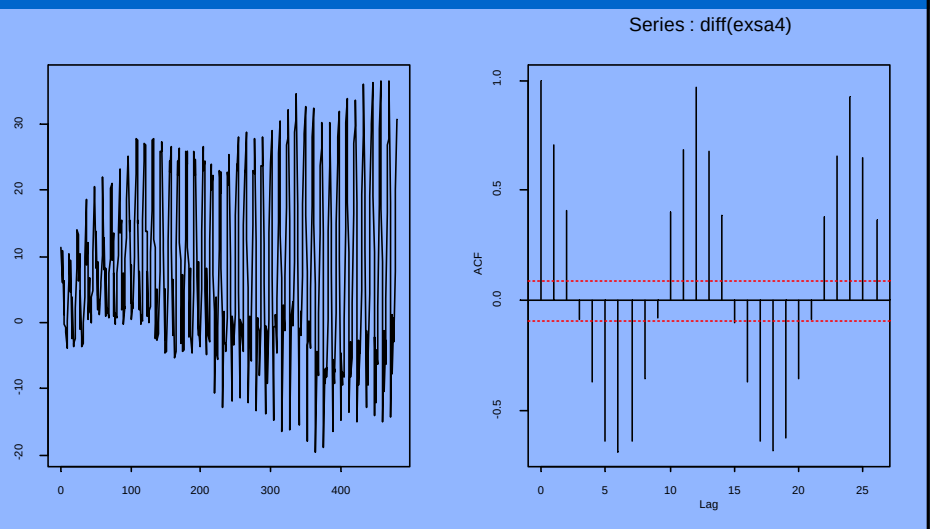
Exemplo 3 - tendência

- Tendência



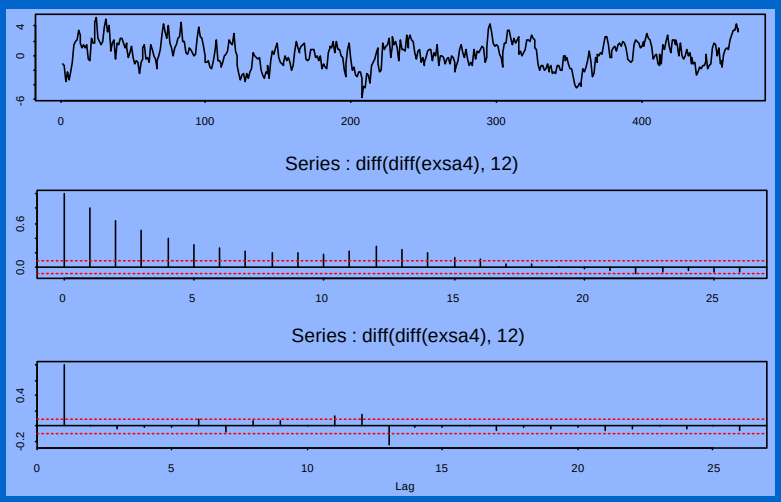
Exemplo 3 - diferenciando

- Removendo tendência - diff=1



Exemplo 3 - identificação

- Identificação da ordem do modelo - dupla diferenciação (1 e 12)

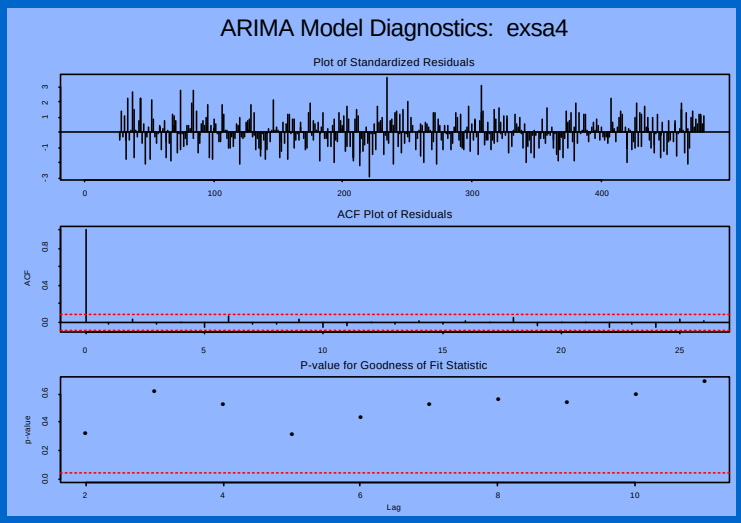


Exemplo 3 - modelagem

- Modelo de ARIMA(1,1,0)*(1,1,0)$_{12}$

$f_1 = 0{,}80$
$F_1 = 0{,}32$
$\nabla \nabla^{12}$
AIC = 1299,492
loglike = 1295,492

$f_1 = 0{,}8$
$q_2 = 0$
$\Phi = 0{,}4$
$\Theta = 0$



Modelo de regressão em ST

- Modelo usual de regressão: $Y_t = C + u_0 X_t + N_t$
- Em séries temporais, pode-se usá-los, porém com as seguintes características:
  - relação entre X e Y apresenta defasagem temporal;
  - entre X e Y existe *feedback*
  - o processo X é ARIMA, e cada ponto NÃO pode ser tomado com independente dos anteriores
  - o processo de X e de Y são correlacionados apenas porque apresentam estrutura temporal semelhante, sem valor explicativo
- Como modelar?

Correlação cruzada

$$Y_t = C + u_0(B) X_t + N_t$$

- Ou seja, modela-se a série Y usando os pontos anteriores de X, exatamente como na modelagem anterior
- Analisar:
  - o atraso
  - *feedback*
  - correlação dos resíduos, removida a estrutura temporal

Outras aplicações

- Análise de intervenção - vários tipos de intervenção existem e pode-se modelar utilizando o ARIMA:
  - pulso - um evento que aparece e desaparece instantaneamente
    - o impacto na violência de "operação" policial em uma favela
    - uma catástrofe natural e as conseqüências nas saúde
  - degrau - um evento que sobe repentinamente e se mantém alto durante um tempo
    - o aumento na cobertura vacinal através de campanhas e seu impacto na incidência da doença
- Predição (*forecast*) - muito usado em econometria. O intervalo de confiança aumenta rapidamente!
- Detecção de *outliers* - útil na detecção epidêmica

Vantagens e desvantagens

VANTAGENS

- Conceitualmente sólido
- O método de estimativa dos parâmetros permite calcular o erro associado e estabelecer intervalos de confiança
- Permite estabelecer relações causais considerando o tempo

DESVANTAGENS

- Modelar requer MUITA experiência
- A previsão em epidemiologia não teve sua utilidade demonstrada
- A escala usual dos SIS mistura processo temporais diferentes - ESPAÇO/TEMPO!

Análise de Dados Espaciais em Saúde

Conceitos Gerais

Há muitos e muitos anos atrás...

Snow, John.

Localização dos casos de cólera e bombas d'água. Londres, 1854

Fonte: *Gilbert, E.W.* **Geographical Journal**, *124:172:183*

O que é análise espacial

- Análise espacial: estudo quantitativo de fenômenos que são localizados no espaço.
- Análise de dados espaciais: em oposição a análise de dados em geral, focaliza-se as técnicas onde de considera explicitamente a localização espacial.
- Defini-se "*análise estatística espacial quando os dados são espacialmente localizados e se considera explicitamente a possível importância de seu arranjo espacial na análise ou interpretação dos resultados*" (Bailey & Gatrell, 1995).
- Neste curso serão abordadas basicamente as técnicas **estatísticas** de análise espacial.
- Diversas operações realizadas em um SIG são também chamadas análise espacial, mas não são objeto deste curso.

Quando usar

- quando o evento em estudo é gerado por fatores ambientais de difícil detecção a nível do indivíduo;
- na delimitação de áreas homogêneas segundo intervenção pretendida;
- quando o evento em estudo e os fatores relacionados têm distribuição espacialmente condicionada;
- no estudo de trajetórias entre localidades.

OBS: O conjunto de técnicas de otimização, análise de redes, rotas não serão abordados neste curso

Dependência espacial

- Quais as possíveis implicações de não considerar a localização espacial na modelagem?
- *"Independência é um pressuposto muito conveniente que faz grande parte da teoria estatística matemática tratável. Entretanto, modelos que envolvem dependência estatística são freqüentemente mais realísticos. Duas classes de modelos que tem sido comumente usados envolvem estruturas de correlação intraclasse e estruturas de correlação serial. Estes oferecem pouca aplicabilidade a dados espaciais, onde a dependência está presente em todas as direções e fica mais fraca a medida em que aumenta a dispersão na localização dos dados."* (Cressie,1991)
- O que aconteceria ao se estimar a associação entre duas variáveis em um estudo ecológico ao não se considerar, por exemplo, a tendência espacial que ambas apresentassem????

Desenho do estudo

- Sensibilidade, especificidade e precisão
- Proporção entre medidas em mapa e medidas reais

	escala	↗
Resolução	capacidade de distinguir pontos adjacentes	↗
Homogeneidade	característica da distribuição estatística	↗
Estabilidade	presença de flutuação aleatória	↘
Dado	disponibilidade	↘

Aplicações - uma revisão recente

- Mapeamento de doenças: descrição do processo espacial de distribuição das doenças, visando vigilância, predição de epidemias, etc.
- Estudos ecológicos: estudar a relação entre incidência de doenças e potenciais fatores etiológicos, seja no campo da análise exploratória visando definir hipóteses (formulação clássica), ou apontar medidas preventivas.
- Cluster: identificação de focos de doença ou avaliação de aumento de risco ao redor de fonte suspeita de risco ambiental.
- Avaliação e monitoramento ambiental: estimativa e monitoramento da distribuição espacial de fatores ambientais relevantes para a saúde.

Tipos de dados

- Dados de casos (eventos) - processos pontuais: usualmente se dispõe da localização pontual (coordenadas) da residência de casos de doença ou de controles da população de risco. Covariáveis do indivíduo podem ser medidas.
- Dados de amostras pontuais - geoestatística: medidas, em geral de natureza ambiental, tomadas em locais amostrados.
- Áreas - pode-se subdividir em dois sub-grupos:
  - Áreas irregulares - em geral contagens de casos ou populações em divisões administrativas, indicadores socioeconômicos
  - Áreas regulares - medidas em grade regular, como nas imagens de satélite

Tipos de dados

- Três tipos básicos de dados:
  - pontos
  - espaço contínuo
  - áreas.
- Eventualmente misturas de diferentes tipos estão presentes em um mesmo estudo.
- Alguns métodos somente são aplicáveis a um tipo de dado, outros a mais de um.
- Em algumas situações pode-se converter o dado de uma para outro tipo

Mapeamento de doenças

- O objetivo geral é avaliar a variação geográfica na ocorrência das doenças visando identificar diferenciais de risco, orientar a alocação de recursos, levantar hipóteses etiológicas.
- Os métodos tem como objetivo produzir um mapa “limpo”, sem o “ruído” gerado pela flutuação aleatória dos pequenos números, e controlando as diferenças na estrutura demográfica.
- São usualmente aplicados aos dados resultantes de contagens de casos em áreas administrativas - taxas.
- Também são aplicados a dados pontuais, usualmente trabalhados sob forma de superfícies de risco, ou de risco relativo.

Estudos ecológicos

- Essencialmente modelos de regressão, onde se busca explicar a variação na incidência da doença através de outras variáveis.
- O modelo se complica pela necessidade de controlar simultaneamente o processo espacial.
- Classicamente aplica-se a dados agregados em áreas.
- Pode-se entretanto considerar também dados pontuais e misturas de diferentes tipos de dados.

Cluster

- “Cluster”: qualquer agregado de eventos.
- Cluster em estatística multivariada é um resultado de classificação onde se busca definir um grupamento de “semelhantes”.
- Cluster espacial é um agregado de eventos no espaço ou a ocorrência de “taxas semelhantes” em área próximas.
- O objetivo da detecção de cluster espacial é estabelecer a significância de um sobre-risco em um determinado espaço ou tempo e espaço.

Cluster (2)

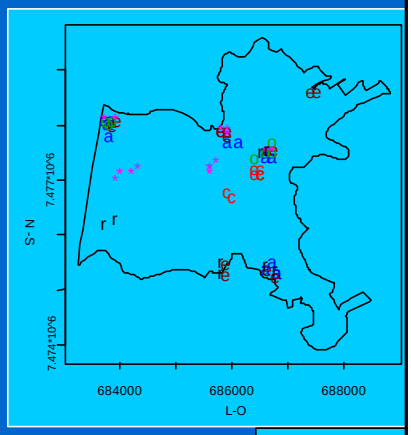
- Estes cluster podem ser causados por diferentes fatores: agentes infecciosos, contaminação ambiental localizada, efeitos colaterais de tratamentos, etc.
- Os estudos podem estar dirigidos a buscar evidência de tendência geral à clusterização, ou a um determinado e predefinido agregado.
- Podem ser usados para pontos ou áreas.
- É indispensável controlar para fatores como a distribuição populacional e outras covariáveis que podem criar agregados.

Monitoramento ambiental

- Acompanhamento de potenciais fontes ambientais de problemas de saúde: poluentes químicos, insolação (Raios UV), vegetação, clima, etc.
- Os modelos estatísticos tem por objetivos fazer a predição espacial ou espaço-temporal. Estes processos em geral tem forte correlação espacial e temporal
- O interesse pode estar voltado para predição de valores extremos.
- A quantidade e disponibilidade de dados nesta área vem crescendo, com ênfase particular para as imagens de satélite, com resolução e acessibilidade cada vez maior.

Análise espacial - análise exploratória

- descrição dos dados de forma a contribuir para o desenvolvimento de hipóteses e modelos;
- caracterizam-se por poucos pressupostos a priori e são resistentes a valores aberrantes (técnicas robustas);
- envolvem, além da visualização, alguma manipulação dos dados, sendo difícil estabelecer a fronteira entre visualização, análise exploratória e modelagem.
- gráficos dinâmicos - selecionar no mapa e analisar estatística, identificar outliers no mapa
- Ex: seleção de sub-regiões, análise de vizinhança.



a = analfabetismo
e = escolaridade
r = renda
o = água
c = c.alugadas
* = esgoto

Conceitos estatísticos fundamentais

- Estacionariedade
  - as propriedades estatísticas da variável independem de sua localização absoluta, ou seja, a média e a variância são constantes em qualquer sub-área e a covariância entre dois pontos quaisquer depende somente de sua localização relativa;
- Isotropia
  - se, além de estacionário, a covariância depende somente da distância entre os pontos e não da direção entre eles.
- Processo de modelagem
  - Transformações visando obtenção de estacionariedade;
  - Ajuste de modelos.

Análise de dados pontuais

(*point pattern*)

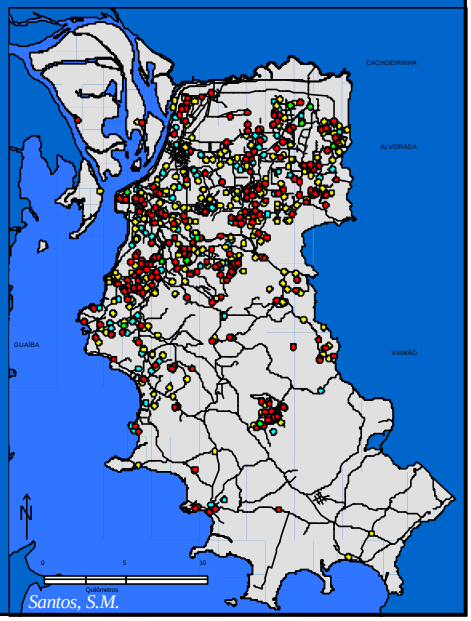
Introdução

- A análise de padrão de pontos, é o tipo mais simples de análise de dados espaciais. Baseia-se na localização dos eventos em determinada área a partir das coordenadas. O objetivo é estudar a disposição espacial dos pontos, a partir de suas coordenadas;
- O modelo básico do banco de dados neste tipo de análise é:

Evento	Coord X	Coord Y
1	4,30	2,45
2	5,39	3,35
3	4,10	3,50

Análise exploratória - mapa de pontos

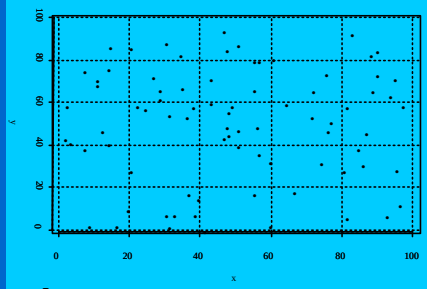
- O tipo mais simples de apresentação de dados espaciais
- Permite comparar a posição relativa dos eventos, inclusive diversos tipos
- Muito usado para localização de prédios, como centros de saúde, escolas, etc.

