Estatística Espacial Aplicada

Renato M. Assunção

LESTE - Laboratório de Estatística Espacial

Departamento de Estatística - UFMG

assuncao@est.ufmg.br

http://www.est.ufmg.br/~assuncao

Introdução

Instrutor: Renato Assunção

Professor da UFMG, Departamento de Estatística.

Coordenador do Laboratório de Estatística Espacial - LESTE

Vice-diretor do CRISP - Centro de Estudos de Criminalidde e Segurança Pública

Público-Alvo:

- Alunos de pós-graduação em Anáalise e Modelagem de Sistemas Ambientais do IGC
- Alunos de pós-graduação, exceto estatística
- Alunos de graduação de estatística e atuária

Aulas toda quarta-feira, de 13:30 as 17:30, sala 1019 no ICEx

Plano da Apresentação

Introdução genérica

Tipologia dos dados espaciais

GIS e algoritmos geométricos

Dados de área

Dados de processos pontuais

Dados de superfícies aleatórias

Dados de interação espacial

Um curso extra muito bom

http://www.dpi.inpe.br/cursos/ser301/

O que é Estatística Espacial?

Toda observação possui referência temporal e espacial.

- Dados obtidos por entrevista sobre indivíduo que vive em Belo Horizonte.
- Dados sobre certo município de Minas.
- Dados sobre rebanho em certa região do Pantanal.
- Dados sobre extração de minério de ferro numa mina nos arredores de Belo Horizonte
- Dados sobre telefonemas entre municípios

Muitos estudos não fazem uso da informação espacial. NO entanto, em alguns casos, essas referências espaciais são importantes na análise.

Estatística Espacial é o conjunto de métodos de análise de dados em que a localização geográfica é usada explicitamente na análise.

Est Esp só se USAR o espaço

Assim, não basta que o dado seja espacial, pois todos os dados, de uma forma ou de outra, possuem uma referência geográfica.

- Por exemplo, a regressão linear do nível de arrecadação de ICMS versus a renda per capita municipal NÃO é parte de estatística espacial.
- Embora os municípios possuam localização espacial, esta localização não é usada na regressão.

O que determina se algo faz parte da estatística espacial é uma propriedade do método de análise, e NÃO do dado utilizado na análise.

Estatística Espacial: quando usar

Se todo dado estocástico possui referência geográfica ... é sempre necessário usar estatística espacial?

Resposta: Deve ser usada se existirem perguntas ou hipóteses sobre o mecanismo gerador dos dados que envolvam alguma característica espacial ou geográfica.

'E ineficiente não usá-la? (Veremos o significado de eficiência mais tarde)

Resposta: Deve ser usada se a correlação espacial aparece como ruído (nuisance) em modelo usual de regressão causados por efeitos de variáveis não-observadas, efeitos de transbordamento (spill-over) por causa do mismatching entre unidades geográficas de mensuração e as unidades geográficas onde o fenômeno ocorre.

Exemplos: Crimes

Crimes não acontecem totalmente ao acaso. É preciso um ofensor, uma vítima potencial e uma oportunidade. Existem grandes diferenças no risco de ser vítima de um crime dependendo da idade, do sexo, da hora do dia, dia da semana, mês no ano, etc.

Existem também grandes diferenças geográficas dentro de uma cidade. Estas diferenças dependem do tipo de crime: crimes contra o patrimônio atingem mais as áreas ricas enquanto que crimes contra a pessoa atingem mais as áreas pobres.

Todos os dados de crimes registrados pela Polícia Militar dentro de Belo Horizonte e Juiz de Fora hoje em dia são georeferenciados ao nível da localização exata (coordenadas latitude-longitude) do evento.

Exemplo: Crimes em BH

Homicídios ocorridos em BH em 1997, região central

Exemplos: Linchamentos raciais nos EUA

Os linchamentos de negros no sul dos EUA nas décadas de 40-50 seguiam algum padrão no espaço e no tempo?

Duas teorias para explicar este comportamento violento de massa:

- um efeito de contágio (real ou aparente). Isto é:
  - um linchamento ocorrendo aqui e agora estimula ou está associado à ocorrência de outros linchamentos nos arredores nos momentos seguintes
- Um efeito de associação negativa. Isto é:
  - um linchamento inibe a ocorrência de outros linchamentos nos arredores nos momentos seguintes
  - a reação seria de procurar não dar motivos para violência adicional ou haveria um maior controle dos mais violentos

Exemplos: Linchamento de Negros nos EUA

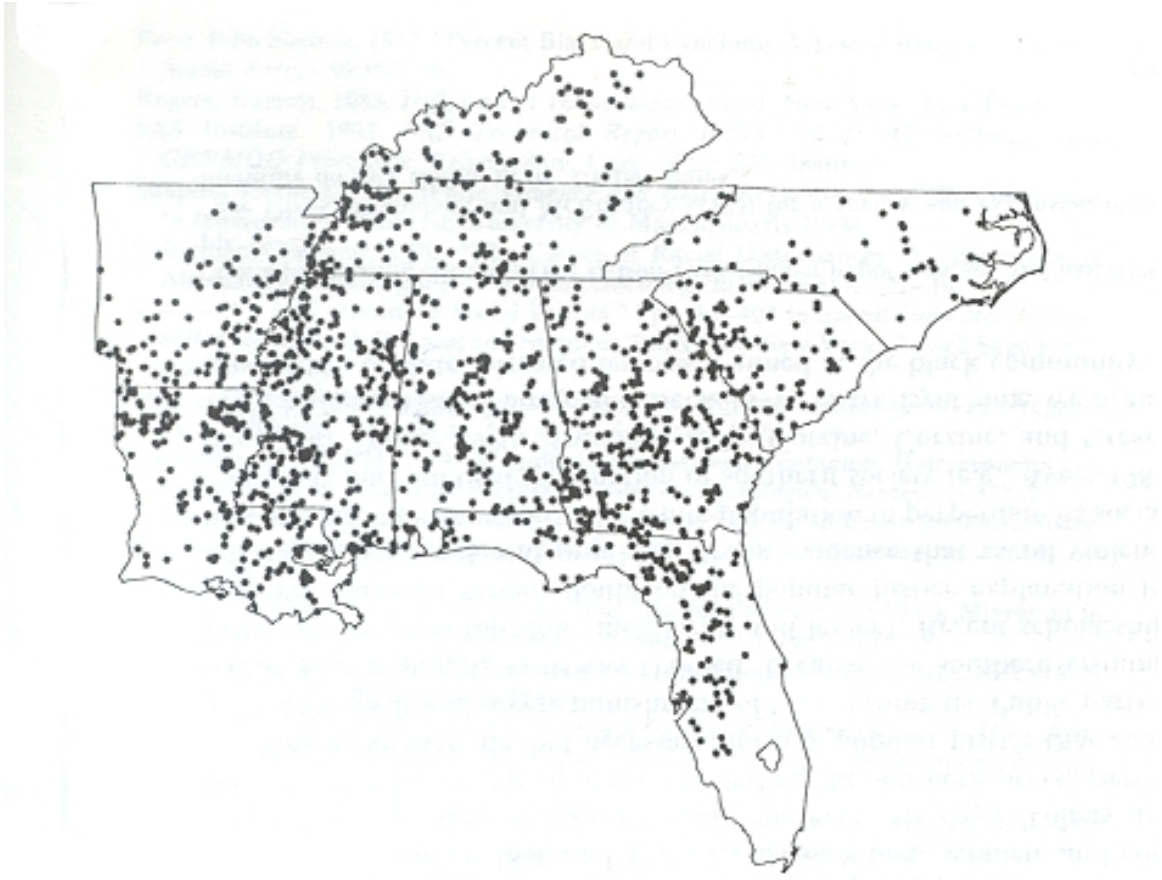


FIG. 1.—Geography of lynchings of blacks by white mobs, 1890–1919

Centróides dos condados onde ocorreram linchamento. REF??

Exemplos: Processos Ecológicos

Processos de colonização por plantas de áreas devastadas.

Árvores tendem a inibir ou a estimular a presença de outras árvores ao seu redor ?

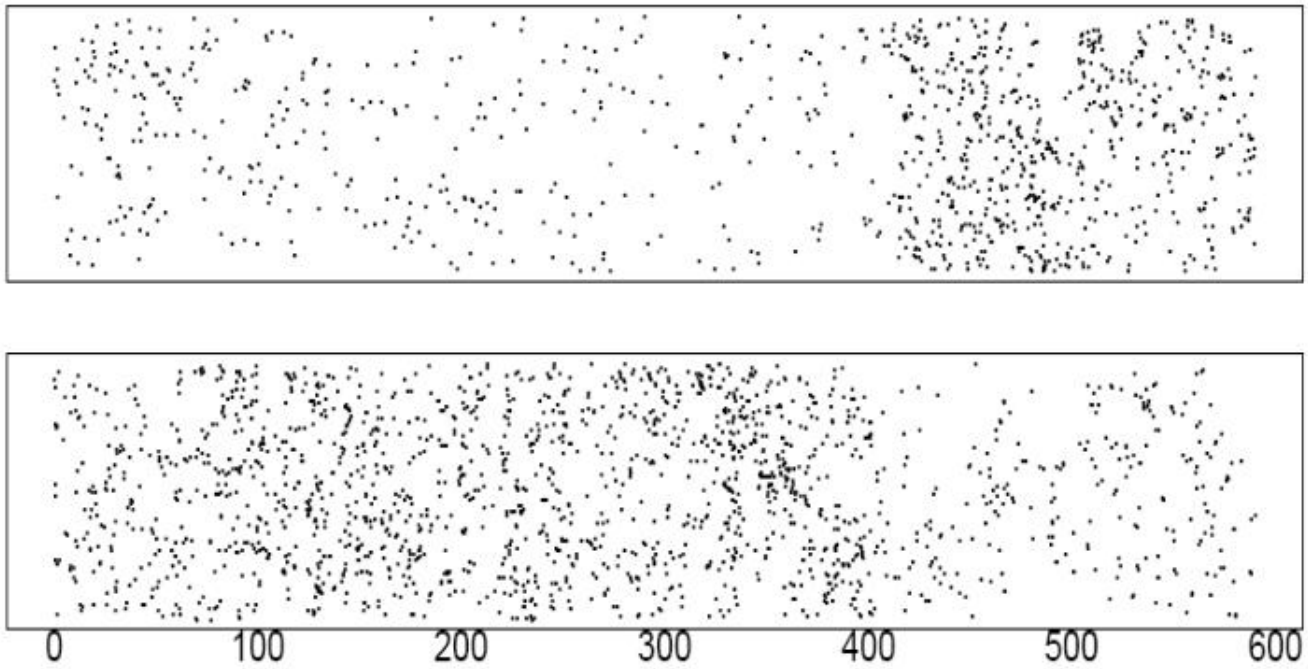
Se existe competição entre as plantas, até que distância esta competição pode alcançar ?

Este padrão espacial depende da idade da floresta ?

Árvores pequenas (mais jovens) tendem a estar próximas de árvores grandes (mais velhas) ?

E se as espécies são diferentes, como é o seu relacionamento ? De competição também ?

Plantas adultas (acima) e Plântulas (abaixo)



Note a interação óbvia entre elas: onde há muita planta adulta, poucas plântulas aparecem. REF??

Exemplos: Epidemiologia Espacial

A distribuição dos casos de uma doença forma um padrão no espaço ?

Descrevendo o desenvolvimento de uma epidemia no espaço e no tempo: sugere formas de controlar e combater.

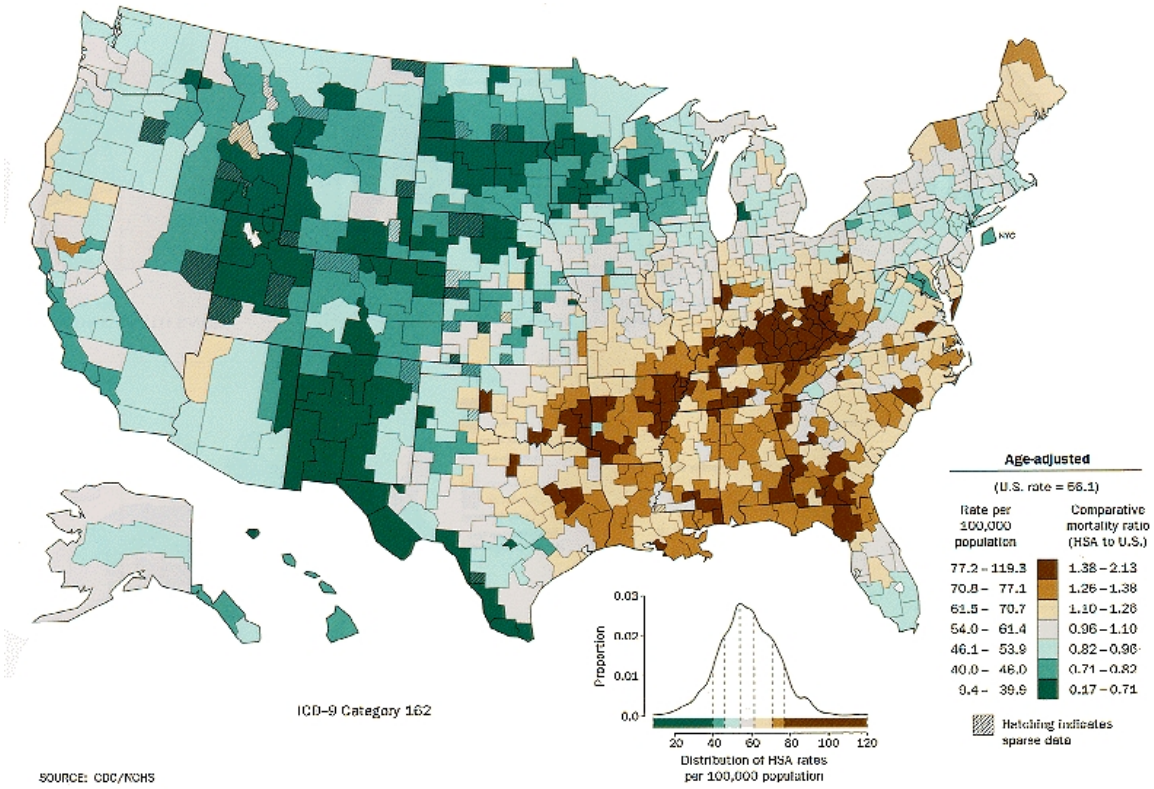
Caracterizando a localização de serviços de saúde: a demanda pelos serviços está sendo atendida adequadamente ?

Poderia ser feita uma nova alocação geográfica de recursos que seja mais eficiente ?

Há alguma associação entre a localização dos casos de uma doença e a posição de alguma fonte suspeita (rio, estação nuclear, fábrica,..) ?

Exemplos: Câncer de Pulmão nos EUA

Publicação do National Institute of Health / National Institute of Cancer



Mapa do Atlas de Mortalidade por Câncer de Pulmão nos EUA, 1950/1994

Exemplos: Geoestatística

Moura et al (2006) estudaram um surto de toxomoplasmose em Santa Isabel do Ivai, no sul do Paraná. A suspeita era que água era o modo de disseminação de Toxoplasma gondii.

Foi realizado um estudos de caso-controle.

Dois reservatórios de água serviam a cidade, cobrindo r egiões distintas.

Era significativamente mais comum que casos consumissem água do reservatório A e que consumissem mais sorvetes que controles.

Odds-ratio = 3.72 com p-valor = 0.016

Reservatório	Casos	Controle	Total
A	152	198	350
B	4	22	26
Total	156	220	376

Exemplos: Geoestatística

Como estimar o volume total de um depósito mineral numa região ?

Conhecemos apenas a densidade num pequeno número de amostras localizadas em alguns poucos pontos do terreno.

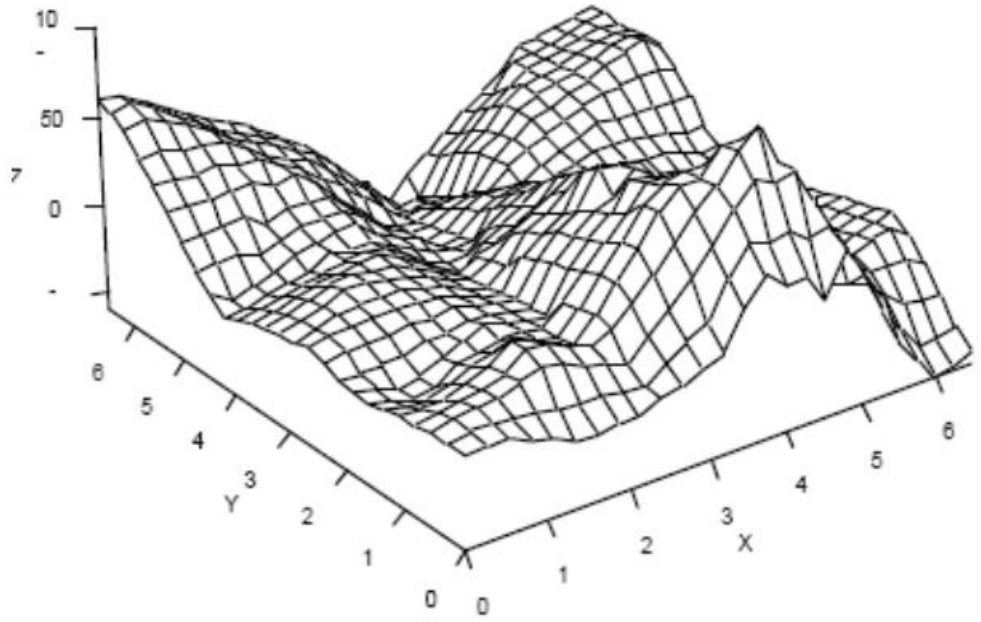
Como prever a precipitação pluviométrica (ou a temperatura) num dado ponto do mapa ?

Possuimos medições apenas em algumas poucas estações espalhadas pelo mapa.

Onde colocar uma nova estação de coleta de medições de forma ótima ?

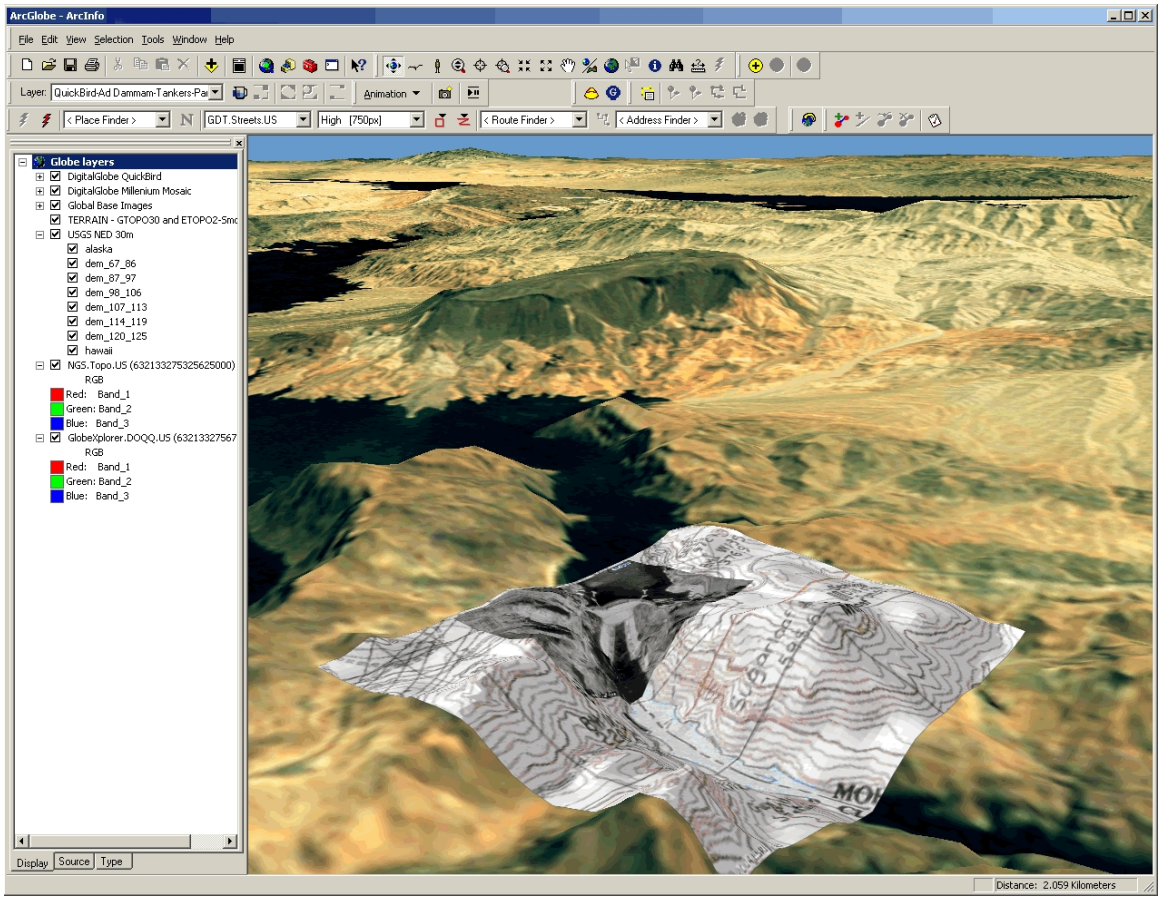
Exemplos: Reconstrução de depósitos minerais

Superfície de densidade de minério de ferro reconstruída por interpolação



Exemplos: Reconstrução - 2

Superfície de curvas de níveis superimposta à imagem do terreno



Exemplos: Espaço, não só geografia

Em um experimento para entender o câncer de colon, todos os animais foram expostos a um carcinoma.