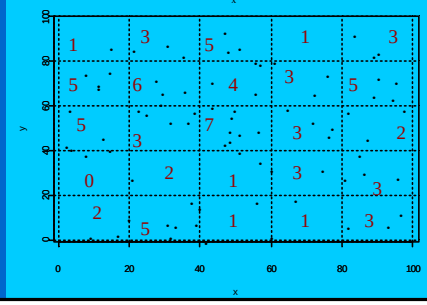


Análise exploratória - quadrat

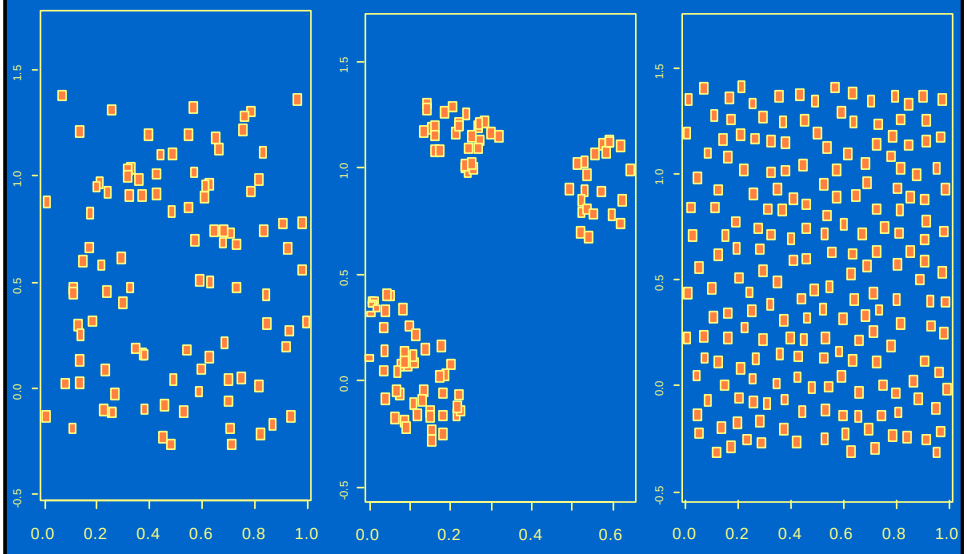
- quadrat: transforma o dado em contagem de pontos por área

Área do Quadrat **Grande ® Pequena**	
Áreas em branco	↗
Total de pontos por área	↘
Resolução	↗
Estabilidade	↘

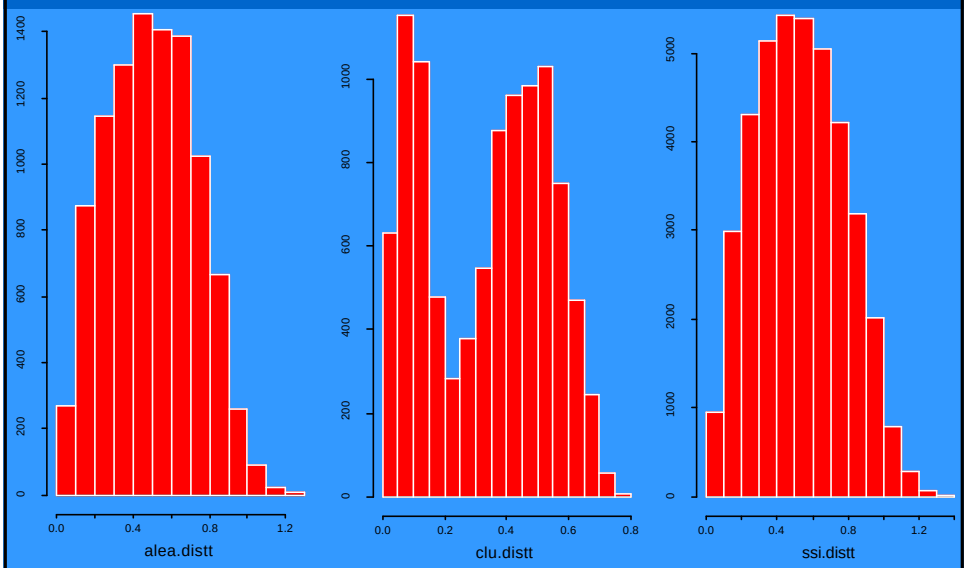




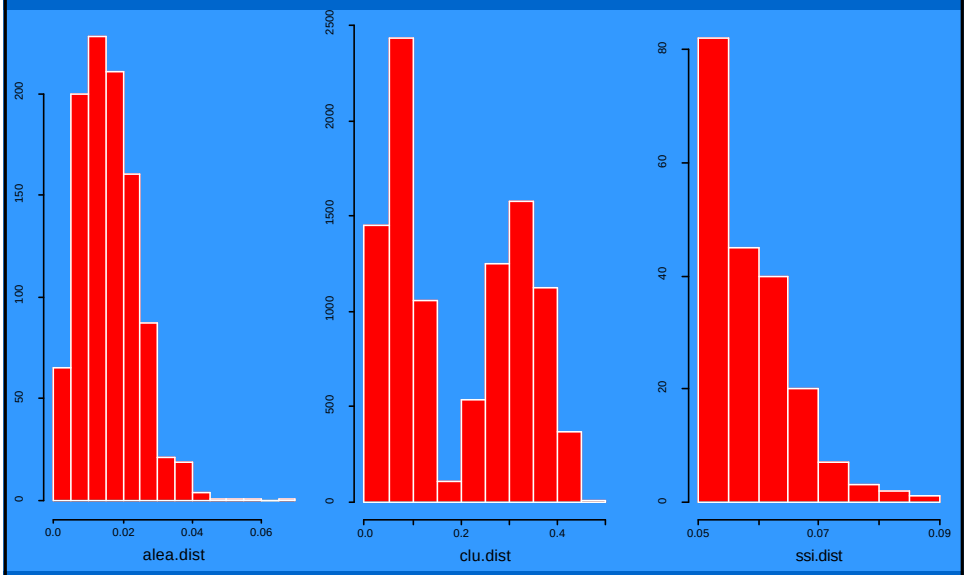
Padrões



Distribuição das distâncias - total

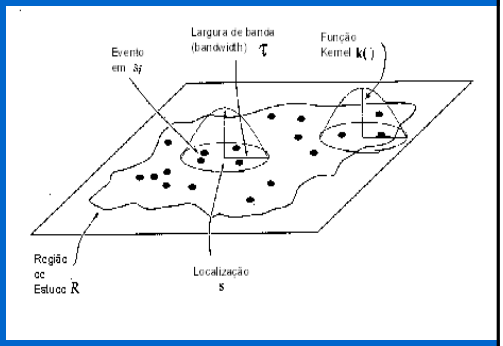


Distribuição das distâncias - 1º vizinho



Kernel

- Técnica de alisamento que utiliza janela móvel e função que dá a cada área um peso variável conforme a distância.
- Foi desenvolvida originalmente para obter uma estimação alisada da densidade de probabilidade uni ou multivariada, ou um histograma alisado.
- Estimar a intensidade de pontos dispostos no espaço é semelhante a estimar uma densidade de probabilidade bivariada.



Kernel

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

l(s) - valor estimado por área;
t - largura da banda (fator de alisamento);
k ( ) - função de ponderação **kernel**;
s - centro da área; s_i - local do ponto.

- Deve-se fazer correção para as bordas
- Calcula-se o volume sob o Kernel que está de fato dentro da região de estudo

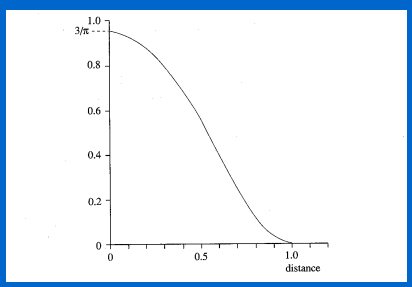
$$\delta_\tau(s) = \int_R \frac{1}{\tau^2} k\left(\frac{(s - \boldsymbol{u})}{\tau}\right) d\boldsymbol{u}$$

- Aplicando a correção das bordas obtém-se um estimador corrigido

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

Kernel

- A função de alisamento escolhida - Kernel - deve ser simétrica à origem
- Ex: Kernel quártico

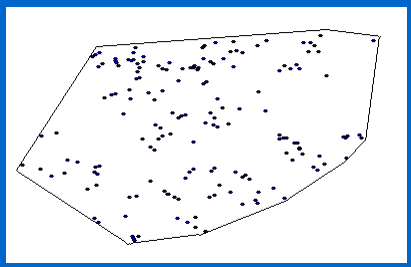


$$k(u) = \begin{cases} \frac{3}{\pi}\left(1 - u^T u\right), \text{para } u^T u \le 1 \\ 0 \quad , \text{caso contrário} \end{cases}$$

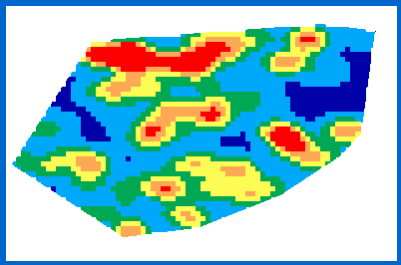
$$\hat{\lambda}(s) = \sum_{k_i \le \tau} \frac{1}{\pi \tau^2}\left(1 - \frac{h_i^2}{\tau}\right)$$

Kernel

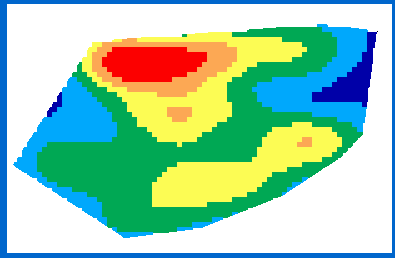
Agressões por adolescentes em Cardiff



BAILEY & GATRELL, 1995

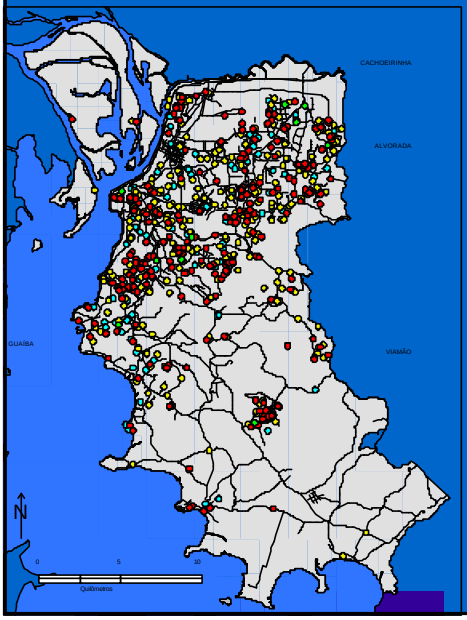


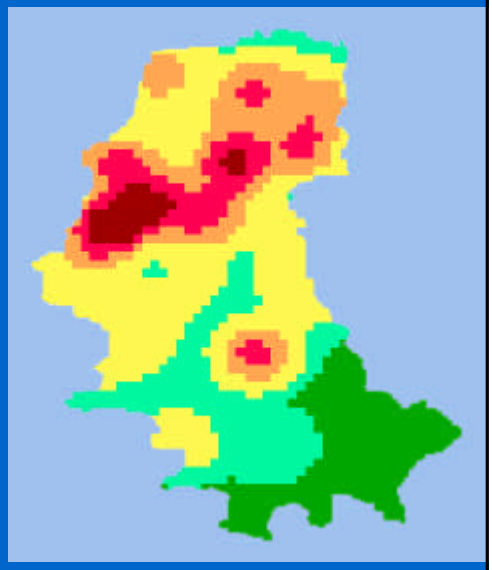
Largura da banda



Largura da banda

Causas Externas - Porto Alegre





Santos, S.M.

Vizinho mais próximo

- Kernel e quadrat permitem explorar a variação da média do processo na região de estudo - propriedade de primeira ordem
- Para investigar propriedade de segunda ordem é necessário observar as distâncias entre os eventos
- Dois tipos de distâncias: *evento-evento* (W) e *ponto aleatório-evento* (X)
- O resultado desta função empírica é o histograma das distâncias para o vizinho mais próximo - cada classe do histograma é uma contagem de eventos que ocorrem até aquela distância

$$\hat{F}(x) = \frac{\#(x_i \le x)}{m}$$

$$\hat{G}(w) = \frac{\#(w_i \le w)}{n}$$

W - *evento-evento*
X - *ponto-evento*
- contagem de pontos onde a condição acontece
n - total de eventos
m - total de pontos aleatórios

Função K

- A funções anteriores somente permitem analisar a distribuição do vizinho mais próximo - pequena escala
- A função K permite analisar as propriedades de segunda ordem de um processo isotrópico

$$\lambda K(h) = E(\#)$$

- é o número de eventos esperados até distância h
λ - intensidade ou média de eventos por unidade de área

Sendo:

λR - nº esperado de eventos na área R
$\lambda^2 R\, K(h)$ - nº pares ordenados até a distância h (por isso entra duas vezes)
$d_{i,j}$ a distância entre os pares i e j

Empiricamente é possível obter a função K.

Função K - estimativa

A função K(h) é, para cada distância h, o somatório do total de pares cuja distância é menor de que h, vezes o inverso do total de pares ordenados existente na região R.

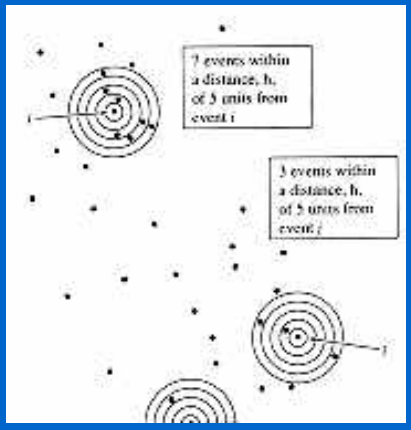
$$\hat{K}(h) = \frac{1}{\lambda^2 R} \sum \sum_{i \neq j} I_h(d_{ij})$$

onde

$$I_h(d_{ij}) = \begin{cases} 1 \text{ se } d_{ij} \leq h \\ 0 \text{ se } d_{ij} > h \end{cases}$$

$I_h(d_{ij})$ é uma função indicador

Esta função também necessita de correção do efeito de borda



Função K e Função L

A função K(h) tem uma distribuição teórica sob condições de aleatoriedade, quando a probabilidade de ocorrência de um evento em qualquer ponto de R é independente da ocorrência de outros eventos e igual em toda a superfície.

Neste caso, o nº de eventos a uma distância h será $\boldsymbol{p}\,\boldsymbol{l}\,h^2$ e $K(h) = \boldsymbol{p}\,h^2$

No caso de distribuição regular, $K(h)$ será menor que $\boldsymbol{p}\,h^2$

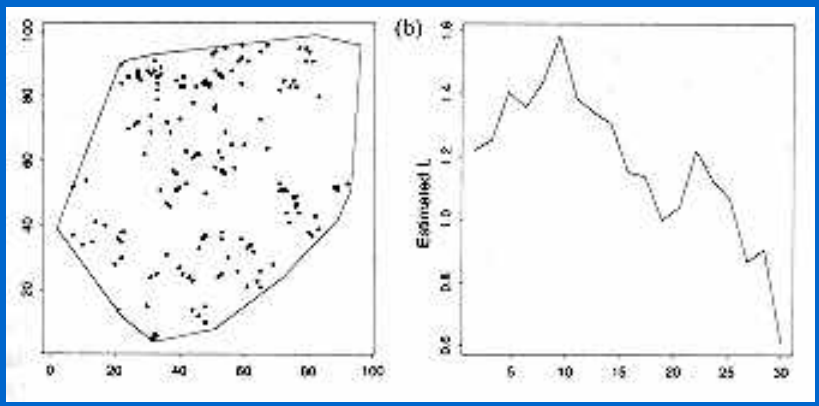
Distribuição em cluster, $K(h)$ será maior que $\boldsymbol{p}\,h^2$

A função $L(h)$ permite comparar entre a função $K(h)$ e $\boldsymbol{p}\,h^2$

$$\hat{L}(h) = \sqrt{\frac{\hat{K}(h)}{\pi}} - h$$

- Picos positivos indicam atração espacial → cluster
- Vales negativos → repulsão espacial ou regularidade

Função L



Há evidência de cluster - todos os valores positivos
Maior em h = 10

Completa Aleatoriedade Espacial

- Para testar se as distribuições observadas na análise exploratória são de fato significativas, é necessário comparar com distribuições teóricas ou simulações que representem a "Completa Aleatoriedade Espacial"
- A hipótese de CAE é que o evento segue um processo de Poisson homogêneo sobre a região estudada, e os testes buscam verificar isso
- Outras modelos podem ser usados: processo de Poisson heterogêneo, processo de Cox, inibição simples, etc.

Testes de cluster espacial e espaço-temporal

- São causas de *cluster*: fonte comum, contagiosidade, acaso
- Para testar se este agregado é acima de um valor esperado, existem diversos testes:
  - Knox - testa um número acima do esperado de pares de casos excessivamente próximos (segundo critério pré-estabelecido) no espaço e no tempo;
  - Mantel - pondera todos os pares pela sua distância espaço-tempo;
  - Cuzick-Edwards - caso-controle onde a coincidência de casos vizinhos aumenta o peso, e a junção controle-controle ou caso-controle tem peso zero; este teste permite considerar a variação populacional.