Metade deles foram também expostos à radiação.

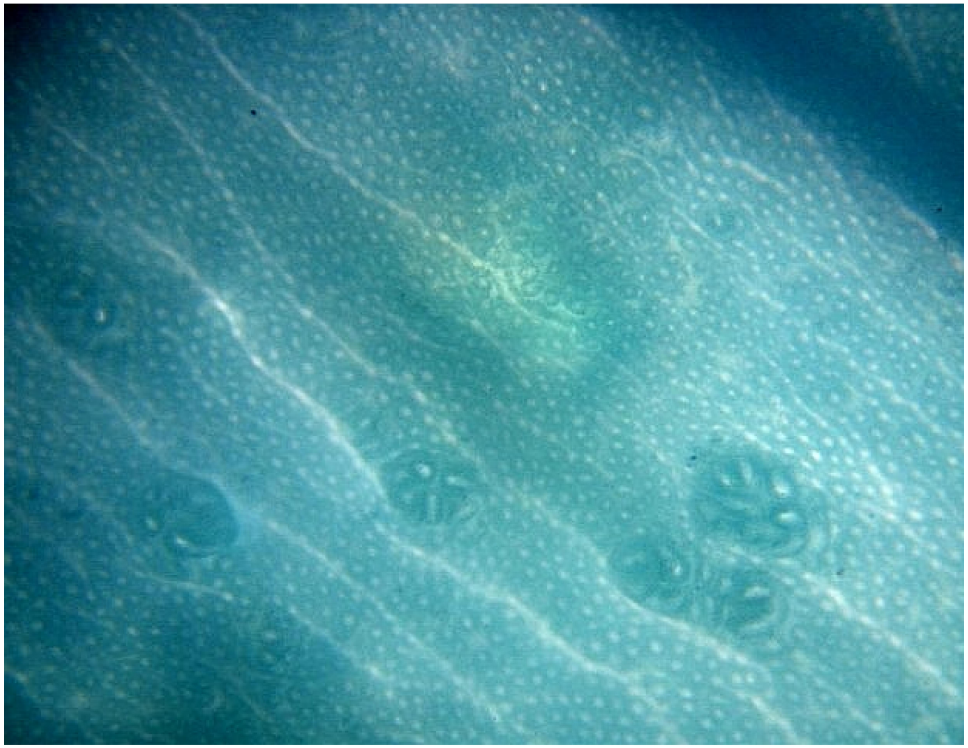
Espacialmente foi medida a existência de focos precursores de cânceres.

A questão de interesse é saber se as localizações desses focos estão espacialmente próximas.

Se sim, então os danos ao colon causado pelo carcinoma e pela radiação são localizados e devem ser tratado levando isto em conta.

Há diferenças nos padrões espaciais de irradiados e não irradiados?

Exemplos: Imagens



Exemplos: Imagens

Imagens de satélite ou fotográficas devem ser "limpas"para obter uma visualização melhor.

A partir das imagens, objetos devem ser identificados e "recortados".

Como fazer isto de forma automática e eficiente?

Exemplos: Espaço não geográfico

Chen e Conley (2001, Journal of Econometrics): A new semiparametric spatial model for panel time series

n agentes econômicos. Para cada agente i, uma série temporal X_{it}, $t = 1, \ldots, T$

Em cada instante t, os valores $X_{1t}, \ldots, X_{nt}$ das séries são correlacionados

Correlação entre séries depende da distância econômica entre os agentes.

Esta distância pode mudar no tempo.

Exemplos:

- *Agentes*: setores econômicos; *proximidade entre i e j*: distância euclidiana entre entre vetores de proporções dos inputs dos setores i e j
- *Agentes*: firmas; *proximidade entre i e j*: overlap das áreas de mercado
- *Agentes*: estados/países/municípios; *proximidade entre i e j*: volume de transações comerciais ou custo de transporte

Exemplos: interação/competição/redes sociais

Indivíduos interagem de formas variadas, entre as quais interações no mercado

Um conjunto de indivíduos que são os nós de uma rede

os arcos ou arestas da rede refletem as relações entre os indivíduos

Indivíduos fazem escolhas e agem a partir de um conjunto de alternativas

Existe incerteza sobre os ganhos obtidos de cada ação

Eles usam informação própria e informação obtida de seus vizinhos, os indivíduos ligados a eles de alguma forma.

Escolhem ação que maximiza utilidade individual

A estrutra da rede (SUA TOPOLOGIA) influencia as decisões individuais e sociais.

A topologia da rede induz distribuições de probabilidade que levam em conta essa configuração espacial de interrelações.

Exemplos: redes

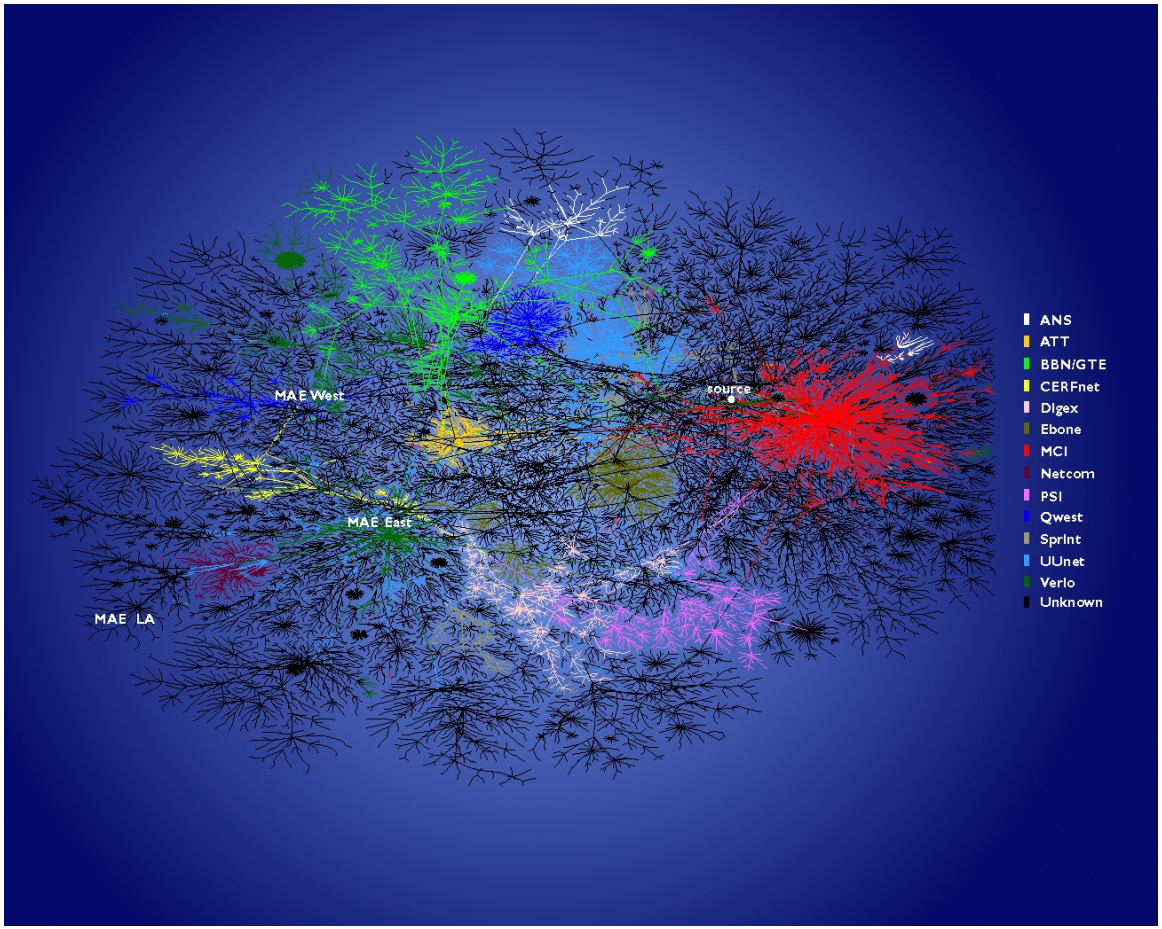
Escolha de produtos por consumidor:

- Decisão sobre que marca comprar
- não possui conhecimento completo sobre alternativas
- preço, características e ... informação de conhecidos

Inovação médica:

- Médicos decidem recomendar produtos sem conhecimento completo
- buscam informação na literatura profissional e de amigos
- ceteris paribus, os médicos mais conectados são aqueles que passam a recomendar produtos melhores mais rapidamente

Exemplos: WWW



- Conectividade da Internet: principais backbone ISPs (Internet Service Provider) coloridos separadamente