Variação da população

- O alisamento Kernel permite estimar "eventos por unidade de área", sem considerar a população
- Pode-se estimar "população por unidade de área", e fazer a razão dos dois obtendo uma estimativa alisada de "eventos por população"

$$\hat{\lambda}_\tau(s) = \sum_{j-1}^{m} \frac{1}{\tau^2} k\left(\frac{(s - s_j')}{\tau}\right) y_j$$

λ' - estimativa população p/ unidade de área
τ - largura de banda
y_j - população em cada ponto

Usa-se atribuir ao centróide do setor censitário ou ao centro populacional o número de habitantes de toda a área

Variação da população - “taxa”

- A criação da taxa é a divisão dos alisamentos:

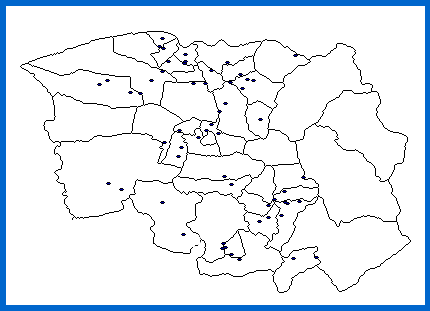
$$\frac{\text{eventos p/ unidade de área}}{\text{população p/unidade de área}}$$

$$\hat{\rho}_\tau(s) = \frac{\sum_{i=1}^{n} \frac{1}{\tau^2} k\left(\frac{(s - s_i)}{\tau}\right)}{\sum_{j=1}^{m} \frac{1}{\tau^2} k\left(\frac{(s - s_j')}{\tau}\right) y_j}$$

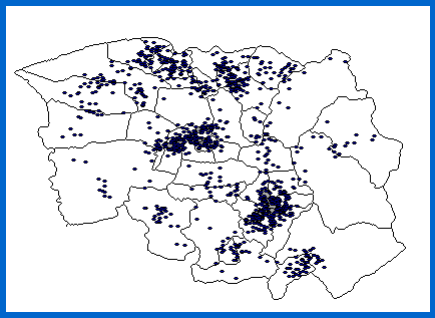
Pode-se usar diferentes larguras de banda (em geral maior no denominador para estabilizar +)

Pode-se usar outro evento como “estimador da população a risco”

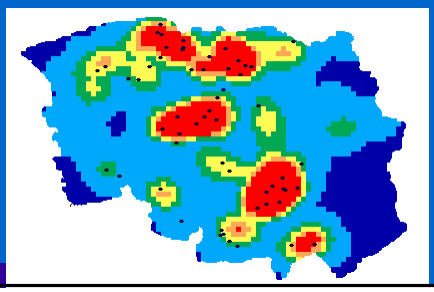
Exemplo



câncer de laringe



câncer de pulmão

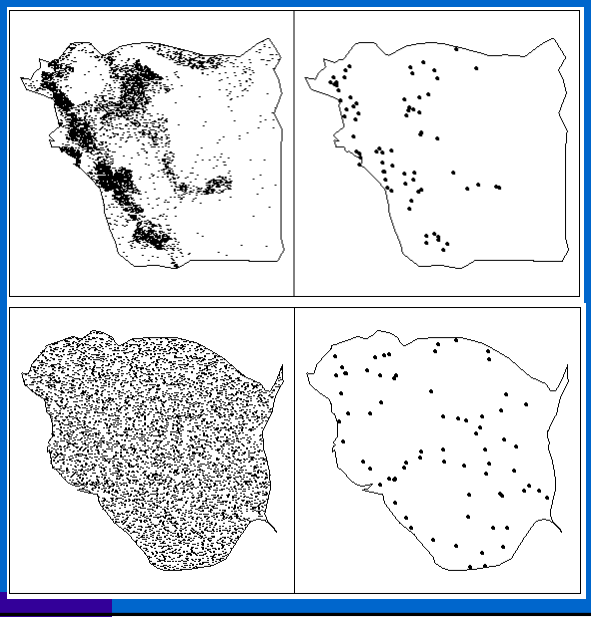


kernel câncer de pulmão
casos de câncer de laringe

BAILEY & GATRELL, 1995

DEM

Investigação de *cluster* de câncer de mama em S.Francisco



SELVIN et al, 1996

Fonte específica

- Cluster ao redor de um ponto ou uma linha
- Compara-se a ocorrência de nº excessivo de "casos" em relação à população a partir de uma função de decaimento em relação à possível fonte

$$\lambda(s) = \rho\lambda'(s) f(h;\theta)$$

$$f(h;\theta) = 1 + \theta_1 e^{\theta_2 h^2}$$

θ - parâmetros a estimar que descrevem como a incidência varia em torno da fonte

$\lambda(s)$ - estimativa do evento p/ unidade de área

ρ - parâmetro que indica a razão entre "casos" e "controles"

$\lambda'(s)$ - estimativa população p/ unidade de área

f - função da distância para a fonte

Análise de dados espacialmente contínuos

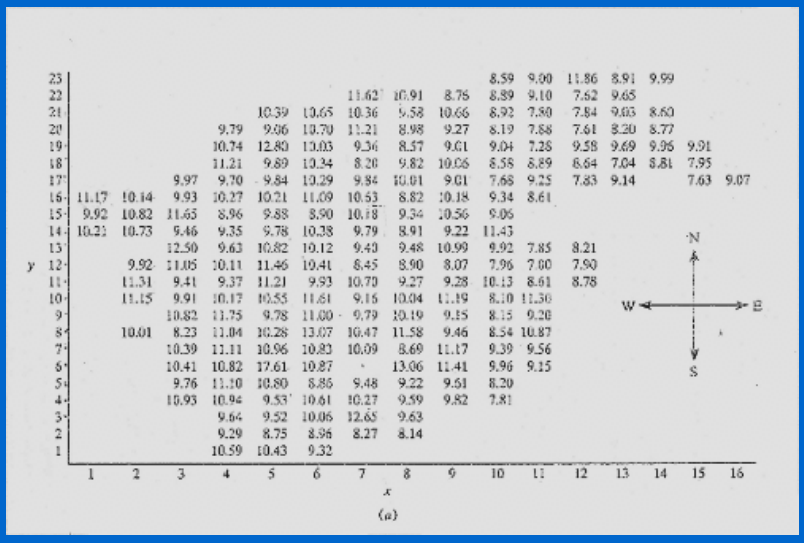
(*geoestatística*)

Introdução

- Na análise de padrão de pontos, o interesse é na localização dos eventos. Na análise de dados espacialmente contínuos, o objetivo é **entender** a distribuição espacial dos valores de um atributo de toda a região estudada, a partir de medidas realizadas em pontos amostrados. As coordenadas dos pontos, neste caso, é apenas a localização onde a variável foi mensurada;
- A estrutura do banco de dados é:

Amostra	**Coord X**	**Coord Y**	**Var1** (°C)	**Var2** (ppm)	**Var3** (p/10^5hab)
1	42°30’	22°45’	32°	0,50	1,7
2	42°39’	22°35’	25°	1,45	2,6
3	42°10’	23°50’	28°	5,87	80,6

Visualização



CRESSIE, 1991 - Fig. 2.2 , pag. 34

Visualização - mapas de símbolos

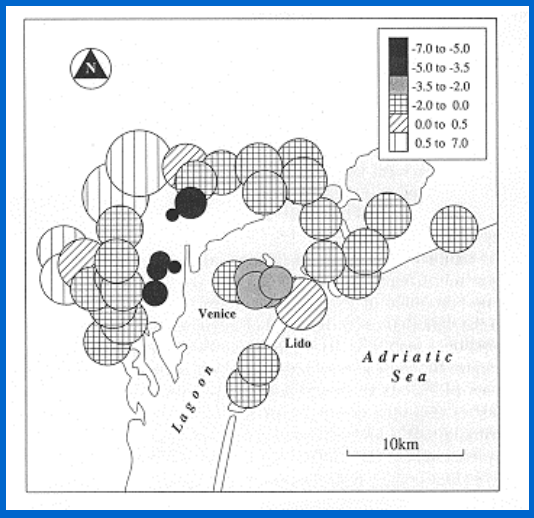


Fig. 5.5 Proporcional symbol map of Venice groundwater, 1973 *BAILEY & GATRELL*, 1995

Análise exploratória: contorno

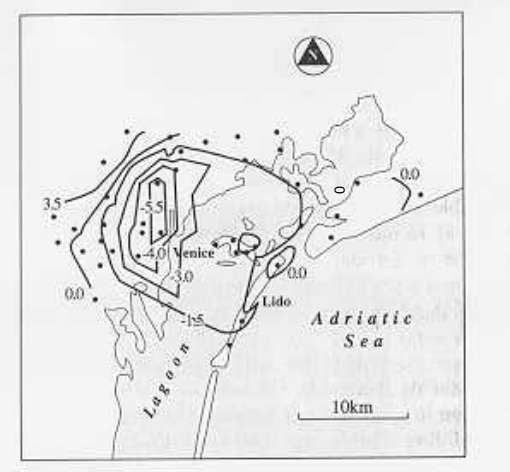
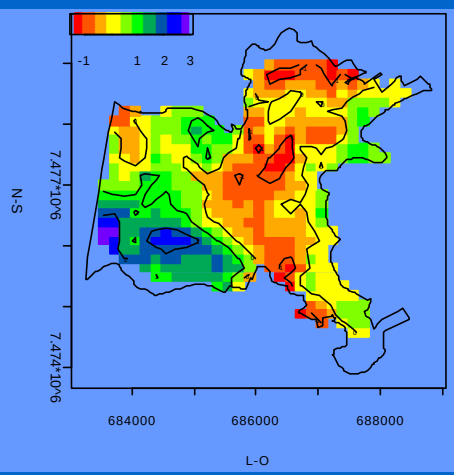
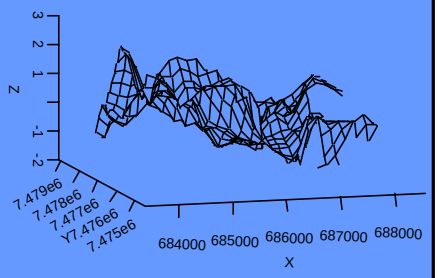


Fig. 5.8 TIN contours of Venice groundwater in 1973

BAILEY & GATRELL, 1995

Mapas de contorno e 3D





CARVALHO, M.S..,1997

Continuidade: mapas de contorno e 3D

- as curvas de contorno, ou isolinhas, delimitam áreas onde a variável tem a mesma grandeza
- no mapa 3D é utilizada uma projeção tridimensional onde a variável em estudo é representada no eixo Z, perpendicular aos eixos X e Y das coordenadas de representação espacial, obtendo-se uma superfície em forma de "cordilheira", com picos e vales representado os diversos valores encontrados em dada área geográfica;
- se for feito um corte em um determinado valor de Z do mapa 3D a visualização em duas dimensões é o mapa de contorno;
- este mapas são construídos por interpolação de valores medidos em diversos pontos;
- poucos programas fazem este tipo de mapa.

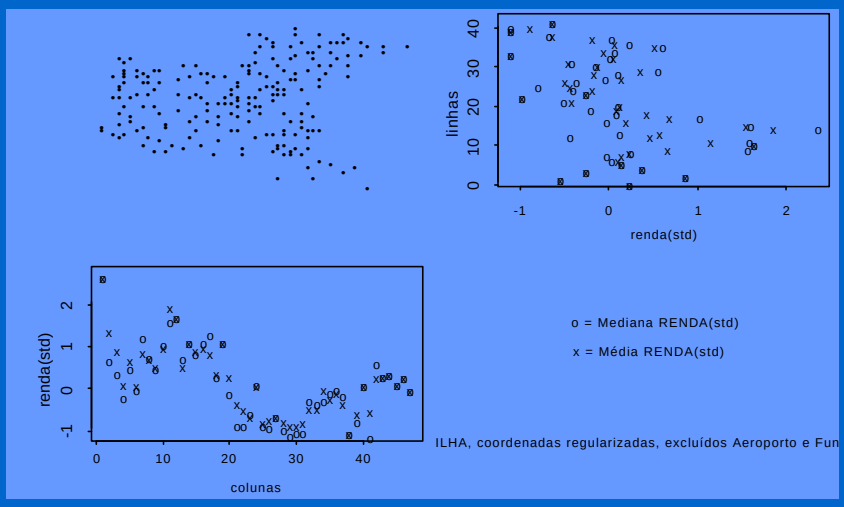
Modelagem 1

- modela-se a variável segundo sua distribuição em larga escala (tendência) e pequena escala (relação com os pontos vizinhos);
- o primeiro passo é transformar a variável buscando aproximar sua distribuição de uma "normal", utilizando transformações (log, exp,...);
- em seguida se estuda a estacionariedade: tendência, outliers, anisotropia.

Tendência e outliers

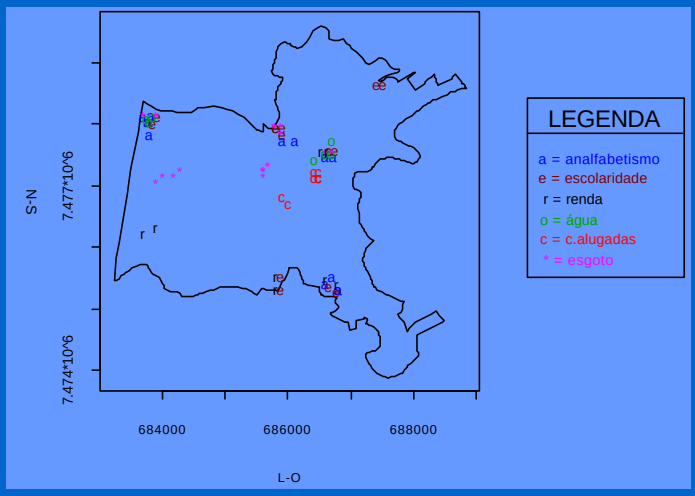
- localização de valores extremos nos mapas
- gráficos de médias e medianas segundo linhas e colunas dos pontos amostrados - permite identificar a flutuação das medidas ao longo de duas direções, permitindo detectar tendência ou valores aberrantes
- saltos no valor das variáveis em pequena distância

Gráfico de médias e medianas direcionais



CARVALHO, M.S.,1997

Mapa de grandes diferenças



CARVALHO, M.S..,1997

Relação entre os pontos - pequena escala

- Variograma amostral: $$\gamma(h) = \frac{1}{2N(h)} \sum_{(i,j)} (v_i - v_j)^2$$

g(h) - valor calculado do semi-variograma amostral para cada intervalo (*h*) entre pares de pontos;
N(h) - total de pares que tem entre si a mesma distância (*h*);
v_i - valor da variável medida em *i* e *j*

- Autocovariância: $$Cov(h) = \frac{1}{N(h)} \sum_{i,j} \left[v_i v_j - \frac{1}{n} \sum_{k=1}^{n} v_k \right]^{\frac{1}{2}}$$

- Autocorrelação: autocovariância normalizada pela variância total.

Variograma

- intervalos (lags):

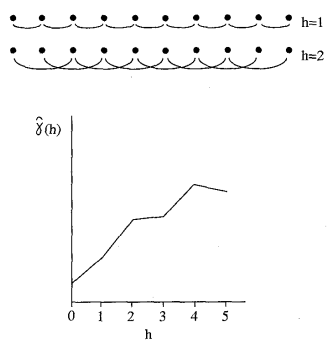
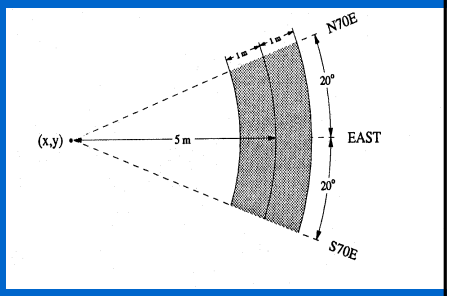


Fig. 5.10 Typical sample variogram

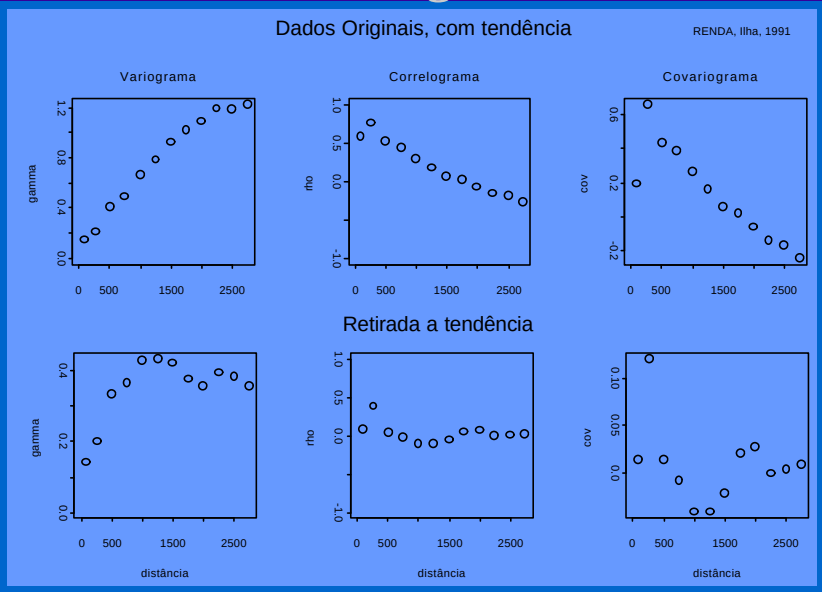
- tolerância:



ISAACS & SHRIVASTAVA, 1989

BAILEY & GATRELL, 1995

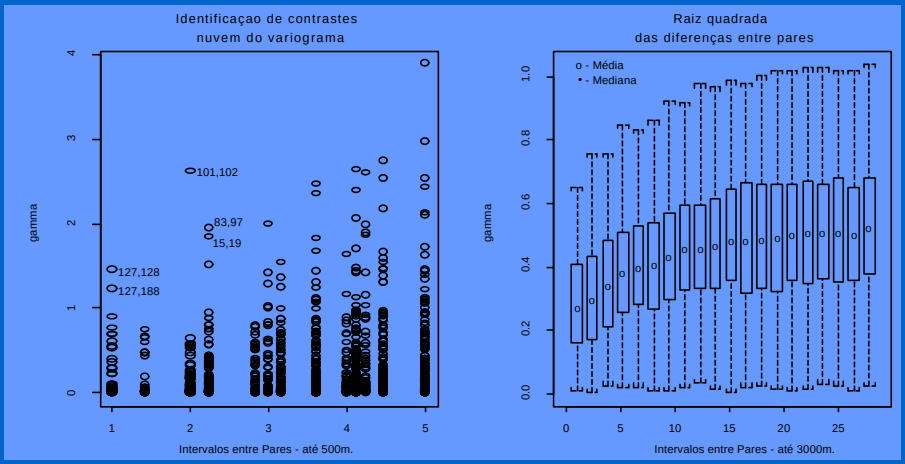
Variograma



CARVALHO, M.S..,1997

Estacionariedade de 2ª ordem

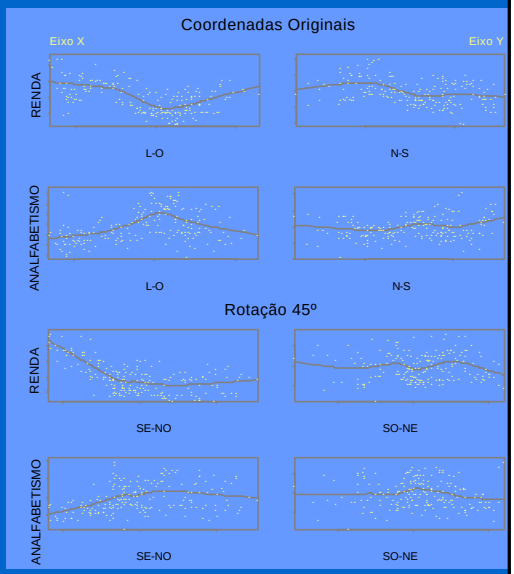
- identificar diferenças aberrantes (utilizando o variograma) entre pares de valores mensurados



CARVALHO, M.S..,1997

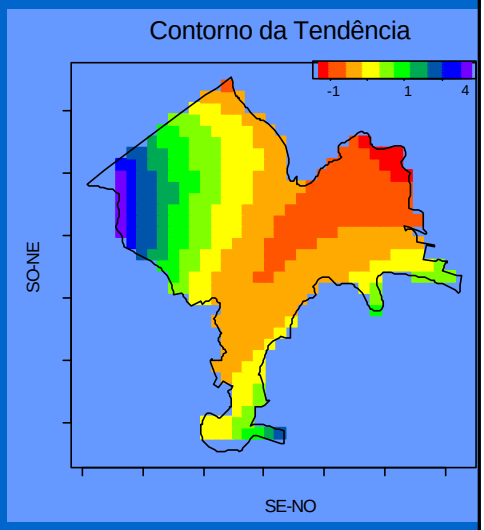
Modelagem 2 - tendência

- A tendência pode ser modelada através de polinômios ou alisamentos "locais"
- O peso das observações diminui à medida em que se afasta do ponto estimado, sendo então uma regressão local ponderada.



Modelagem 2 - *loess*

- Ao invés de se calcular a média em cada janela, se estima, por mínimos quadrados, os parâmetros de um plano.
- Depois de modelada, se retira a tendência e examina os resíduos



Isotropia

- Quando a variabilidade espacial de um fenômeno em estudo é a mesma em todas as direções, diz-se que o fenômeno é ISOTRÓPICO

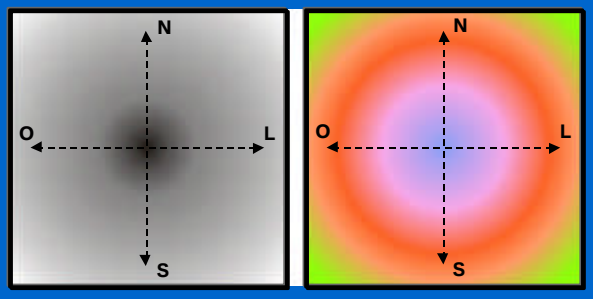


Imagem nível de cinza Composição Colorida

Anisotropia

- Quando a variabilidade espacial de um fenômeno em estudo não é a mesma em todas as direções, diz-se que o fenômeno é ANISOTRÓPICO.

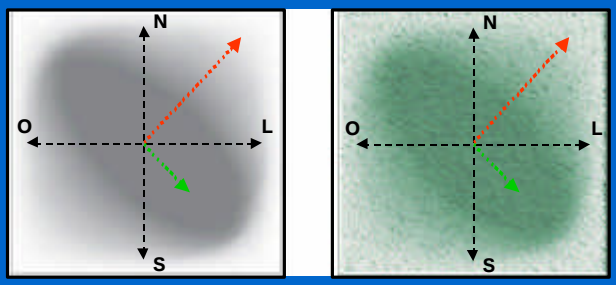
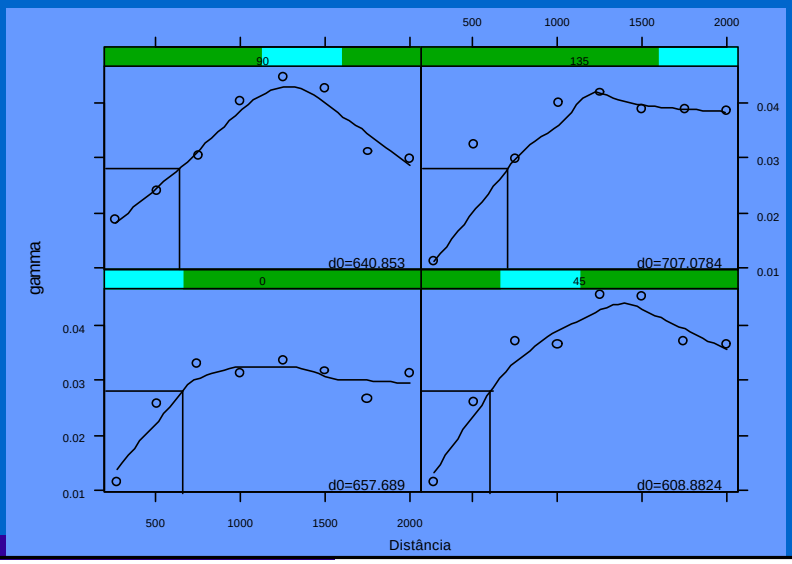


Imagem nível de cinza Composição Colorida

Modelagem 3 - anisotropia

- Se houver anisotropia, é necessário corrigí-la



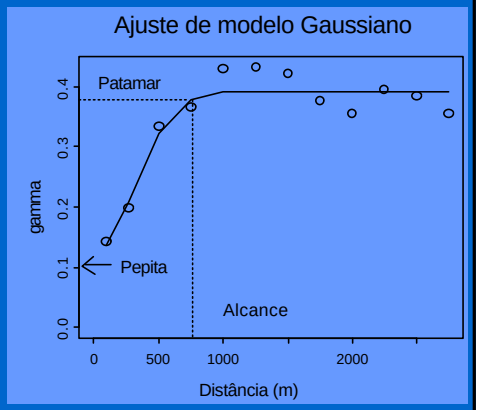
Modelagem 4 - variograma

- Somente então é possível modelar a variação em pequena escala através de ajuste de uma função ao variograma amostral
- Modelando como se dá a relação entre os pontos próximos é possível interpolar o valor da variável em qualquer ponto, e investigar a relação entre diversas variáveis que ocorrem de forma contínua na região

Ajuste do variograma

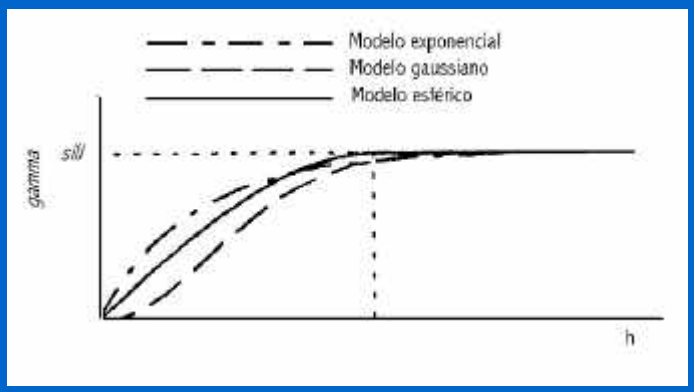
- os principais parâmetros a serem ajustados são:

- função da curva: esférica, gaussiana ou exponencial;
- patamar (*sill*): valor máximo atingido;
- alcance (*range*): distância até onde existe correlação entre os pontos;
- pepita (*nugget*): valor inicial, que representa a diferença medida onde a distância tende a 0.



CARVALHO, M.S..,1997

Modelos de variograma

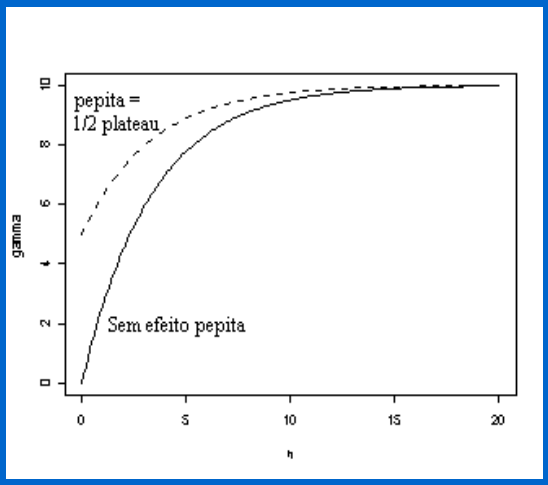


ISAACS & SHRIVASTAVA, 1989

Modelos de variograma

EQUAÇÃO	OBSERVAÇÕES
$g(h)=\begin{cases}1{,}5\frac{h}{a}-0{,}5\left(\frac{h}{a}\right)^3 & ,\text{se } h\le a\\ 1 & ,\text{caso contrário}\end{cases}$	• comportamento linear próximo à origem
$g(h)=1-\exp\left(\frac{-3h}{a}\right)$	• atinge o plateau assintoticamente, na prática, considera-se o valor de a onde o variograma atinge 95% do plateau.
$g(h)=1-\exp\left(\frac{-3h^2}{a^2}\right)$	• também assintótico, com crescimento parabólico próximo à origem

Variações nos modelos de variograma

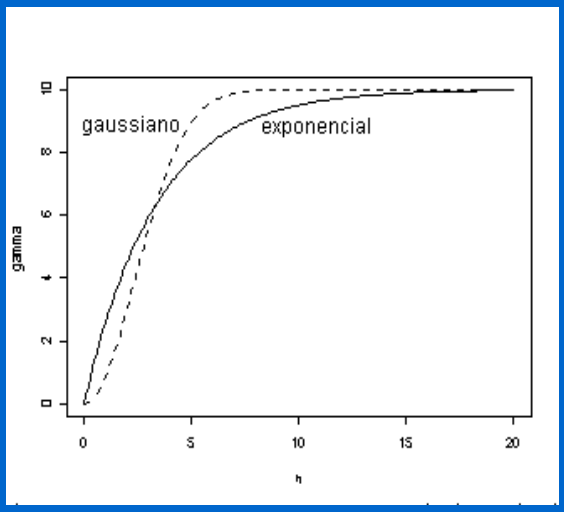


Efeito Pepita

$$\gamma(h) = 10\left(1 - \exp\left(\frac{-3|h|}{10}\right)\right)$$

$$g(h) = \begin{cases} 5 + 5\left(1 - \exp\left(\frac{-3|h|}{10}\right)\right) \\ \text{ou } 0 \text{ , se } h = 0 \end{cases}$$

Variações nos modelos de variograma

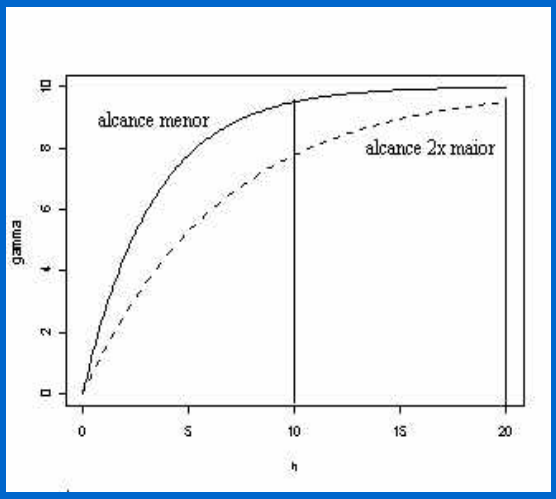


Modelo

$$g(h) = 10\left(1 - \exp\left(\frac{-3|h|}{10}\right)\right)$$

$$g(h) = 10\left(1 - \exp\left(\frac{-3|h|}{10}\right)^2\right)$$

Variações nos modelos de variograma



Alcance

$$g(h) = 10\left(1 - \exp\left(\frac{-3|h|}{10}\right)\right)$$

$$g(h) = 10\left(1 - \exp\left(\frac{-15|h|}{100}\right)\right)$$

Interpolando: Krigeagem

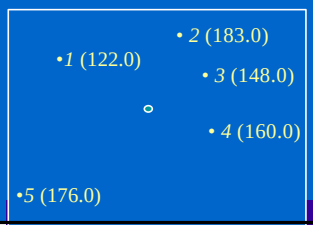
- a krigeagem (*kriging*) é o método para interpolação de qualquer ponto, utilizando os pontos vizinhos e relacionando-os através do modelo de variograma;
- a krigeagem atribui pesos diferentes conforme a distância entre o ponto a ser estimado e os pontos amostrados:

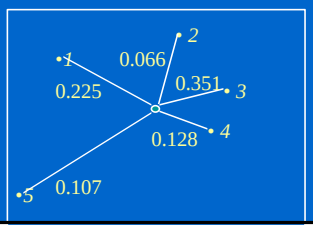
$$\hat{\upsilon}(s) = \sum_{j=1}^{n} (\omega_j . \upsilon_j)$$

$\upsilon(s)$ - valor estimado

υ_j - medidas nos pontos j

w - matriz de pesos, estimada a partir do modelo do variograma

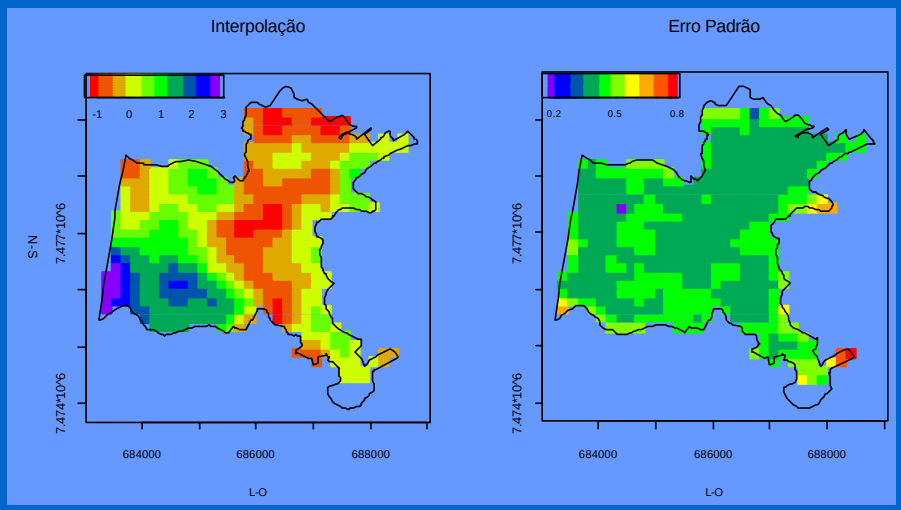




Interpolando: Krigeagem

- O kriging é denominado blue - “best linear unbiesed estimator” pela forma como é calculada a matriz de pesos
- No ponto onde houve medida o kriging garante com que o valor medido seja mantido
- O kriging permite estimar o erro padrão associado ao modelo
- A tendência deve ser reincorporada, sendo possível também estimar simultaneamente tendência e variação em pequena escala através da krigeagem universal
- Neste caso a tendência só pode ser modelada com polinômios

Krigeagem

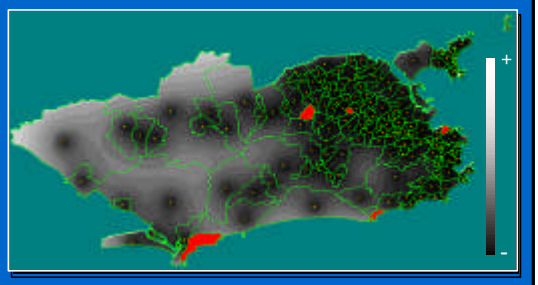


CARVALHO, M.S..,1997

Krigeagem

Imagem da variância de krigeagem relativa as proporções de nascidos com Apgar bom, no Município do RJ.

Imagem da variabilidade espacial das proporções de nascidos com Apgar bom, no Município do RJ, proveniente da Krigeagem.



Dados: D'Orsi,l.1995

Análise: equipe SPRING/INPE, 1999

Potencialidade

- aplicações mais indicadas:
  - ambiente e saúde;
  - identificação de características de ocupação do solo e imagem de satélite.
- precisão X facilidade;
- métodos alternativos:
  - alisamentos não paramétricos;
  - interpolação linear simples.
- outros recursos:
  - co-variograma e co-krigeagem;
  - análise multivariada.

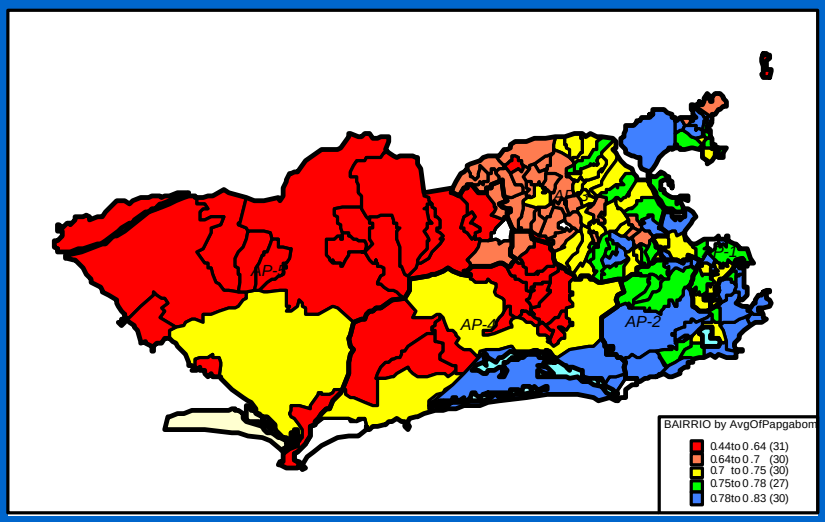
Análise de áreas

Introdução

- Na análise de áreas, ao invés de variar continuamente no espaço, o atributo estudado apresenta valor constante sendo medida de síntese;
- O objetivo não é a predição para pontos não mensurados, mas a detecção e explicação de padrões e tendências observados nas áreas;
- área é definida por um polígono cuja forma pode ser complexa bem como as relações de vizinhança;
- O modelo básico do banco de dados:

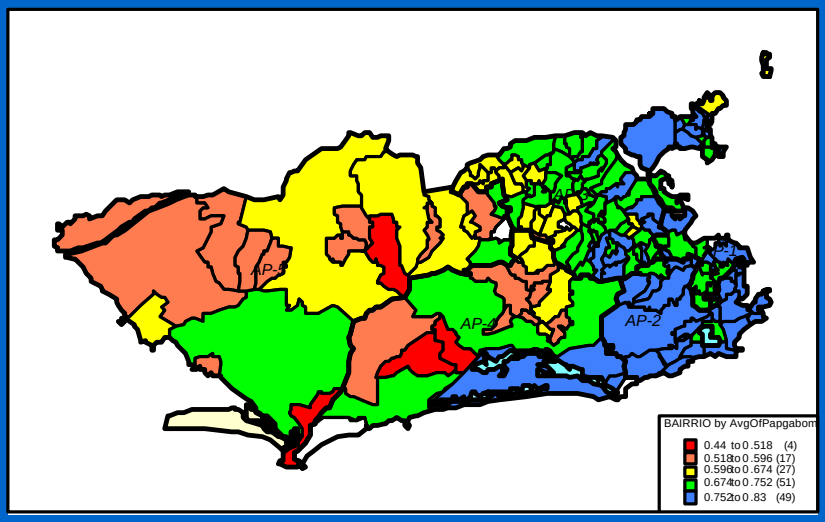
Local	Casos	População	Med/1000 hab.
Rio Bom	41	3209	5,4
Serra Verde	320	16897	2,6
Poço Fundo	67	2569	1,3