Exemplos: Sociedade Adolescente

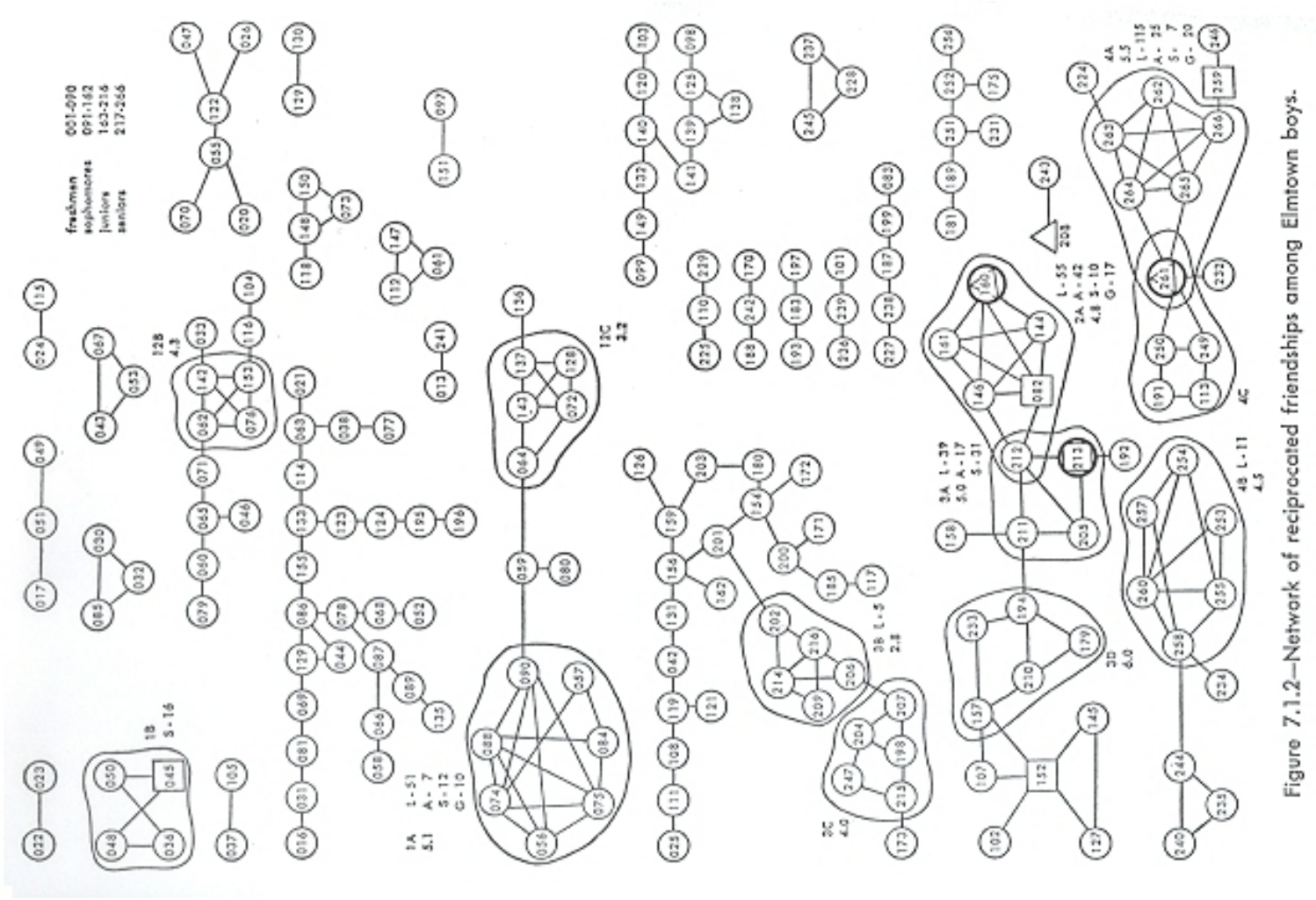


Figure 7.1.2—Network of reciprocated friendships among Elmtown boys.

Topologia: Grafo de amizade recíproca entre meninos de uma escola americana

Exemplos: Sociedade Adolescente - 2

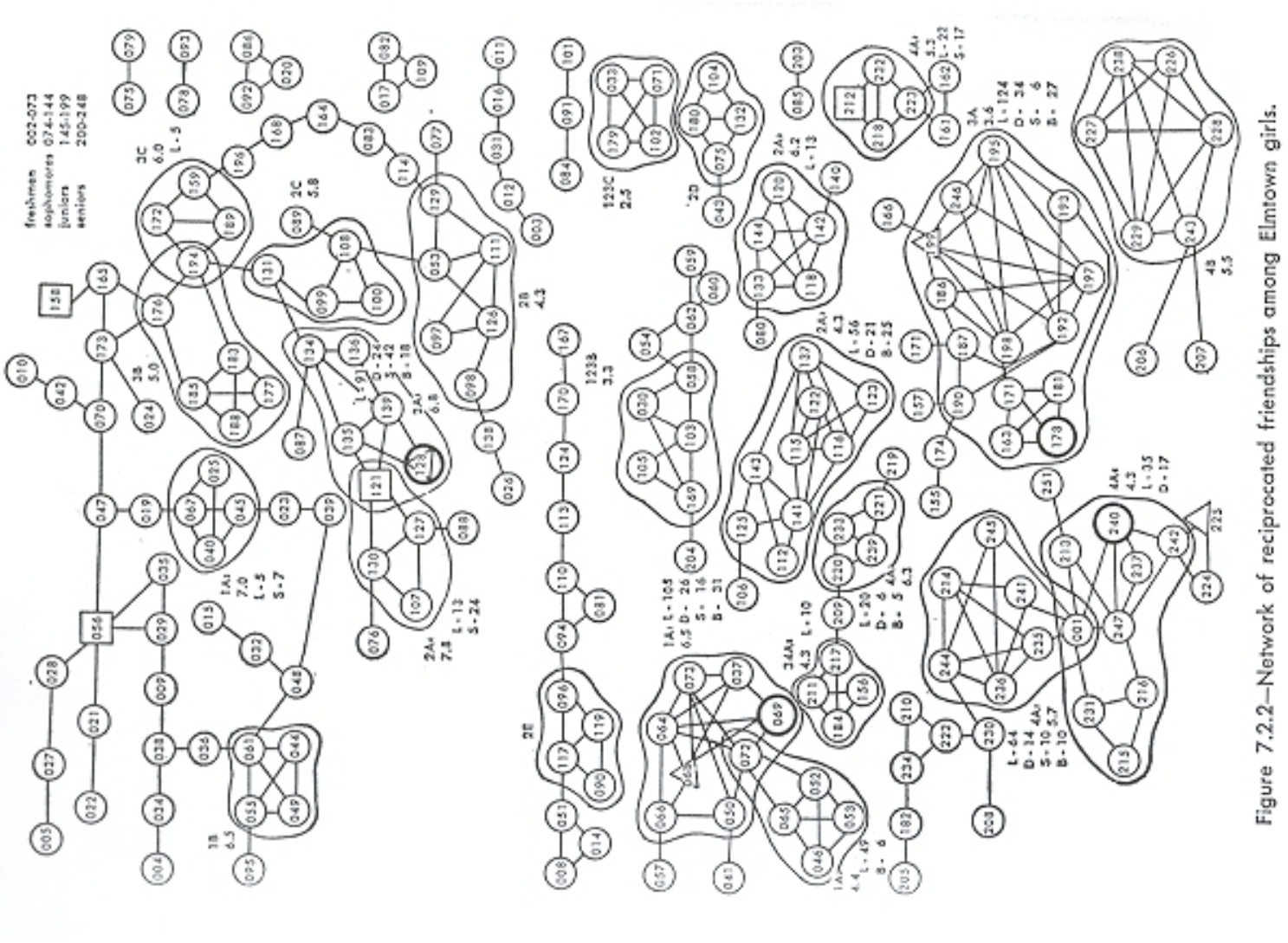


Figure 7.2.2—Network of reciprocated friendships among Elmtown girls.

- Grafo de amizade recíproca entre meninas de uma escola americana

Tipos de Dados Espaciais

Taxonomia dos Dados Espaciais

O fundamental é identificar o que é o componente aleatório em cada tipo de dado.

Este componente aleatório é que será modelado com distribuições de probabilidade.

Os 4 Tipos de Dados Espaciais:

- Dados de Superfície aleatória
- Dados de Processos Pontuais
- Dados de Área
- Dados de Interação Espacial

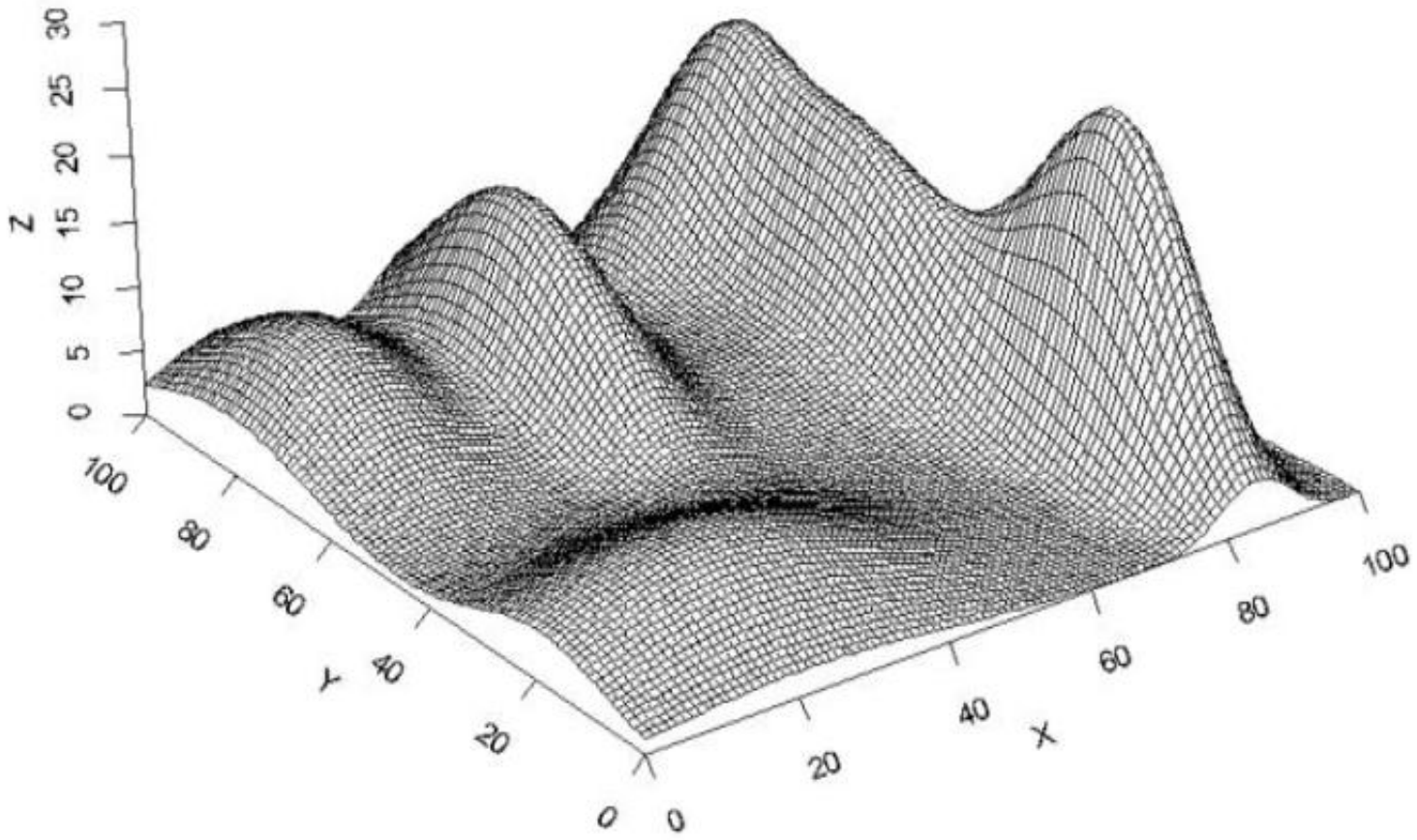
Dados de Superfície aleatória

Dado típico de estudos ambientais, geológicos e de ciências naturais.

Superfície $\boldsymbol{Y(s)}$ DEFINIDA em todo ponto $\boldsymbol{s} = (s_1, s_2)$ de uma região do plano.

Exemplos: Temperatura, Ph de água de lago, acidez do solo...