Forma de representação: Mapa de Padrão



D'ÓRSI,E., 1996

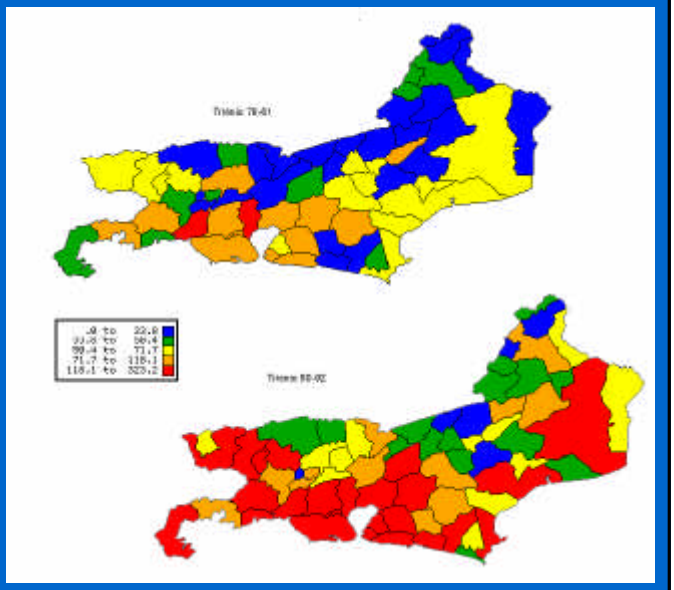
Forma de representação: Mapa de Padrão



Pontos de corte

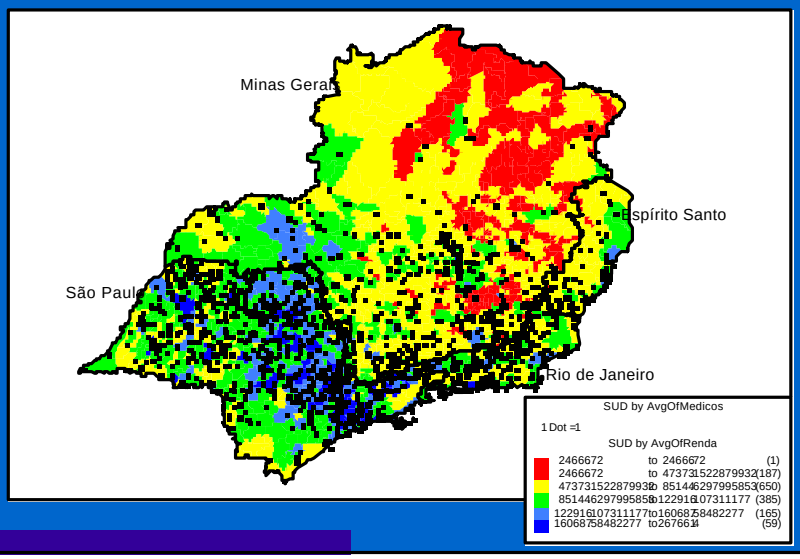
Mortalidade por Homicídios: triênios 79-81, 90-92

Estado do Rio de Janeiro

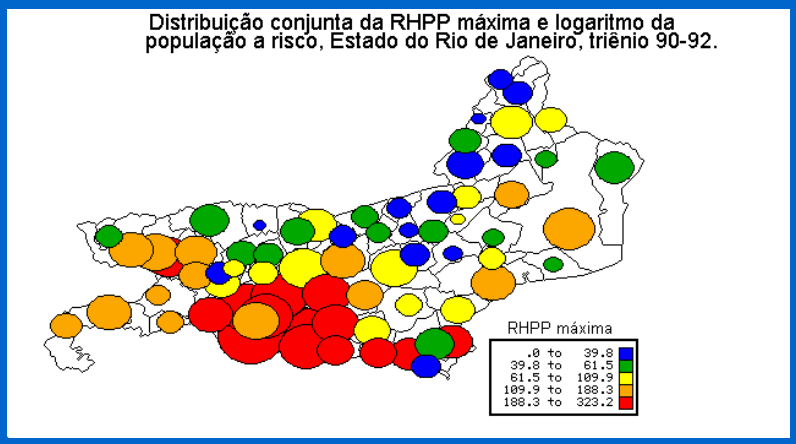


CRUZ,O.G.,1996

Análise exploratória bivariada - Pontos



Análise exploratória bivariada: Símbolos



CRUZ,O.G., 1996

Análise exploratória - alisamentos

- Médias móveis:

$$\hat{m}_i = \frac{\sum_{j=1}^{n} w_{i,j} y_i}{\sum_{j=1}^{n} w_{i,j}}$$

$w_{i,j}$ é a ponderação obtida da matriz de vizinhança
y_i é o valor do atributo na área

- Polimento pela mediana (median polish):

$Y_{i,j}$ é o valor do atributo na área, que pode ser decomposto em:
m - média global da área
m_i - efeito em linhas
m_i - efeito em colunas
e_{ij} - erro alatório

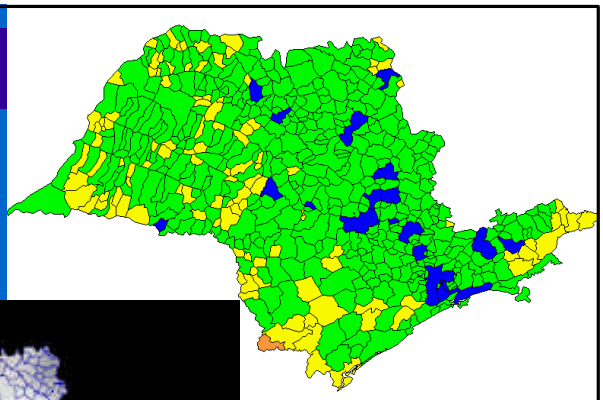
$$y_{i,j} = m + m_i + m_j + e_{i,j}$$

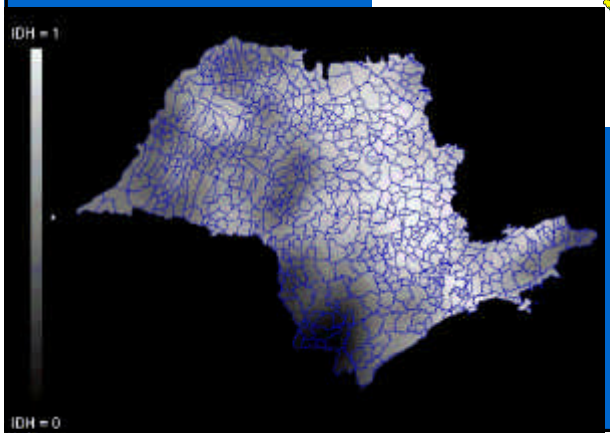
Equivale à análise de variância onde os grupos são as linhas e as colunas, mas utilizando medianas ao invés de médias

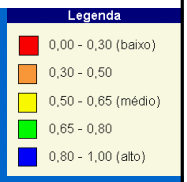
Entidades

ou

Superfícies ?

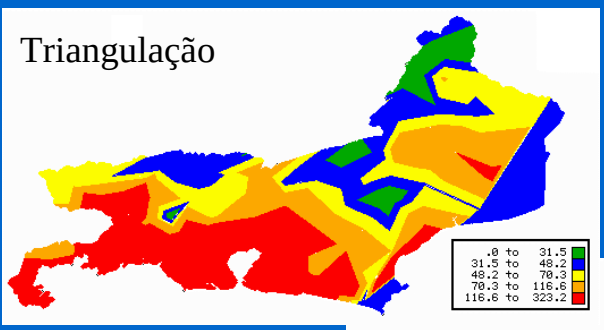




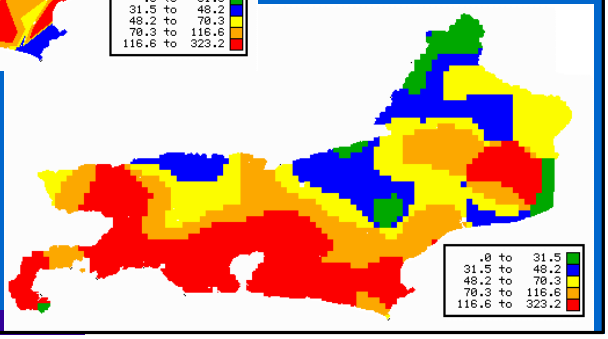


Interpolação em áreas

Triangulação



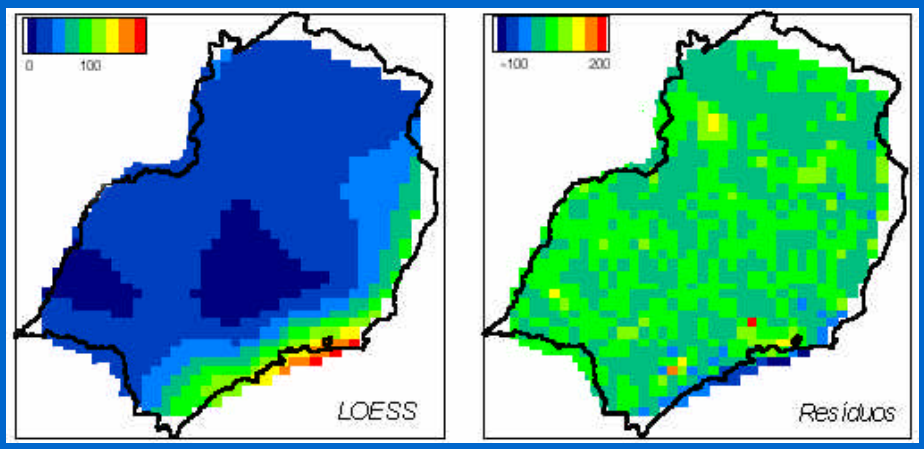
Kernel



CRUZ,O.G.,1996

Interpolação em áreas

Interpolação LOESS com janela de 30% dos pontos



Kernel de áreas

- Utiliza-se para áreas alocando o valor do atributo a um ponto da área - centróide geométrico, populacional

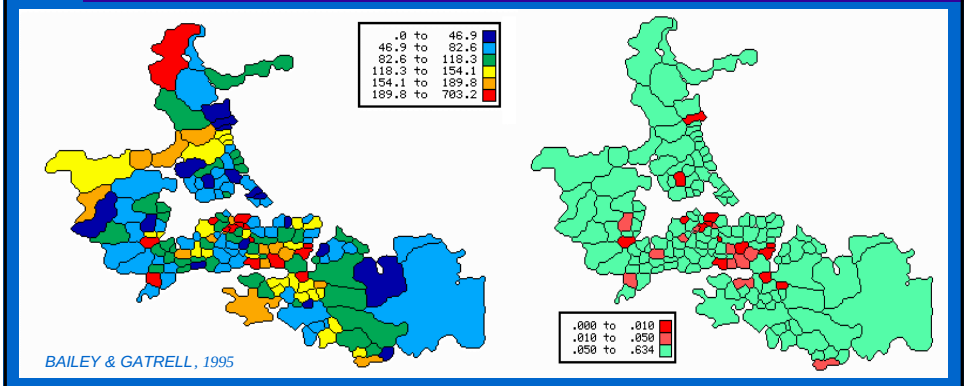
$$\hat{p}_t(s) = \sum_{j=1}^{n} k\left(\frac{s - s_i}{t}\right) p_i$$

- Para o kernel de população, cada ponto receberá o atributo p_i (população) alisado pela função k, e largura de banda t

$$\hat{m}_t(s) = \frac{\sum_{j=1}^{n} k\left(\frac{s - s_i}{t}\right) y_i}{\sum_{j=1}^{n} k\left(\frac{s - s_i}{t}\right)}$$

- No kernel de um atributo contínuo (por ex., indicadores), inclui-se no denominador o kernel da distribuição dos centróides das áreas
- Obtém-se portanto a média do atributo na região e não uma contagem de eventos por unidade de área
- Correção para efeitos de borda

Mapa de probabilidades



Estimativa da média em i

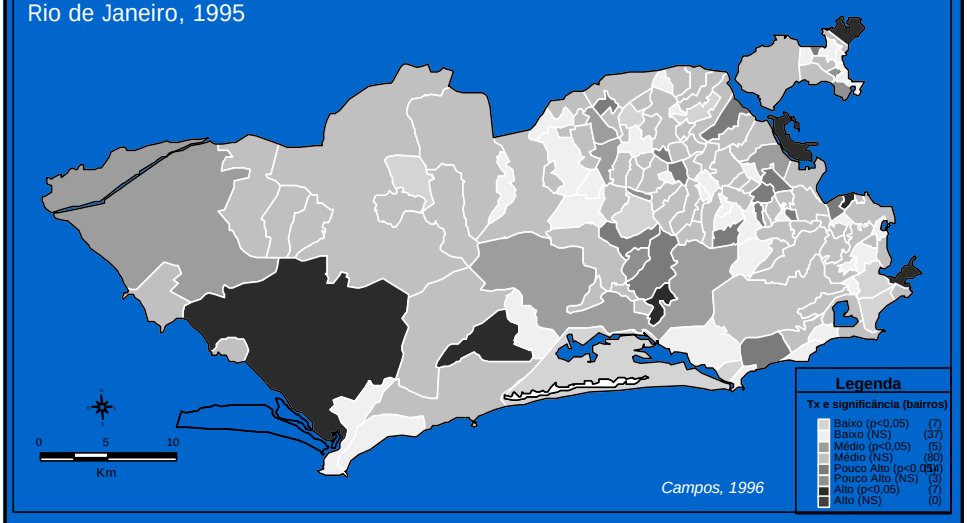
$$\hat{m}_i = n_i\left(\frac{\sum y_i}{\sum n_i}\right)$$

$$p_i = \begin{cases} \sum_{x \geq y_i} \frac{\hat{m}_i^{x} e^{-\hat{m}_i}}{x!} & \text{para } y_i \geq \hat{m}_i \\ \sum_{x \leq y_i} \frac{\hat{m}_i^{x} e^{-\hat{m}_i}}{x!} & \text{para } y_i < \hat{m}_i \end{cases}$$

p_i é a probabilidade de encontrar o valor y_i em cada área

Mapa de probabilidades

Mortalidade posneonatal (taxa e significância)
Rio de Janeiro, 1995



Cluster em áreas

- diz-se que existe um cluster entre áreas quando áreas com valores semelhantes ocorrem próximas no espaço;
- ou quando existe uma quantidade "excessiva de eventos" na mesma área
- são causas de cluster: fonte comum, contagiosidade, acaso;
- para testar se este agregado é acima de um valor esperado, existem diversos testes que procuram verificar a medida da autocorrelação espacial, testando se significativa:
- os resultados de qualquer destes métodos depende diretamente dos pesos da matriz de vizinhança.

Matriz de vizinhança

- utiliza-se matriz W , onde cada elemento w_{ij} representa medida de proximidade espacial entre as áreas A_i e A_j;
- a escolha de w_{ij} depende do tipo de dado, de região, dos mecanismos particulares da dependência espacial;
- vizinhos podem ser de primeira ordem, segunda até n.

Possíveis Critérios:

$$w_{ij} = \begin{cases} 1 & \text{centróide de } A_i \text{ é o mais próximo de } A_j \\ 0 & \text{caso contrário} \end{cases}$$

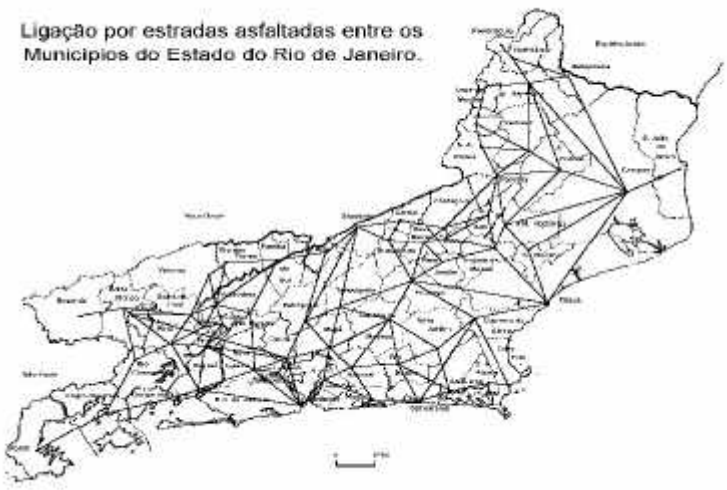
$$w_{ij} = \begin{cases} 1 & \text{centróide de } A_i \text{ dentro de distância especificada de } A_j \text{ (buffer)} \\ 0 & \text{caso contrário} \end{cases}$$

$$w_{ij} = \begin{cases} 1 & A_i \text{ tem fronteira comum com } A_j \\ 0 & \text{caso contrário} \end{cases}$$

$$w_{ij} = \frac{l_{ij}}{l_i}$$

l_{ij} é o comprimento da fronteira comum entre com A_i e A_j e l_i é o perímetro de A_i

Matriz de vizinhança



Testes de Cluster

Moran I

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

- W_{ij} é a matriz de vizinhança
- Relaciona-se à auto-correlação
- Média $\bar{y}$ suposta constante: processo estacionário

Geary C

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

- Relaciona-se ao variograma
- Outros testes: Moran Ipop, Assunção

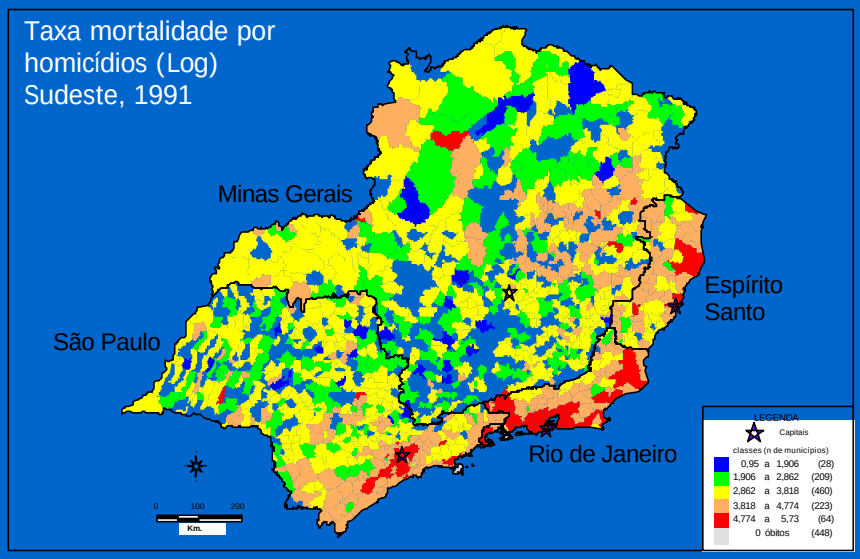
Função de autocorrelação

Moran no lag k

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

- Desta forma se constrói a função de autocorrelação para cada lag
- A significância estatística pode ser calculada por permutação ou, caso a variável tenha distribuição normal, por teste Z