Uma superfície aleatória: temperatura



Amostrando uma Superfície Aleatória

Superfície $Y(\boldsymbol{s})$ é DEFINIDA em todo ponto $\boldsymbol{s} = (s_1, s_2)$ da região.

Mas... OBSERVADA apenas em alguns locais: n estações de coleta ou monitoramento.

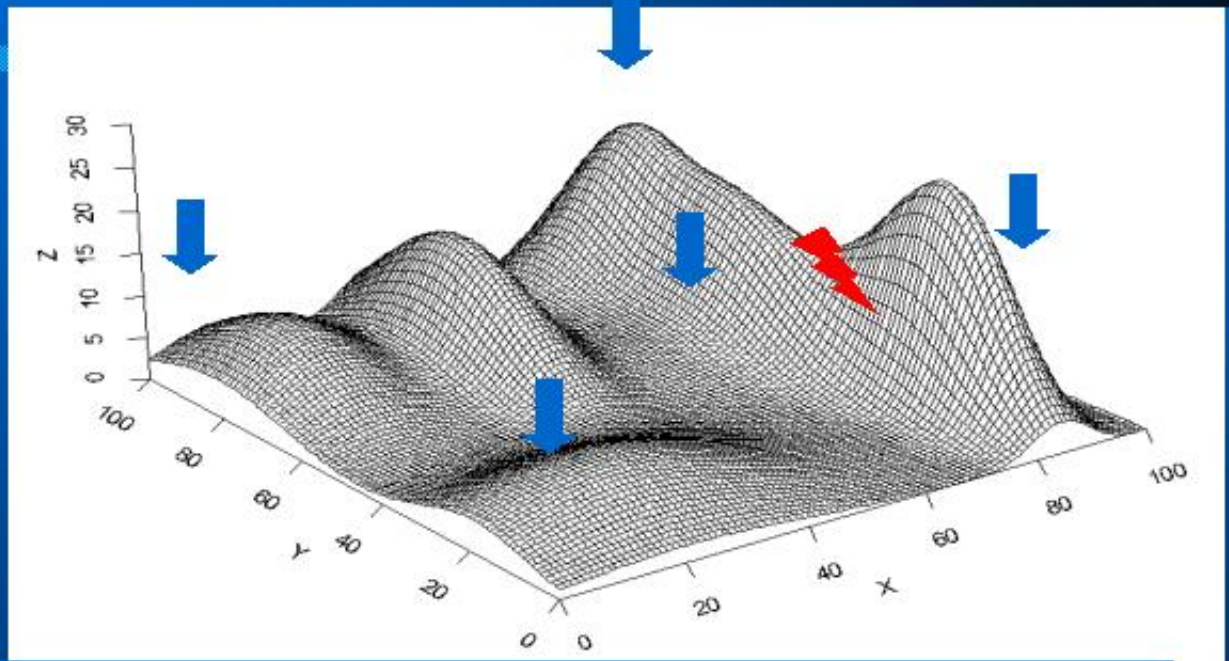
Estações $i = 1, \ldots, n$ em posições FIXAS e CONHECIDAS (não-aleatórias).

Estação i está localizada em $\boldsymbol{s}_i = (s_{1i}, s_{2i})$ e $Y(\boldsymbol{s}_i)$ é o valor da superfície na estação.

Interesse em $Y(\boldsymbol{s})$ onde $\boldsymbol{s}$ é localização não monitorada.

Aleatório é o valor da superfície.

Dados de Superfícies Aleatórias



Temperatura, Ph de água de lago, acidez do solo...

Estações de coleta nas locações marcadas por

Localização a predizer: marcada por

Problemas típicos

predizer superfície em posições novas

interpolação

escolher posição para instalar uma nova estação.

Krigagem é o método chave: regressão com erros correlacionados por distância.

Correlação de erros é definida pelo variograma (ou correlograma): função $\rho(d)$ que mede o grau de independência (correlação) entre os erros de acordo com distância d entre posições.

A função de correlação $\rho(d)$ deve satisfazer restrições severas para que, dado qualquer conjunto de n posições no plano, a matriz de correlação $n \times n$ resultante seja definida positiva

Mostra-se que $\rho(d)$ deve ser representada como uma integral de uma função de Bessel generalizada

Mapa topográfico de vulcão na Nova Zelândia. Pontos são os locais onde existe uma medição aproximada.

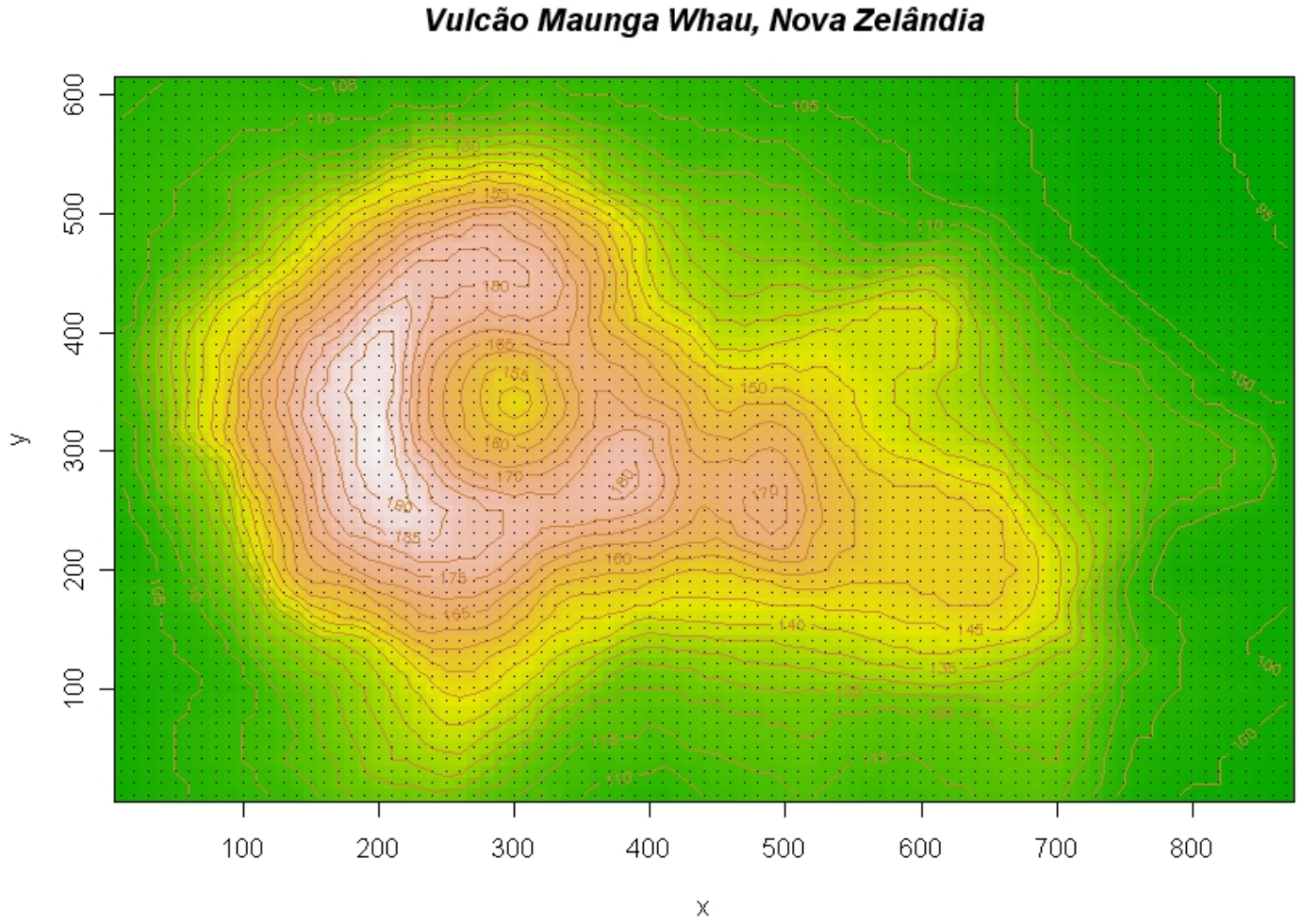
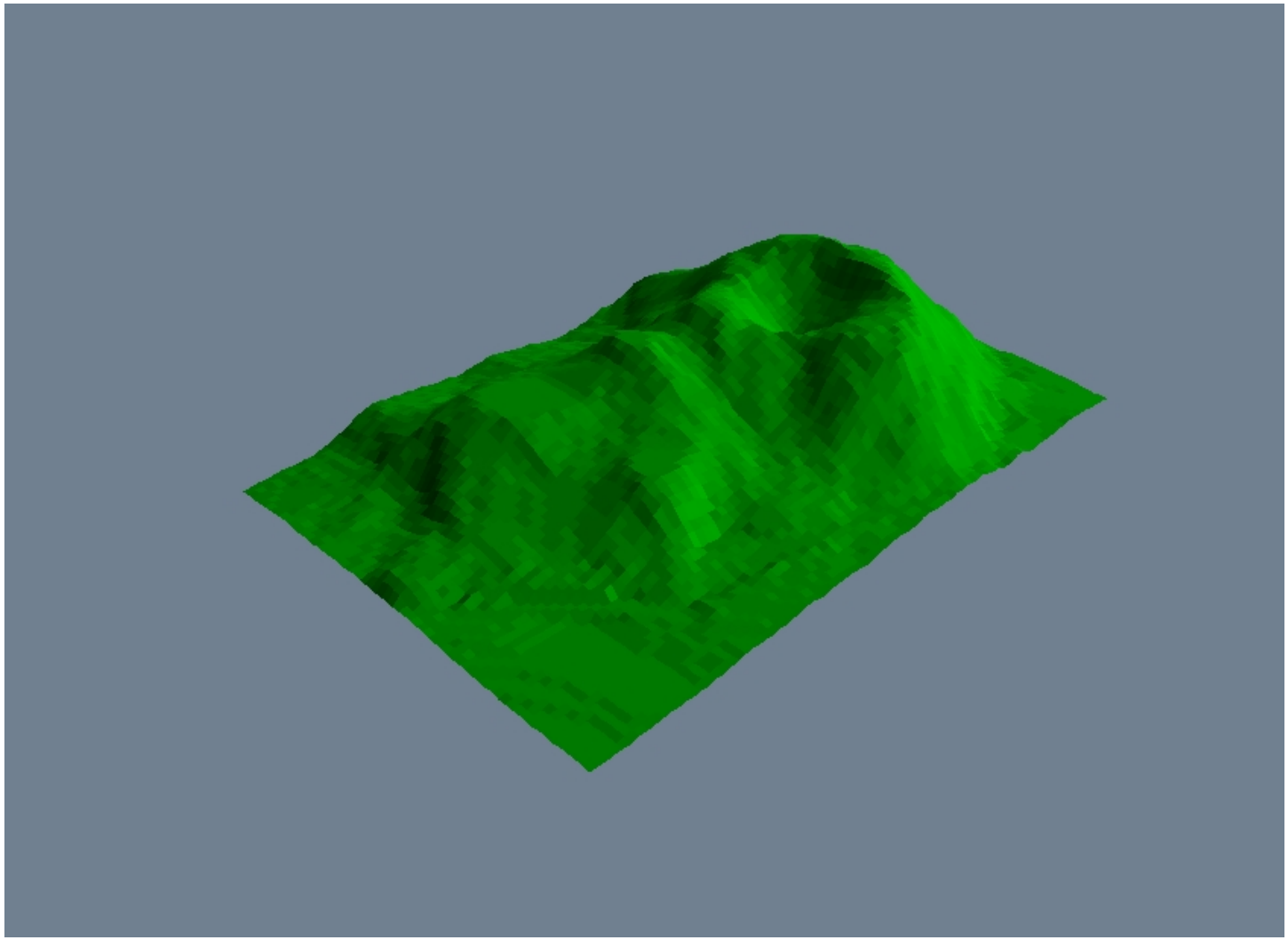
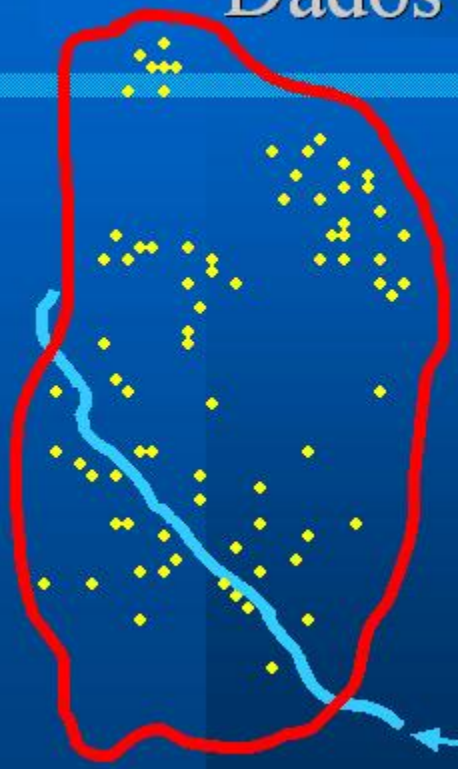


Imagem a partir do mapa topográfico de vulcão na Nova Zelândia.



Dados de Processos Pontuais



Localização de crimes numa certa região

O que é aleatório aqui ?

A posição espacial dos crimes

Avenida

Padrões de Pontos Aleatórios

Num padrão espacial de pontos, o que é aleatório ? A própria posição dos pontos ou eventos.

Modelo estocástico deve explicar a configuração espacial dos eventos: há interação entre os eventos tal como atração ou inibição ?

Problemas Típicos

- Existe aglomeração de casos ou eles estão dispostos ao acaso (proporcional à população sob risco) ?
- Estar próximo ao rio aumenta o risco de tornar-se um caso?
- Interação espaço-temporal: Eventos estão em 3 dimensões incluindo o tempo. Casos próximos no espaço tendem a estar próximos no tempo também ?
- Eventos de dois tipos diferentes. Por exemplo: casos e controles ou homicídios e roubo. Análise compara os padrões espaciais de cada tipo e testa se eles são similares.

Processo Pontuais na prática - TEMPO

Caso uni-dimensional: "espaço"é a reta representando o tempo.

Processos pontuais no tempo: eventos aleatórios no tempo. Muita teoria E muitas aplicações em diversas áreas.

Teoria moderna usa a estrutura ordenada do tempo e as ferramentas são martingalas, filtragens, etc.

Valores Extremos ou Sinistros no tempo: abordagem de processos pontuais em Finanças, seguros, hidrologia, estudos ambientais, etc.

Ocorrências raras em muitas pessoas: estudos de dados longitudinais.

Demografia: tempos entre sucessivos nascimentos de mulheres de uma população. Análise com base numa amostra de mulheres.

Economia: períodos de alternância entre emprego e desemprego.

Epidemiologia: tempos entre reinternações sucessivas.

Processo Pontuais na prática - espaço

No espaço: Produção teórica menor que no tempo: implica em menos aplicações também.

análise ambiental: previsão e controle de queimadas em florestas.

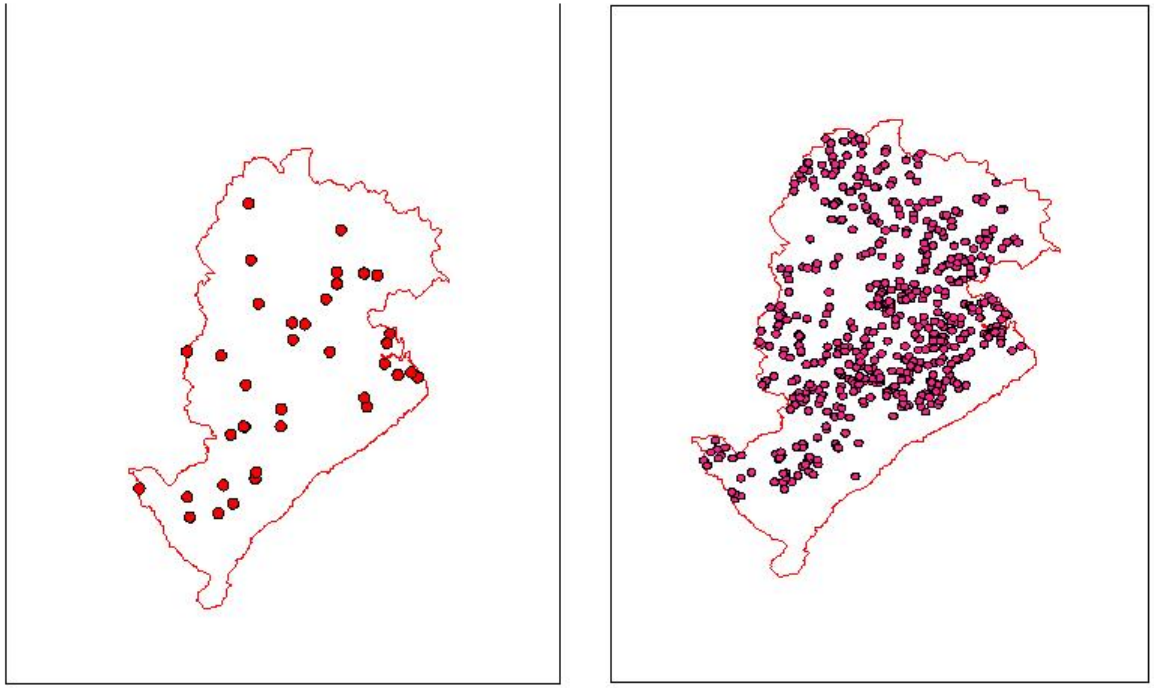
Previsão de terremotos, enchentes ao longo de rios ...

Análise áreas de maior incidência de crimes.

Epidemiologia: padrão espacial de uma doença. Existem áreas de maior incidência? Comparação de dois padrões.

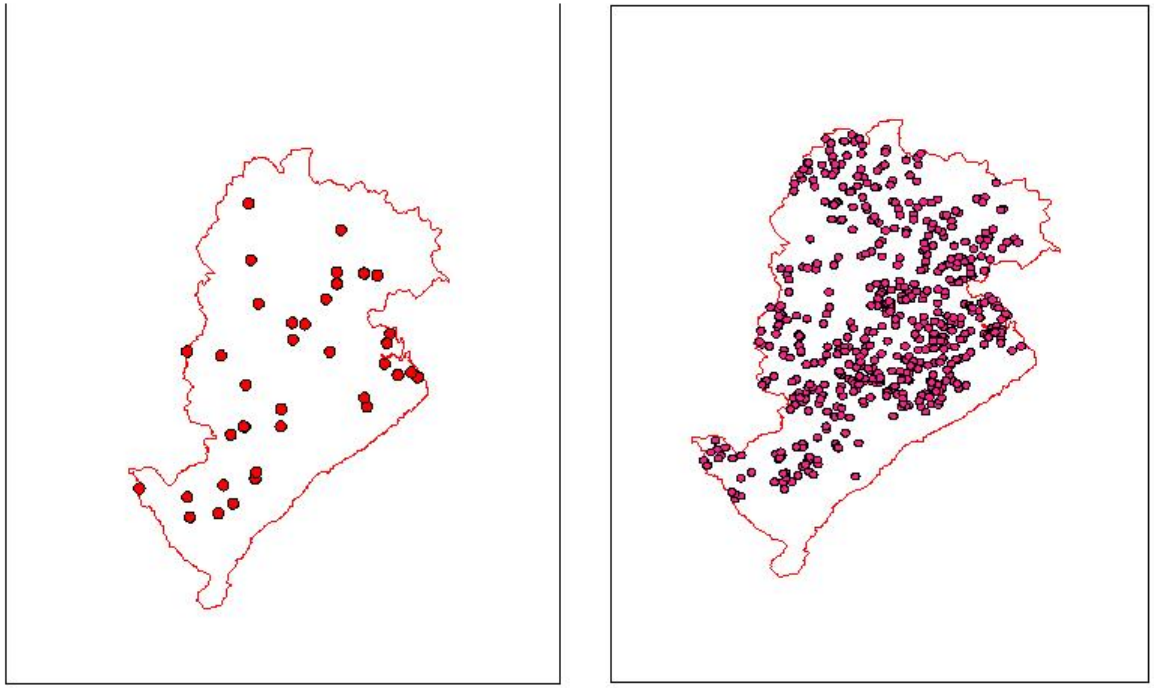
HTLV (esq) e controles (dir) em BH

Os casos de HTLV possuem a mesma dispersão espacial que os controles (pessoas sem o vírus)?



HTLV (esq) e controles (dir) em BH

Os casos de HTLV possuem a mesma dispersão espacial que os controles (pessoas sem o vírus)?



Dados de Interação Espacial

Exemplos:

- Migração de mão de obra
- Fluxo de bens entre centros urbanos
- Tráfego de comunicação Web numa rede
- Rede social trafegando informação, status, etc.