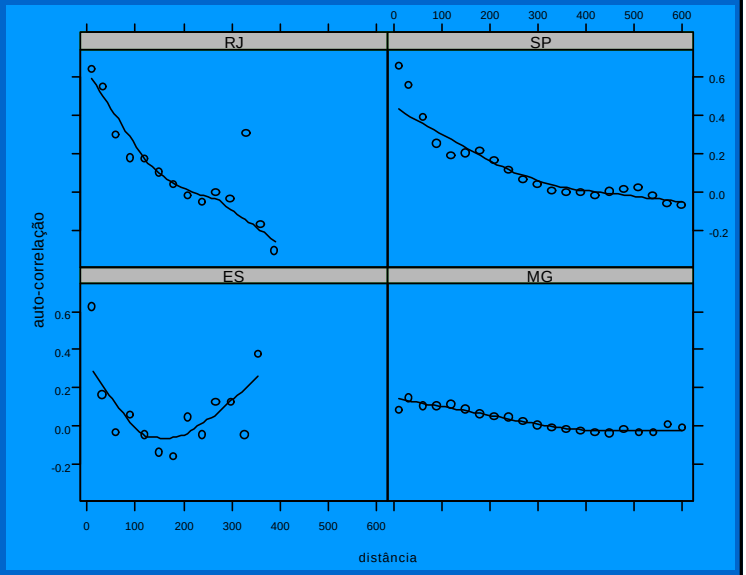
Autocorrelação



CARVALHO & CRUZ,1998

Correlograma

Correlograma da taxa mortalidade por homicídios por UF



Indicadores locais

- Permitem encontrar os “bolsões” de dependência espacial não evidenciados nos índices globais
- permitem identificar:
  - agrupamentos de objetos com valores semelhantes (cluster)
  - objetos anômalos
  - existência de mais de um processo espacial
- A significância estatística também é calculada por permutações e supõe-se normalidade da variável.
- Existem dois índices locais:
  - LISA (Anselin, 1996)
  - Índice G_i e G_i^* (Getis e Ord, 1992)

Indicadores Locais

LISA - Indicador local de autocorrelação espacial

$$I_i = \frac{z_i \sum_j w_{ij} z_j}{\sum_{i=1}^{N} z_i^2}$$

- Z_i - desvio de i em relação a média global
- Z_j - média dos desvios dos vizinhos de i
- Média constante: processo estacionário
- Significância semelhante a I - permutação ou normalidade

Indicadores Locais

Estatística G

$$G_i(d) = \frac{\sum_{i=1}^{n} w_{ij}(d) y_j}{\sum_{j=1}^{n} y_j}, i \neq j$$

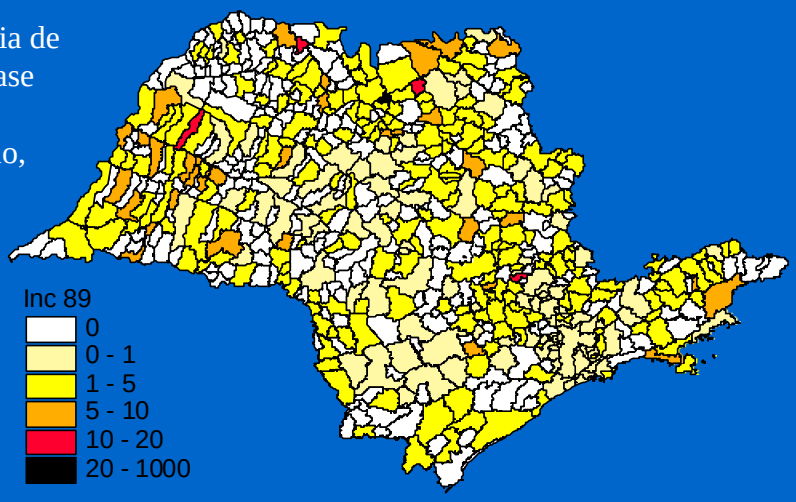
- No numerador: somatório de todos os valores de todos os vizinhos dentro de distância d ponderados por W_{ij}
- G_i^* inclui também o ponto i no cálculo
- Valores positivos: cluster de valores altos;
  negativos: cluster baixo
- Interpretação mais direta que o LISA

Mapa LISA

Moran I = 0.07185

Incidência de Hanseníase

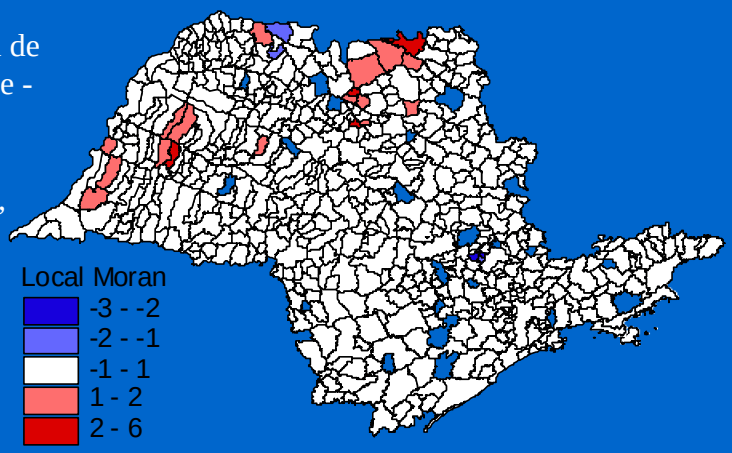
São Paulo, 1989



Mapa LISA

Incidência de Hanseníase - LISA

São Paulo, 1989



Modelos de regressão NÃO espacial

- Na investigação sobre causas de diferenças entre áreas é possível utilizar modelos multivariados não espaciais (estudos ecológicos clássicos).

$$y_i = b_0 + b_1 x_1 + \ldots + b_k x_k + e_i$$

y - estimativa da var. resposta;
b_i - coeficiente de regressão;
x_k variável explicativa;
e é erro aleatório

- As hipótese básicas deste modelo são:
  - As variáveis explicativas são linearmente independentes;
  - $E(\varepsilon)=0$
  - $V(\varepsilon)=\sigma_\varepsilon^2$
  - $\varepsilon \sim (0, \sigma_\varepsilon^2)$
- Embora úteis, se existir forte tendência ou correlação espacial, os resultados serão influenciados, apresentando associação estatística onde não existem (e vice-versa).

Modelos de regressão espacial

- Novamente o objetivo é modelar simultaneamente a variação em larga escala e em pequena escala.

$$Z_i = m_i + d$$

Z_i - processo espacial
μ_i - Estimativa da média em i

$\delta \sim N(0,\Sigma)$, onde Σ é a matriz de covariância das variáveis aleatórias nos locais

- A variação na média - larga escala - pode ser modelada em função das coordenadas (superfície de tendência)
- Em pequena escala ajusta-se um modelo autorregressivo ou de médias móveis à Σ

Modelo de superfície de tendência

- Pode-se incluir no modelo de regressão comum as coordenadas geográficas de cada ponto como variáveis independentes, inclusive ao quadrado e seus produtos - neste caso se modela a superfície de tendência.

$$m_i = b_{10}x_i + b_{20}x_i^2 + b_{11}x_i y_i + b_{01}y_i + b_{02}y_i^2 + e_i$$

Modelos de regressão espacial - SAR

- Suponha que a variável y_i depende dos valores da variável independente nas áreas vizinhas a i:

$$y_i = b_0 + b_1 x_{1,i} + \ldots + b_k x_{k,i} + t\sum_j w_i x_j + e$$

t - parâmetro de associação entre o valor da variável explicativa nas áreas vizinhas e a variável resposta;
W_i- conjunto de áreas adjacentes i

- Suponha que a variável y_i é autocorrelacionada:

$$y_i = b_0 + b_1 x_{1,i} + \ldots + b_k x_{k,i} + r\sum_j w_i y_j + e$$

r - parâmetro da função de autocorrelação

Modelos de regressão espacial - CAR

- A medida que aumenta a complexidade, a estimativa de cada componente isolado do modelo torna-se impossível
- Utiliza-se então os modelos CAR (Conditional Autorregressive) que são modelos de regressão com erros espacialmente correlacionados

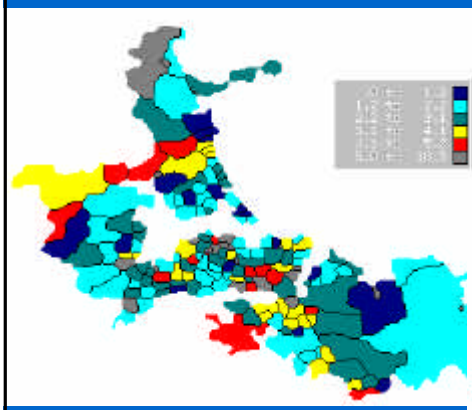
$$y_i = b_0 + b_1 x_{1,i} + \ldots + b_k x_{k,i} + e$$

Onde:
o erro tem matriz de covariância igual a s^2V
V é uma matriz não diagonal que descreve a dependência espacial dos erros

Modelagem Bayesiana

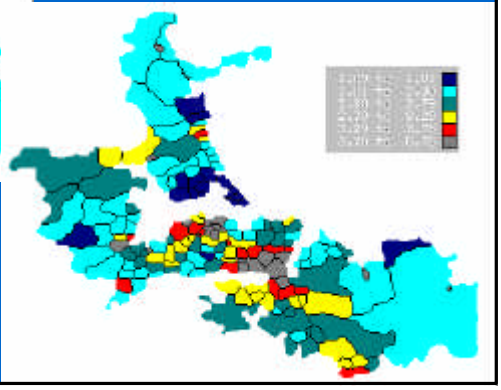
- Basicamente técnicas de "mapeamento de doenças" onde se incorpora o conhecimento "a priori" do investigador
- A principal característica é procurar identificar o processo que gerou aqueles dados, e não apenas a realização os dados, filtrando a variabilidade aleatória
- Para isso incorpora-se informação das áreas vizinhas: "vizinhos são parecidos"
- O mais utilizado método de estimativa - Markov Chain Monte Carlo (MCMC) - através de simulações permite estimar não só o valor esperado da distribuição da variável estudada em cada área, mas outros parâmetros também.

Ex: Bayesiano empírico



Taxa bruta de mortalidade infantil

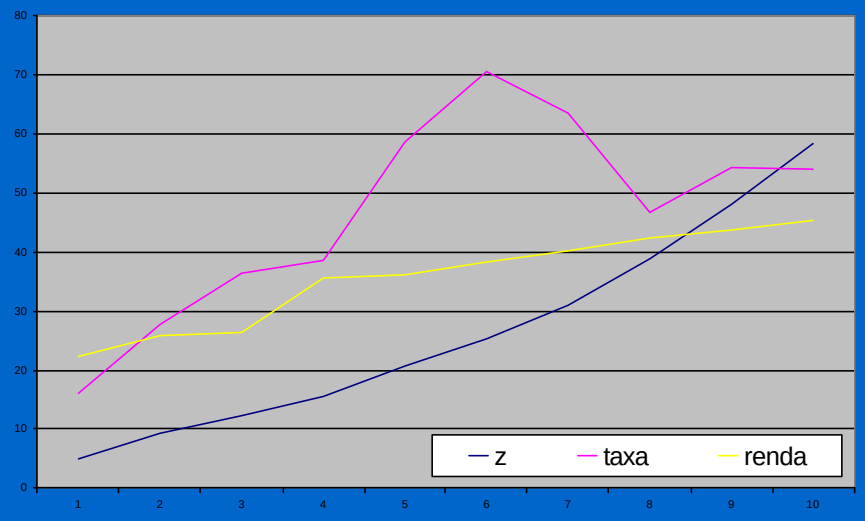
Alisamento Bayesiano empírico da mortalidade infantil



BAILEY & GATRELL, 1995

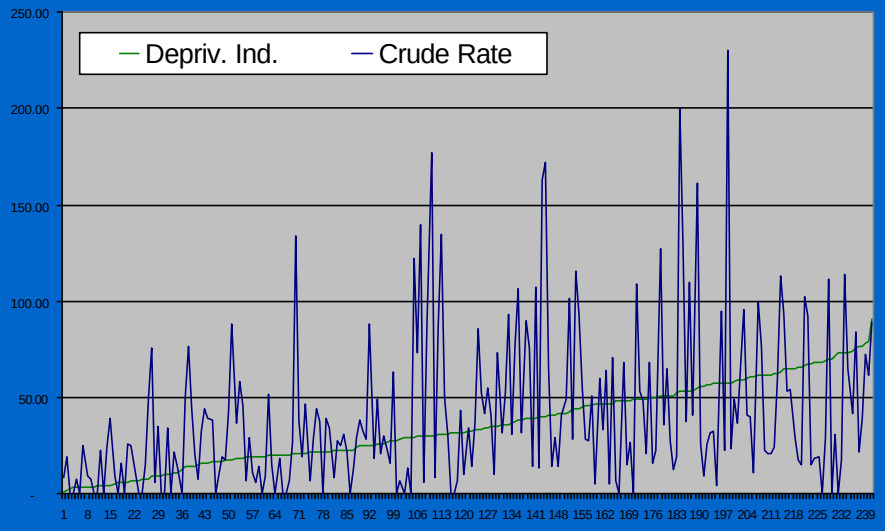
Ex: Hanseníase em Olinda

Setores censitários ordenados segundo decis de indicador de carência (z)



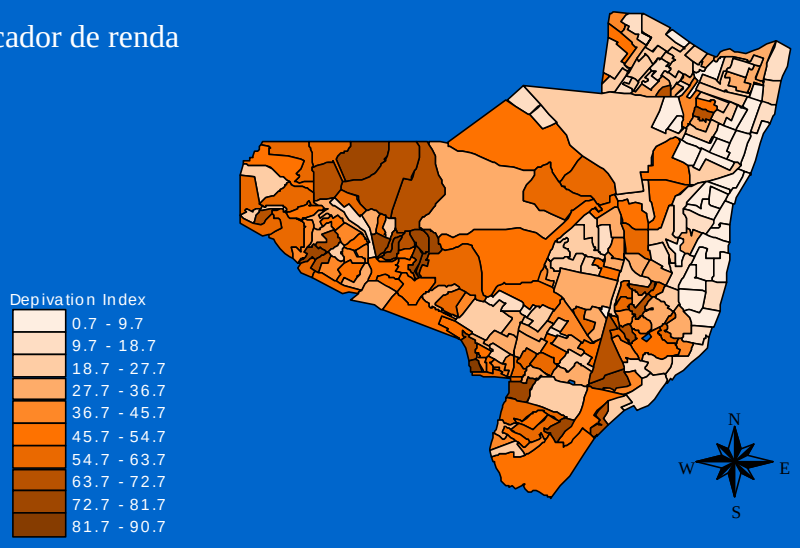
Renda X Hanseníase

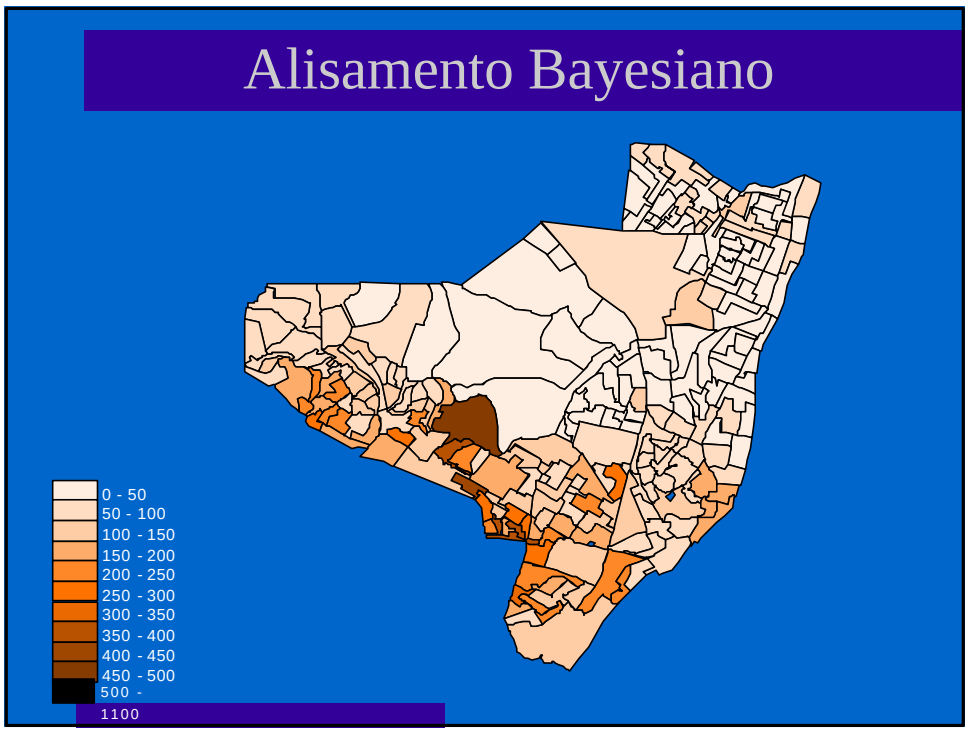
Indicador de renda: % chefes com renda < 1 salário mínimo