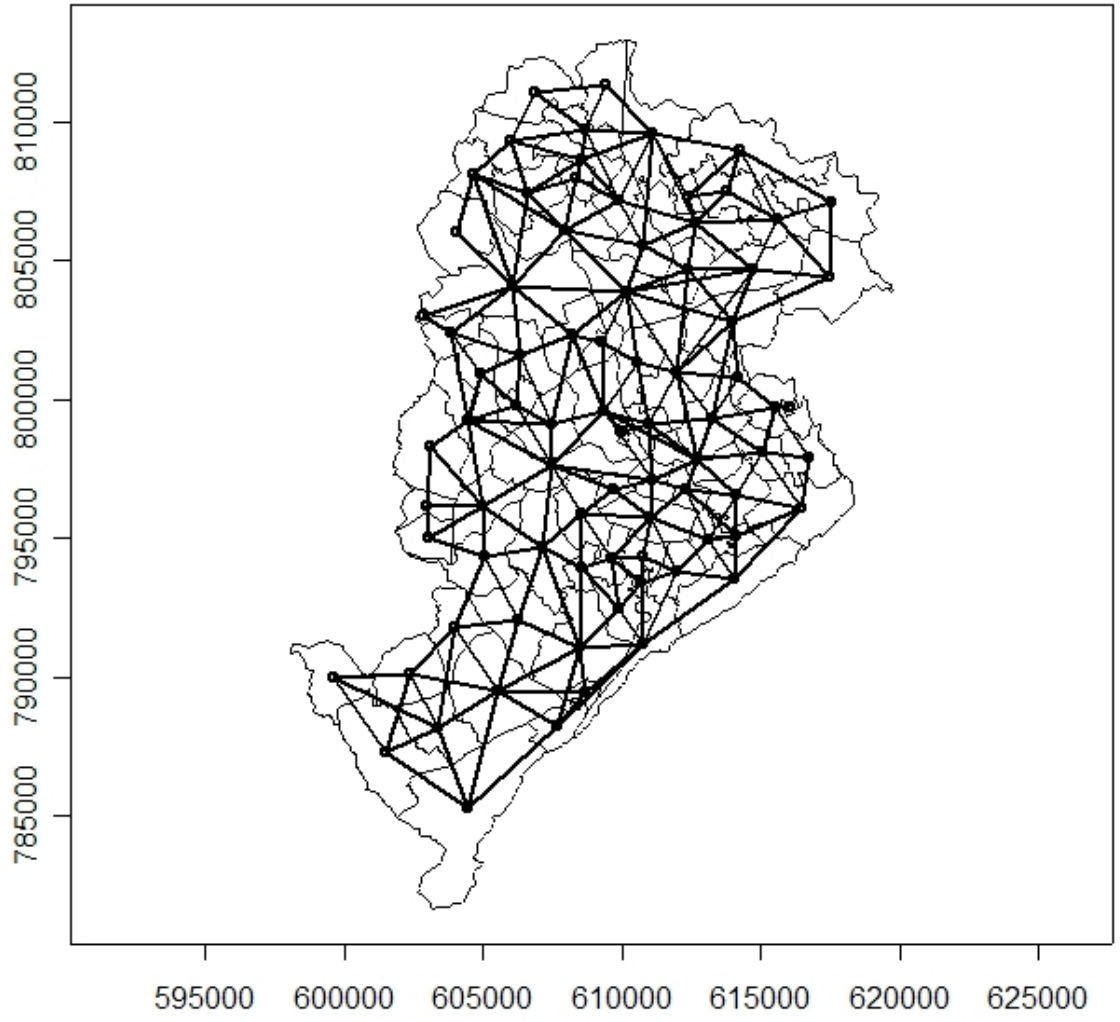
Muito comum em Economia Urbana, Economia Regional, etc.

De Origem i flui $\longrightarrow$ para Destino j

Fluxo aleatório Y_{ij} entre posições i e j

Cada medição Y_{ij} refere-se a um PAR ORDENADO de posições (i, j) e NÃO apenas a um local específico i

Topologia é um grafo: vizinhança dos bairros de Belo Horizonte



Interação Espacial

Problemas típicos:

- Que características de i e j determinam o volume do fluxo?
- Como os fluxos podem afetar características do local i ?
- Onde colocar novo centro para minimizar custo ?
- Como um tipo de fluxo afeta outros tipos de fluxos ?

Modelo típico é o gravitacional

- $E(Y_{ij}) \propto g(\boldsymbol{x}_i)\, g(\boldsymbol{x}_j)/d_{ij}^{\alpha}$
- onde $\boldsymbol{x}_i$ são as características de i
- $g(\boldsymbol{x}_i)$ é uma função das características da área i
- d_{ij} é a distância entre i e j.

Outra abordagem possível mas pouco usada: Campos aleatórios de Markov

Dados de Área

Região $\mathcal{R}$ é particionada em n áreas

Em cada área é feita uma observação aleatória Y_i

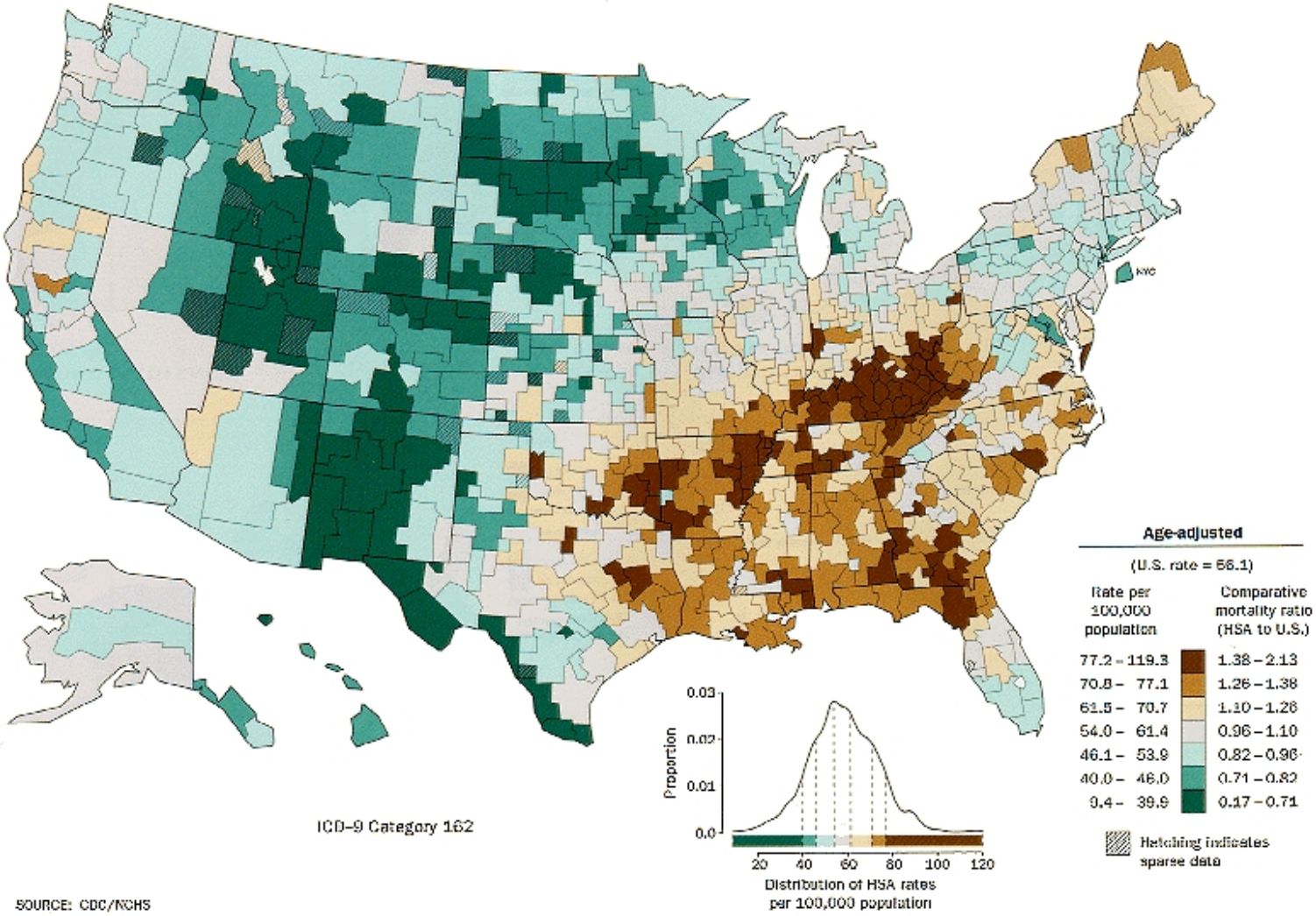
Exemplos:

- PIB per capita no município i
- número de desempregados no município i
- preço médio de imóvel de certo perfil no bairro i
- número de crimes no bairro i

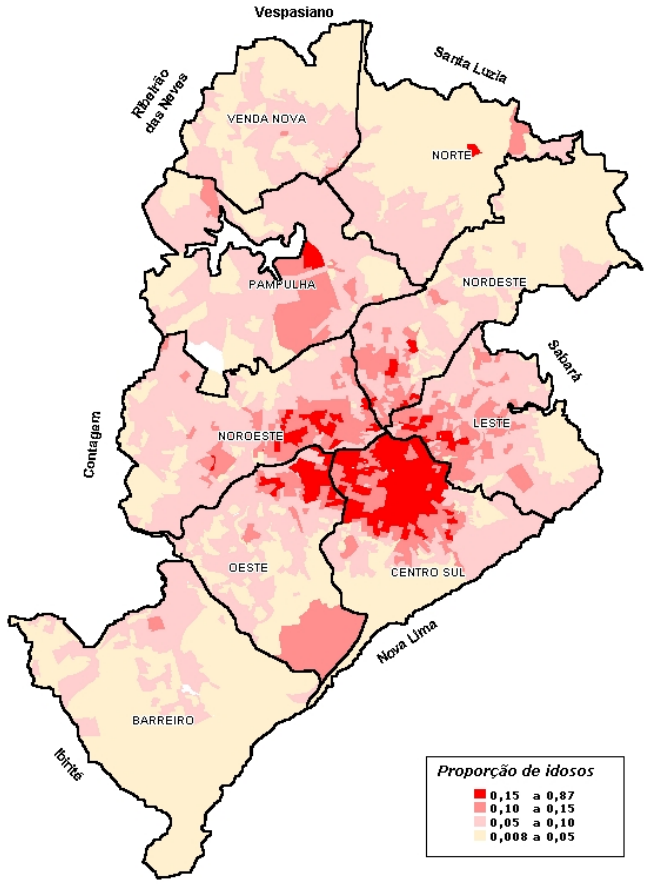
Este é o tipo de dado mais comum em econometria espacial