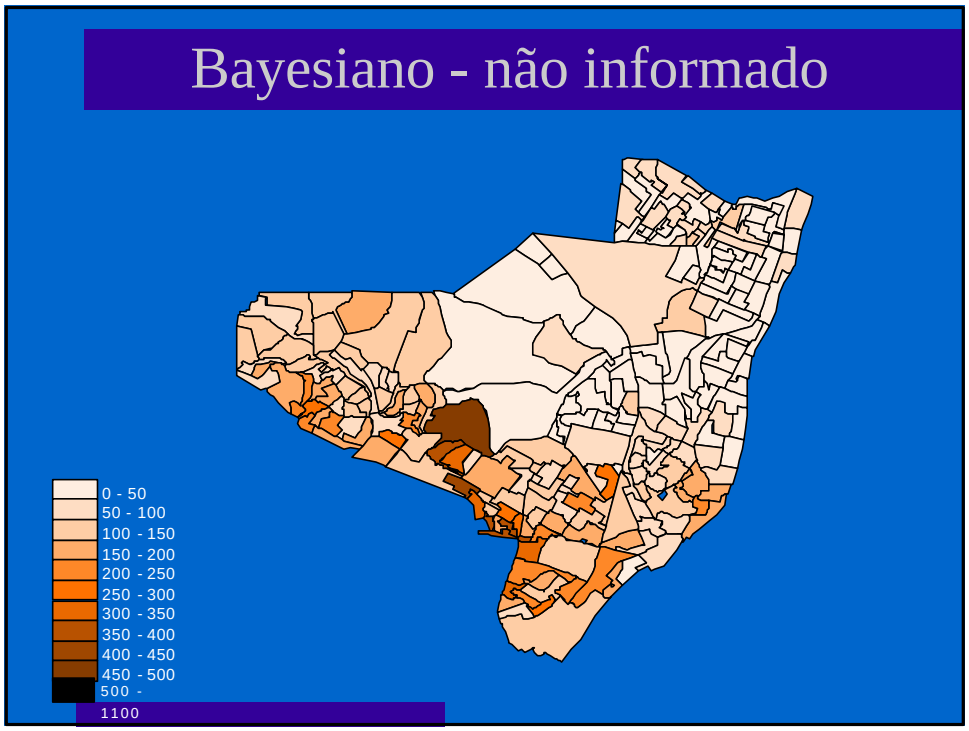


Ex: Hanseníase em Olinda

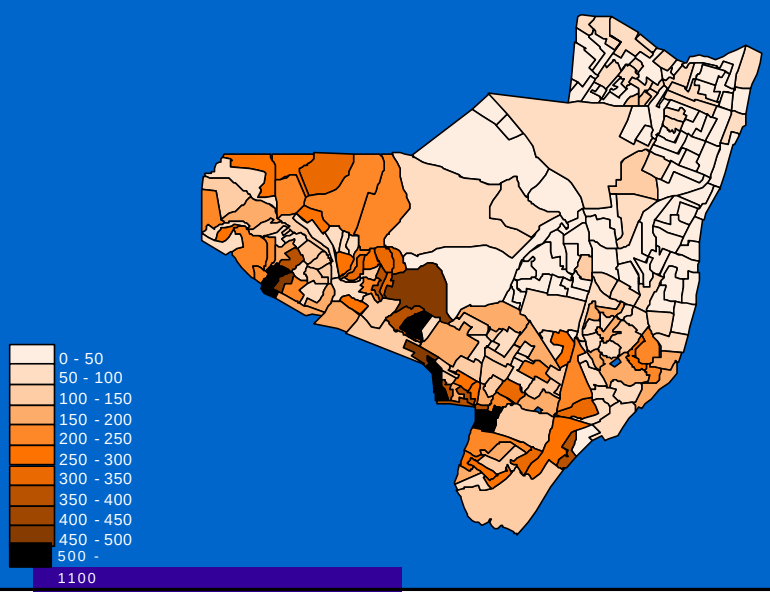
Indicador de renda



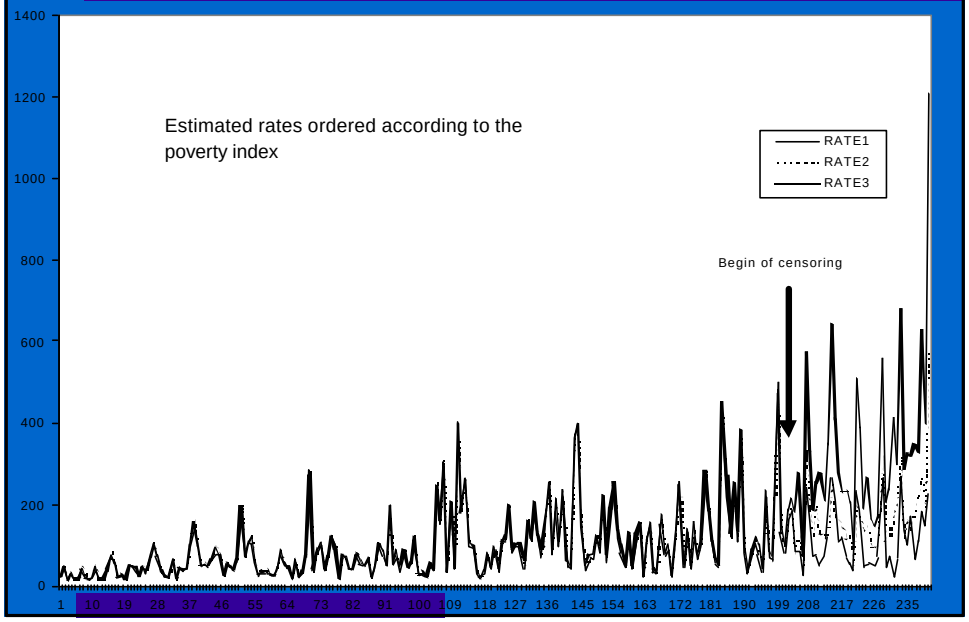

Alisamento Bayesiano
0 - 50
50 - 100
100 - 150
150 - 200
200 - 250
250 - 300
300 - 350
350 - 400
400 - 450
450 - 500
500 - 1100


Bayesiano - não informado
0 - 50
50 - 100
100 - 150
150 - 200
200 - 250
250 - 300
300 - 350
350 - 400
400 - 450
450 - 500
500 - 1100

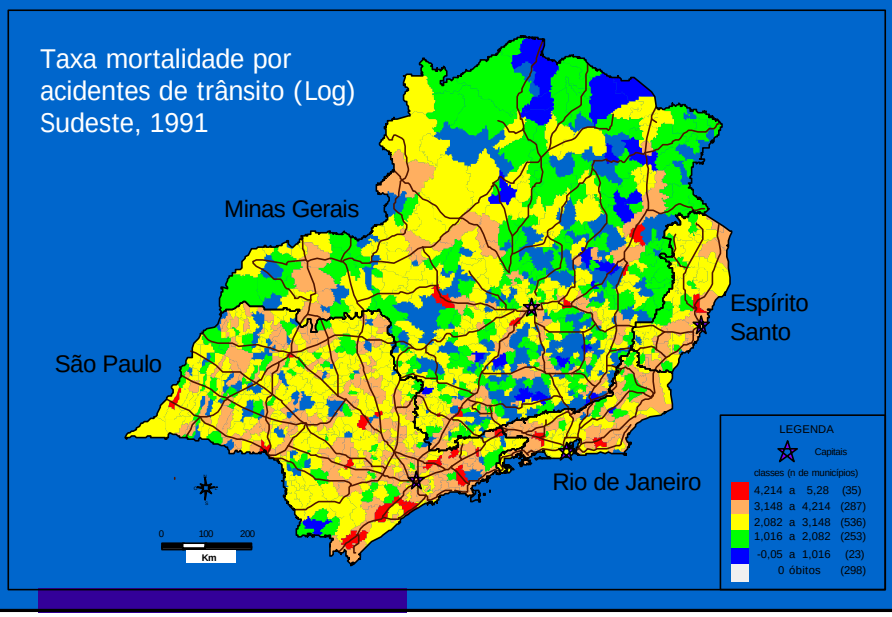
Bayesiano - Dado Censurado



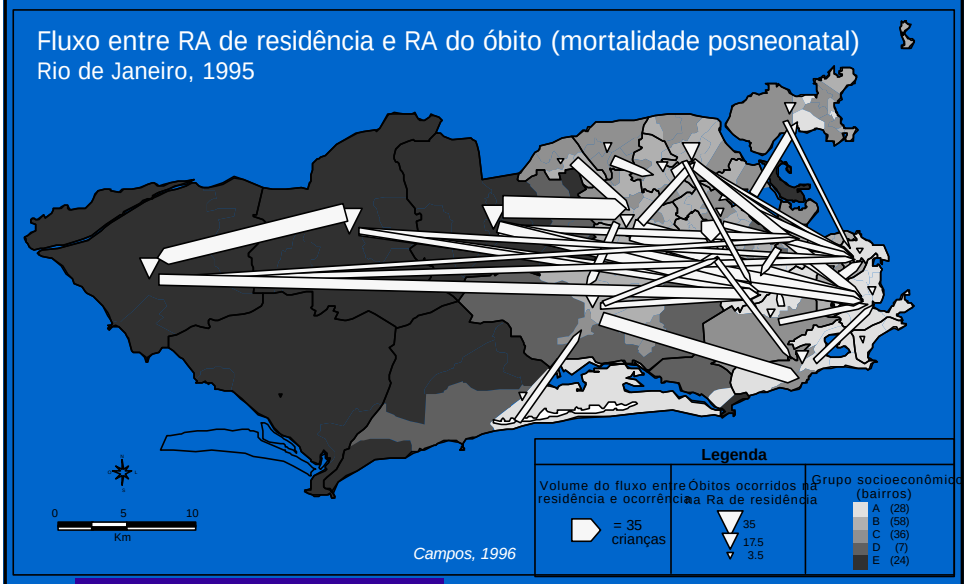
Ex: Bayesiano



Linhas



Fluxo

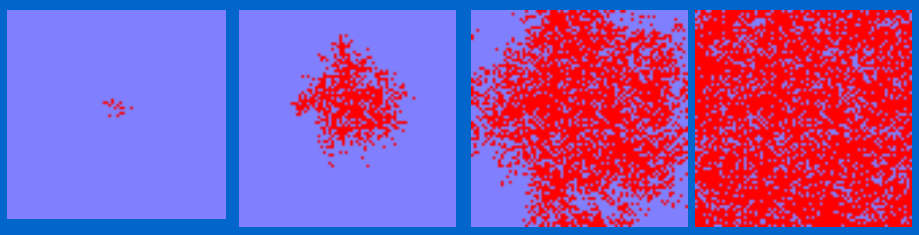


Modelos Espaço-Tempo

Modelos Espaço-Tempo

- O interesse na distribuição espaço-temporal esteve presente desde os primórdios da epidemiologia
- Entretanto só recentemente tem sido utilizadas técnicas que permitem a incorporação das dimensões **tempo**, **espaço** e somente **muito recentemente** a interação **espaço-tempo**.

Modelos de difusão das doenças



Simulação por multi-agentes

Difusão Espacial

- Difusão - dispersar a partir de um centro, disseminar, propagar, espalhar.
- Em geografia utiliza-se 2 conceitos
  - Difusão por expansão - quando um material, uma informação, etc... se espalha de um local p/ o outro, permanecendo (ou até mesmo se intensificando) na região inicial. Ex: doença transmissível
  - Difusão por realocação - quando o material difundido deixa a área original e se move p/ novas áreas. Ex: Movimentos migratórios

Difusão Espacial

- Difusão por expansão também pode ocorrer através de uma seqüência de classes locais - neste caso é denominado espalhamento hierárquico (ex: moda, bens de consumo), que tendem a passar por classes sociais diferentes e se irradia a partir das grandes metrópoles.
- Difusão em cascata é um caso particular onde a difusão sempre se dá dos grandes centros p/ os menores.

Difusão Espacial

- Na geografia houve um grande interesse por modelos de difusão desde a década do inicio do século, e se intensificou a partir da década de 50 com o trabalho pioneiro de Hägerstrand. Um dos principais interesses dos geógrafos eram os modelos de difusão de inovações tecnológicas.
- Na epidemiologia por outro lado, devido a influencia de Ross e Hamer, os modelos compartimentais foram mais utilizados.

Espacializando a difusão das doenças

- O principal desafio é como introduzir as dimensões **espaço e tempo** na modelagem da difusão das doenças transmissíveis .
- Qual o impacto do espaço e da interação espaço-tempo?
- Eles são capazes de mudar parâmetros inferidos para a interação das populações envolvidas?
- São capaz de trazer novidades na análise e interpretação de resultados?
- Quais os possíveis modelos e maneiras de incorporar o **espaço e tempo** ?

Processo de difusão das doenças

- Sob o ponto de vista da Ecologia, mais especificamente da dinâmica de populações uma doença transmissível, é o resultado da interação entre pelo menos 2 espécies (parasita X hospedeiro).
- Também na ecologia os modelos de crescimento populacional, interação entre espécies, competição etc... apesar de utilizarem o tempo em suas equações de crescimento (Lotka-Volterra) não incorporam a dimensão espaço.
- A introdução do espaço nos modelos, mesmo com uma única espécie, e capaz de alterar a inferência a respeito da dinâmica desta espécie.

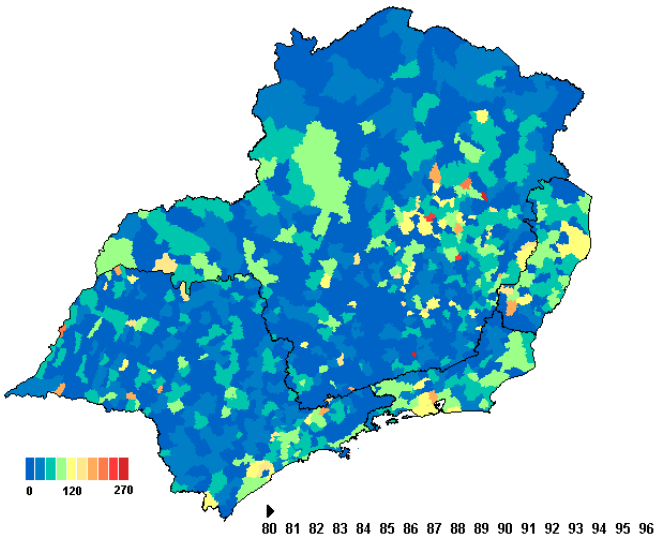
Modelos de difusão de doenças

- No início do século XX surgiram os primeiros modelos de transmissão de doenças , que consideravam que o curso de uma epidemia deveria depender do número de suscetíveis, das taxas de contato entre os indivíduos infectados e suscetíveis e do número de infectados.
- A partir daí diversos modelos determinísticos foram empregados modelando parâmetros de epidemias.
- À medida em que o interesse se volta para pequenas populações e eventos raros, foram introduzidos modelos estocásticos.

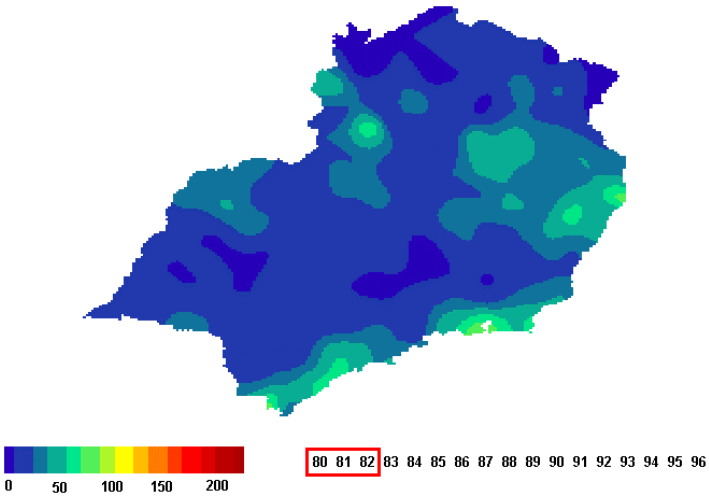
Modelos de difusão das doenças

- No entanto, a medida que se adiciona variáveis (por exemplo, estrutura etária, populações de vetor) esses modelos se tornam muito complexos, dificultando ou impossibilitando uma solução analítica.
- Neste contexto pode-se empregar métodos numéricos ou simulações na estimação parâmetros, no entanto a inclusão do espaço e tempo inviabilizam a convergência mesmo p/ essa classe de modelos
- Os avanços recentes na modelagem espaço-temporal empregam modelos bayesianos (MCMC) espaço-temporais.

Modelagem Estatística Espaço-temporal



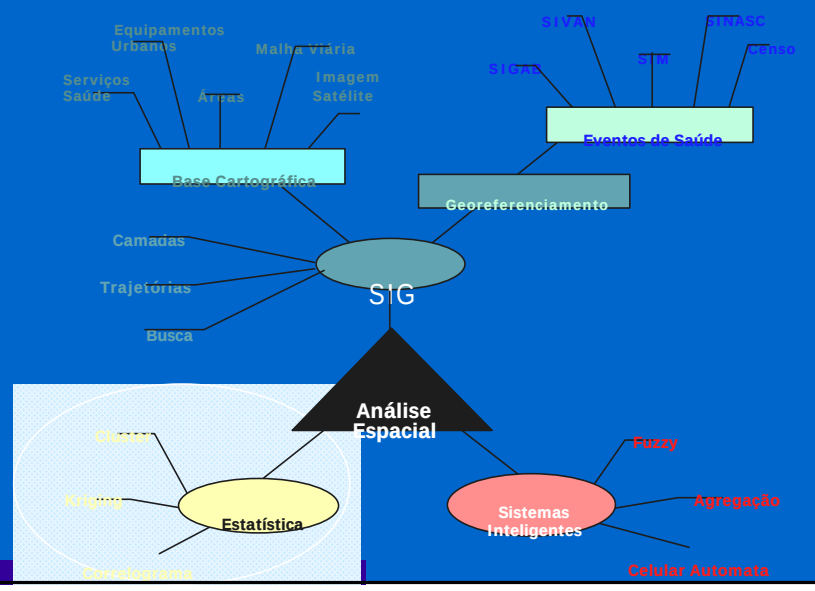
Modelagem Estatística Espaço-temporal



Análise de Dados Espaciais em Saúde:

avanços recentes no Brasil

Modelo Integrado de Análise Espacial



Geoprocessamento e SIG - definições

Geoprocessamento - processamento digital de dados geográficos, incluindo qualquer tipo de manipulação de informações geográficas, da imagem de satélite à restituição de fotos aéreas.
SIG - sistema digital de captura, armazenamento, recuperação, análise e apresentação de dados espaciais.
Georeferenciamento - localização de informações não geográficas (por ex. eventos em saúde) em base cartográfica. Pode ser feito em diversas escalas - município, bairro, estado, setor censitário - e de diversas formas.

SIG - Sistema de Informações Geográficas

- Sistema constituído de "*hardware*" e "*software*" que permite armazenar, gerenciar e editar bases cartográficas e acoplá-las a dados não gráficos, realizando análises espaciais e apresentando graficamente os resultados.
- A estrutura de dados gráficos armazena informações sobre localização, escala, dimensão, etc. A estrutura de dados não-gráficos informa sobre os objetos ou ligações entre eles.
- A base gráfica pode ser vetorial ou matricial (“raster”).
- Estruturas vetoriais podem ser armazenadas de duas formas:
  - espaguete - todas as feições do mapa são arquivadas como sucessão de pontos (lista de coordenadas), linha a linha (típica de CAD). Este tipo não permite reconhecer a relação espacial entre os objetos
  - topológica - permite mais facilmente a análise de dados, pois armazena, além do componente locacional e dos atributos dos dados, informações sobre a interligação entre os objetos.

Análise espacial - funções do SIG

- Operações características de SIGs:

interseção de linhas - análise de redes, otimização de rotas;
pontos pertencentes a áreas - georeferenciamento, densidade;
operações de camadas - tratamento simultâneo de diversas informações;
buffer - área de influência.

- Atributos dos dados no SIG:
  - vizinhança
  - conectividade
  - distância.

Cartografia - produção de mapas

- A elaboração de mapas envolve diversas técnicas, mais ou menos sofisticadas, desde o simples esboço rascunhado por um agente de saúde na Região Amazônica, até técnicas sofisticadas utilizando satélites e arquivamento digital.
- O processo de mapeamento formal começa com o sensoriamento remoto - informação do terreno coletada à distância. Dois tipos de métodos são usados: fotografias aéreas e imagens de satélites.
- As imagens obtidas com qualquer método são interpretadas a partir de outras informações, de contexto (por exemplo, uma linha com estruturas de casas dos dois lados deve ser uma rodovia, e não um rio ou fenda geológica) ou de campo. O trabalho de campo pode ser feito através de inspeções pontuais, onde se visita determinado ponto para obter informações específicas, ou sistematicamente por amostragem. Dependendo do detalhamento da informação, o trabalho de campo pode ser uma das atividades mais custosas de toda a elaboração dos mapas.

Cartografia - aerofotogrametria

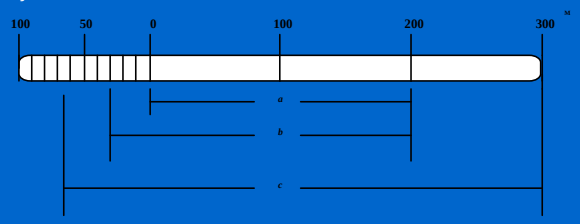
- E a tecnologia mais tradicional de mapeamento sistemático, e envolve vôos especiais quanto à trajetória e altitude, onde são tiradas fotos com câmeras e angulação apropriadas, em sucessão, com distância conhecida e área de superposição definida rigorosamente. Este tipo de técnica permite recuperar o relevo, através da comparação entre as imagens de superposição a partir de ângulos diferentes e conhecidos, de forma semelhante à visão estereoscópica do olho humano. Os mapas elaborados a partir destes dados podem ter grande resolução, e, como são fotografias utilizam toda a faixa visível simultaneamente, do azul ao vermelho, também o infra-vermelho próximo.
- É a técnica que permite o desenho de cartas urbanas com grande resolução (1:2.000). O custo e tempo de execução são muito elevados.

Cartografia - imagem de satélite

- As imagens de satélite são obtidas a partir de equipamentos colocados em órbita a partir de 640Km., que enviam uma sucessão de imagens digitais para uma rede de recepção distribuída em diversos pontos de sua trajetória. O satélite tem uma órbita absolutamente regular, passa sobre o mesmo ponto exatamente à mesma hora, com o sol à mesma altura e sombreamento mantido, garantindo uma repetitividade rigorosa.
- Diversos satélites cobrem o Brasil: os mais usados são o Landsat5 e o Spot.
- As imagens são gravadas em diferentes canais de acordo com o comprimento de onda, incluindo além da faixa visível uma de infra-vermelho (calor). A resolução varia de 30m. no Landsat 5 à 10m. no Spot, quando utilizadas várias bandas. Assim, considerando que uma rua normal tem 17m., as imagens do Landsat não permitem mapear cidades, exceto como grandes manchas e suas principais vias. Já existem satélites comerciais com resolução de 1m.

Bases gráficas - escala

- Razão entre medidas no mapa e medidas reais:
  - 1:2.000 1 cm. no mapa equivale a 2.000 cm. no local (20m.)
  - escala **pequena**: razão é pequena, logo quando o denominador é grande
- A escala pode ser apresentada sob forma numérica ou gráfica:
  - 1:2.000 - representação numérica
  - escala gráfica:



- **Mapa**: abrange grandes extensões, portanto em escalas pequenas
- **Carta**: regiões menores, escalas médias (carta do Estado da Bahia)
- **Planta**: regiões ainda menores (por ex. 1:2.000), escalas grandes (planta de uma cidade)

Desenvolvimento recente da análise espacial

Como as técnicas necessárias à análise espacial têm origem em diferentes disciplinas e os dados são produzidos por diversas fontes, a construção de projetos de análise espacial depende da constituição de um amplo fórum de instituições, onde o setor saúde é essencialmente usuário das informações, ainda que possa representar papel articulador. Nacionalmente, o Comitê Técnico Interinstitutional de Geoprocessamento e Dados Espaciais (CTI-GEO) da Rede Interagencial de Informações para a Saúde (RIPSA) vem trabalhando no sentido de articular os diversos setores de produtores e usuários de dados espaciais.

Os avanços recentes estão relacionados à construção de bases cartográficas, à incorporação do georreferenciamento aos Sistemas de Informação em Saúde, à possibilidade de utilização de novos (para a Saúde) SIGs e ao crescimento da capacidade analítica.

Bases Cartográficas

- Disponibilização da malha municipal oficial do Brasil:
  - CD-ROM do DECAR/IBGE;
  - no formato do TabWin (http://www.datasus.gov.br)
- Digitalização dos mapas topográficos no IBGE
- Base Territorial do Censo 2000 – Vertente Urbana" - criação de biblioteca CAD com a malha censitária dos 1058 municípios com mais de 25.000 habitantes (75% da população brasileira) pronta em dezembro de 1999:
  - escala 1:5.000;
  - com hidrografia básica, quadras, toponímia, principais edificações
- Malha censitária rural praticamente pronta, porém não há integração entre as diferentes escalas.
- Disponibilização a ser discutida em conjunto MS/IBGE.

Georreferenciamento: área mínima

- Características necessárias:
  - Maior homogeneidade possível, quanto a situação (rural/urbana), ocupação do espaço urbano (favela/área urbanizada) e demais indicadores socioeconômicos
  - Continuidade das feições espaciais, sem interrupção por acidentes geográficos ou construções
  - Continuidade histórica
- Propôs-se o setor censitário do Censo 2000, como tendo potencialmente estas características.
- Outros níveis de agregação de dados podem ser construídos, através da combinação de setores censitários.
- O uso dessa unidade espacial mínima de referência permitirá também o acompanhamento histórico dos municípios desmembrados, sendo necessário para isso que se busque manter os limites dos setores censitários.

Georreferenciamento - SIS

- A base cadastral de endereços para georreferenciar os dados dos SIS para setor censitário será o Cadastro de Segmentos de Logradouros por Setor Censitário, componente da base territorial do Censo 2000.
- Os sistemas de informações de bases nacionais deverão permitir o georreferenciamento para setor censitário, utilizando este cadastro e formato de entrada de endereço compatível com o fornecido/acordado com o IBGE.
- Este formato poderá ser alterado pelos municípios que dispuserem de outros cadastros e métodos de localização. Por isso, o DATASUS incorporará o georreferenciamento sob a forma de um módulo separado, disponibilizando, caso solicitado, o programa fonte.
- O teste deste módulo será feito possivelmente em Campinas, Porto Velho e Goiânia

Georreferenciamento - SIS

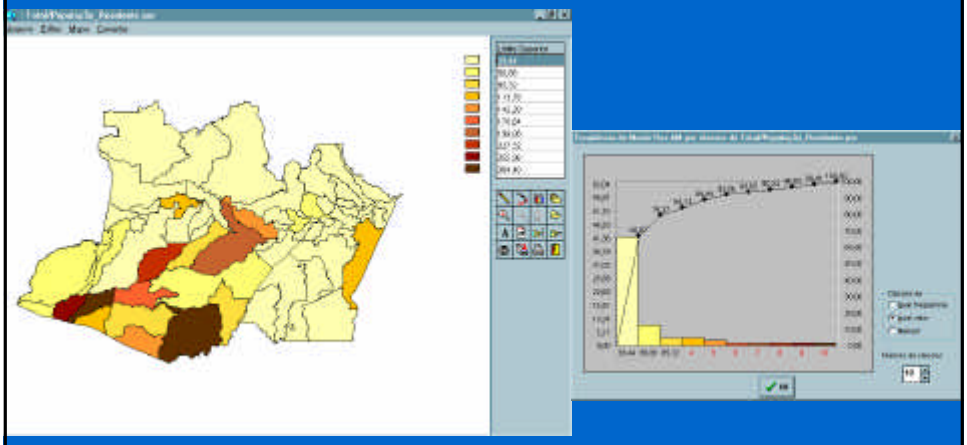
- Considera-se que georreferenciar os SIS para setor censitário utilizando o cadastro do IBGE será um avanço, ainda que a utilização de cadastro localmente gerados tenha provavelmente melhores resultados. Entre os problemas considerados, que deverão receber particular atenção estão:
  - Atualização apenas decenal (ou nas contagens rápidas no meio do período)
  - Incompatibilidade deste sistema com as formas de endereçamento usadas pela população de áreas faveladas, de expansão, rural e invasões
  - Qualidade precária de preenchimento dos endereços nos registros de saúde

Georreferenciamento: soluções locais

- Atualização dos cadastros localmente, em parceria com o IBGE, e utilizando o potencial do setor saúde e sua relação com as comunidades.
- Desenvolvimento de métodos de georreferenciamento em áreas faveladas (e similares) através de identificação de grupamentos de setores favelados contíguos com o mesmo endereço urbano de referência – entrada da favela, associação de moradores, comissão de luz.
- Integração SIG e o Programa de Agentes Comunitários de Saúde (PACS)/Programa de Saúde da Família (PSF), aproveitando as informações geradas por estes programas na localização de populações de risco - a experiência do Juá/Caruaru.

Software - mapeamento básico

- A função de cartograma desenvolvida no TAB-WIN (DATASUS) possibilita a análise exploratória dos dados dos SIS.



Software - análise espacial

- Considerando o alto custo dos aplicativos comerciais de mapeamento e a necessidade de desenvolver nos municípios habilidades de análise espacial, se desenvolvem parcerias com autores de software não comerciais:
  - SPRING - INPE (http://www.dpi.inpe.br/)
  - SIG-EPI - COPPE/UFRJ
- Além disso alguns software estatísticos de domínio público tem funções de análise espacial desenvolvidas:
  - R (http://www.ci.tuwien.ac.at/R/)
  - WIN-BUGS (http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.shtml)

Análise espacial

- Potenciais inovações
  - mapas animados - espaço & tempo;
  - GAM (geographic analysis machine), etc.
- Integração estatística/SIG
  - Exportação automática de camadas do SIG para o software estatístico e importação dos resultados da análise;
- Métodos inteligentes e SIG
  - Lógica difusa (fuzzy);
  - Cellular automata;
  - Técnicas de otimização (redes neurais, algoritmos genéticos, simulated annealing).

CONASEMS e RIPSA

Visando orientar a utilização dos recursos dos Sistemas de Informações Geográficas - SIG - para a gestão da saúde, foi lançada esta publicação voltada para os Secretários Municipais de Saúde.

Outro documento, mais completo, incluindo modelos de licitação para aquisição de bases cartográficas está em fase de preparação.

Recursos - Internet

- lista de discussão sobre análise de dados espaciais em saúde:
  - http://www.ensp.fiocruz.br/servicos/ades-l.html
- lista nacional de discussão sobre geoprocessamento:
  - http://www.lampada.uerj.br/fgeorj
- lista da OMS:
  - HEALTH-GIS@who.ch
- páginas para navegar:
  - http://www.geocities.com/Tokyo/Flats/7335/medical_geography.html
  - http://curie.ei.jrc.it/ai-geostats.htm
- CTI-GEO da RIPSA:
  - e-mail: georipsa@procc.fiocruz.br
  - http://www.procc.fiocruz.br/~marilia/

Marilia Sá Carvalho
Escola Nacional de Sáude Pública/FIOCRUZ
http://www.procc.fiocruz.br/~marilia/
marilia@procc.fiocruz.br

Oswaldo Gonçalves Cruz
Programa de Computação Científica/FIOCRUZ
http://www.procc.fiocruz.br/~oswaldo/
oswaldo@procc.fiocruz.br

"Análise de dados espaciais em saúde"

Navegue:
http://www.procc.fiocruz.br/ades/

Bibliografia

As principais referências para **estudos ecológicos** são:

Editorial (1994), *AJPH*, 84(15):715-716

SCHWARTZ (1994), The fallacy of Ecological fallacy: the potencial misuse of a concept and the consequences. *AJPH*, 84(15):819-824

SUSSER (1994), The logic in ecological: I The logic of analysis. *AJPH*, 84(15):825-829

SUSSER (1994), The logic in ecological: II The logic of design. *AJPH*, 84(15):830-835

EVANS – 1° Capítulo do livro *Why are some people healthy and others not ?*

ROSE (1985) Bol. Epidemiol. OPS, 6(3):1-8, 1985 ou Sick Individuals and Sick Population. Int. J. Epidemiology 14:32-38

Morgenstern, H. (1998) Ecologic Studies. In Rothman, K.J. & Greenland, S. *Modern epidemiology*, 2ª Edição

Para análise de **séries temporais**:

Martinez-Schnell, B. & Zaidi, A. (1989) Time series analysis of injuries, *Statistics in Medicine*, 8:1497-1508.

Morettin,P.A. & Toloi, C.M.C. (1987) Previsão de Séries Temporais. Atual Editora, 2ª Edição.

Morettin,P.A. & Toloi, C.M.C. (1987) Séries Temporais. Atual Editora, 2ª Edição. (versão reduzida, sem modelagem ARIMA)

Diggle, P. (1990) Time Series : A Biostatistical Introduction. Oxford Statistical Science Series, No. 5 (ISBN: 0198522266)

Para análise **espacial**:

Bailey, T.C. & Gatrell, A. (1995) *Interactive Spatial Data Analysis*, Longman Scientific & Technical.

Cressie, N.A.C. (1991) *Statistics for Spatial Data*, John Wiley, Chichester.

Haining, R. (1990) *Spatial Data Analysis in the Social and Environmental Sciences*, Cambridge University Press.

Isaaks, E. H. and Srivastava, R. M. (1989) *An Introduction to Applied Geostatistics*, Oxford University Press, Oxford.

Câmara, G (org.). Geoprocessamento: Teoria e Aplicações. Livro em preparação, disponível “on line” em http://www.dpi.inpe.br/ ~gilberto/livro/

Os trabalhos nacionais apresentados e os dados utilizados nos exercícios foram, na maioria, provenientes de trabalhos de tese de diversos pesquisadores:

Eleonora d’Orsi (1996). *Perfil de nascimentos e condições sócio-econômicas no Município do Rio de Janeiro: uma análise espacial*. Tese de mestrado aprovada pela Escola Nacional de Saúde Pública/Fundação Oswaldo Cruz. (orientação de Marilia Sá Carvalho e Maria Zulmira de Araújo Hartz).

Enirtes Caetano Prates de Melo (1996) *Heterogeneidade do padrão da doença isquêmica do coração na Região Sudeste - Brasil: mortalidade e utilização de serviços hospitalares*. Tese de mestrado aprovada na ENSP/FIOCRUZ. (orientação de Marilia Sá Carvalho e Maria Zulmira de Araújo Hartz).

Marilia Sá Carvalho (1997) *Identificação de áreas segundo risco: uma análise espacial*. Tese de doutorado aprovada pela COPPE/UFRJ, orientação de Flávio F. Nobre. (http://www.procc.fiocruz.br/~marilia/)

Mirian Carvalho de Souza (em curso) - *O problema da escala na análise de dados espaciais - aplicações em epidemiologia*. Tese em andamento na ENSP/FIOCRUZ, orientação de Marilia Sá Carvalho e Oswaldo G. Cruz.

Oswaldo Gonçalves Cruz (1996) Homicídios no Estado do Rio de Janeiro: análise da distribuição espacial e sua evolução. Tese de mostrado aprovada pela FSP/USP, orientação de Maria Lúcia Lebrão. (http://www.procc.fiocruz.br/~oswaldo/)

Simone Maria dos Santos (1999) *Análise da distribuição espacial das nmortes violentas em Porto Alegre, no ano de 1996, e do seu contexto social*. Tese de mestrado aprovada pela Escola Nacional de Saúde Pública/Fundação Oswaldo Cruz. (O projeto foi selecionado em concurso promovido pela OPAS para financiamento de projetos de pos-graduação na América Latina e Caribe - orientação de Marilia Sá Carvalho e Christovam Barcellos).

Tatiana Campos (1997) *Perfil de nascimentos e óbitos infantis: a busca da assistência*. Tese de mestrado aprovada pela Escola Nacional de Saúde Pública/Fundação Oswaldo Cruz. (orientação de Marilia Sá Carvalho e Christovam Barcellos)