Restante desse curso só trata desse tipo de dado

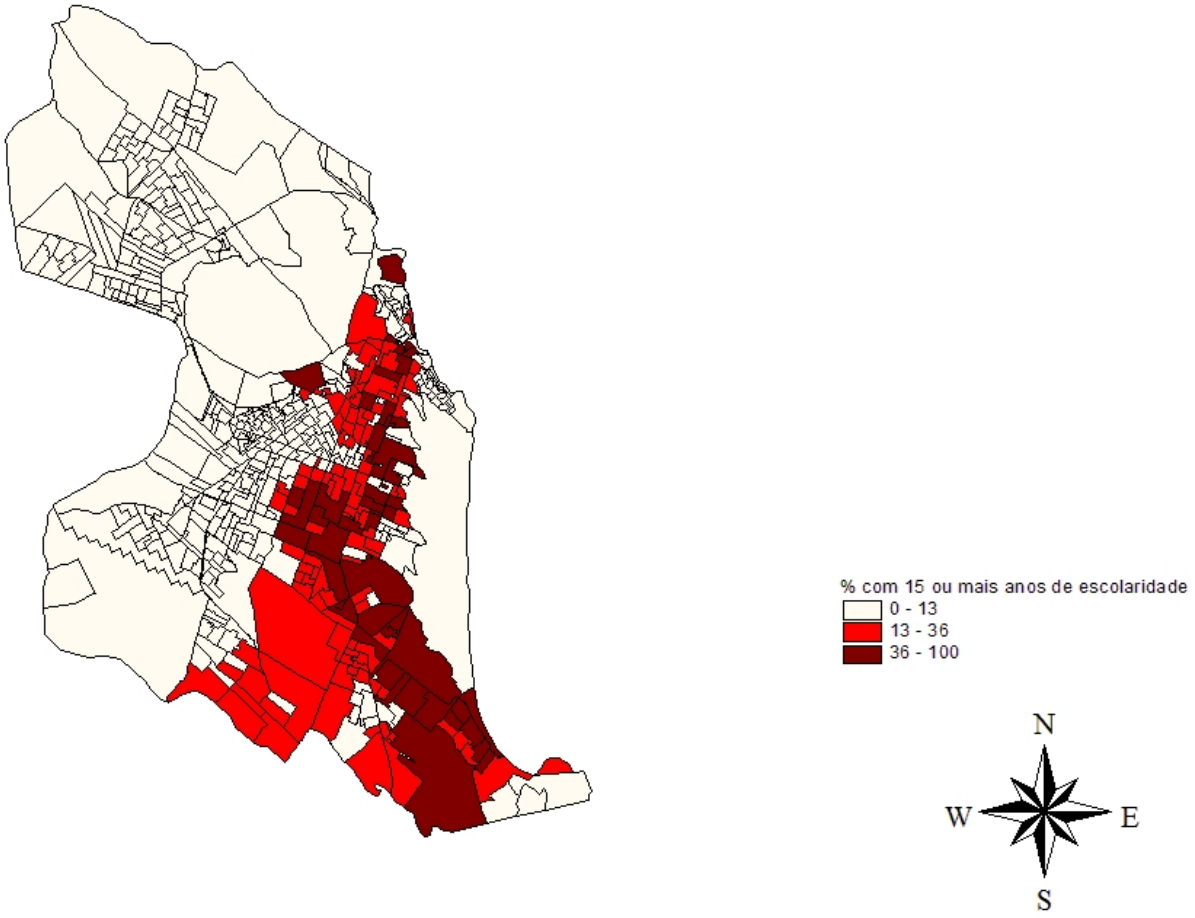
Câncer de pulmao:EUA



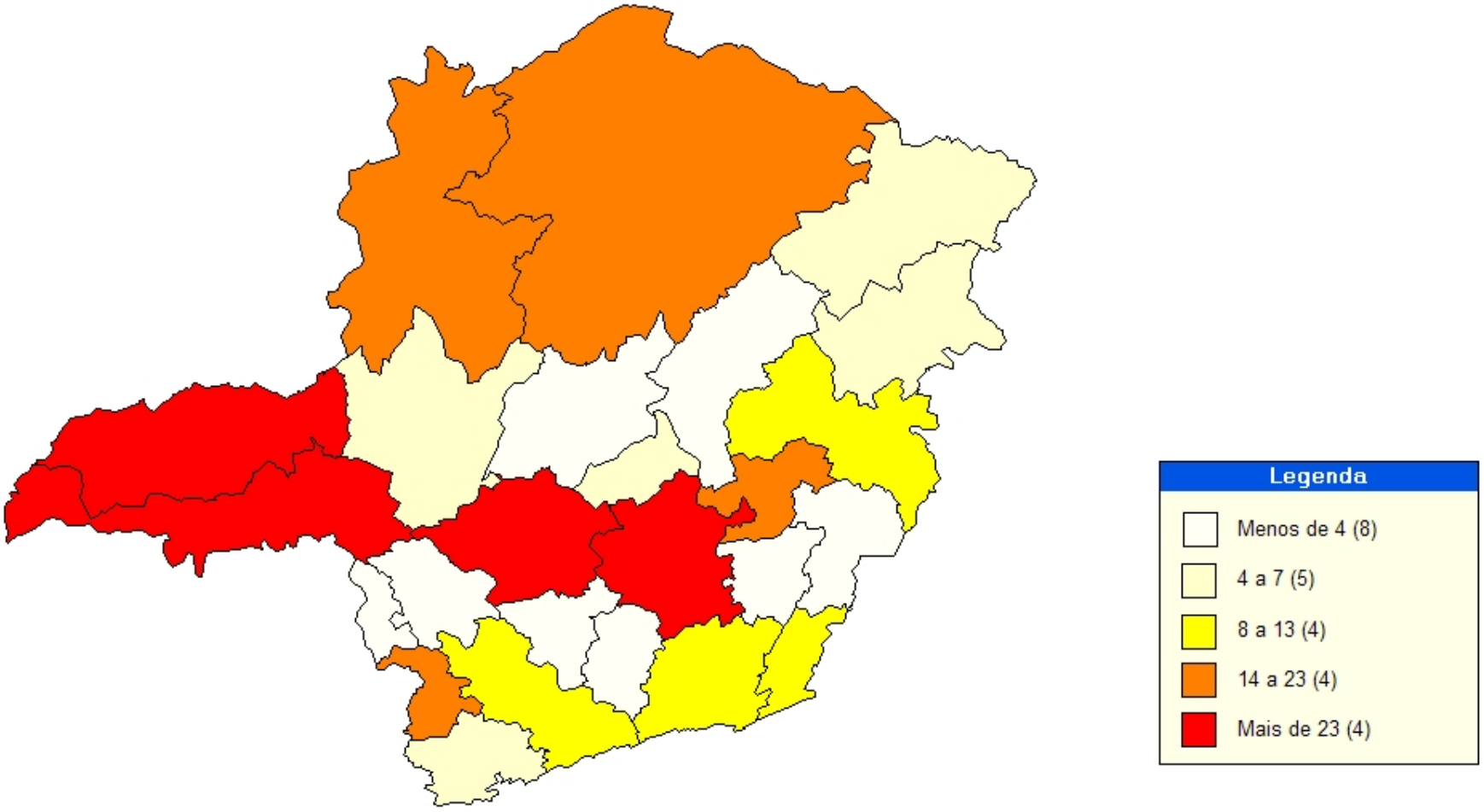
Idosos por setor censitario em BH



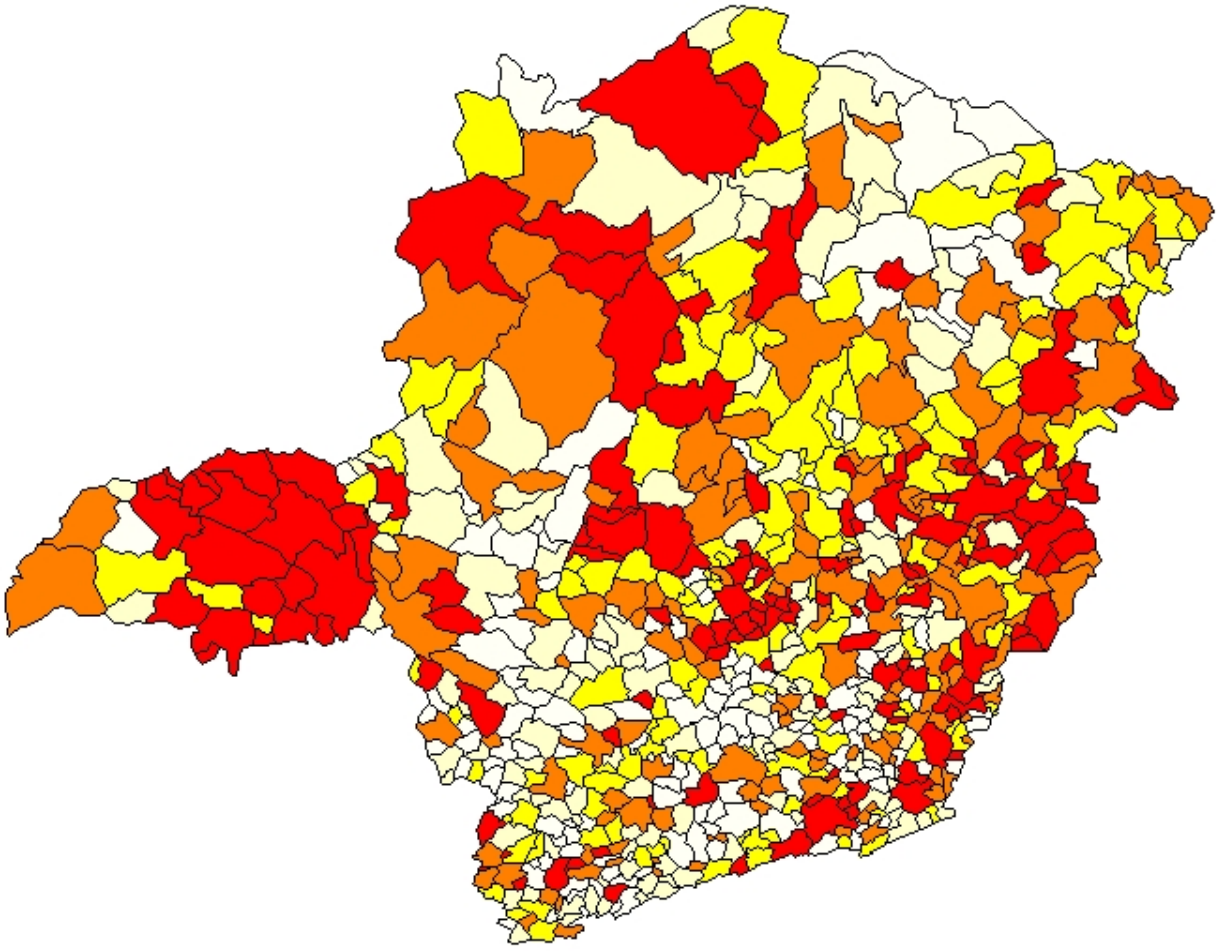
Natal: % com 15 ou mais anos de escolaridade



Minas Gerais
Roubo de Veículo, 1997
(Número de Ocorrências)



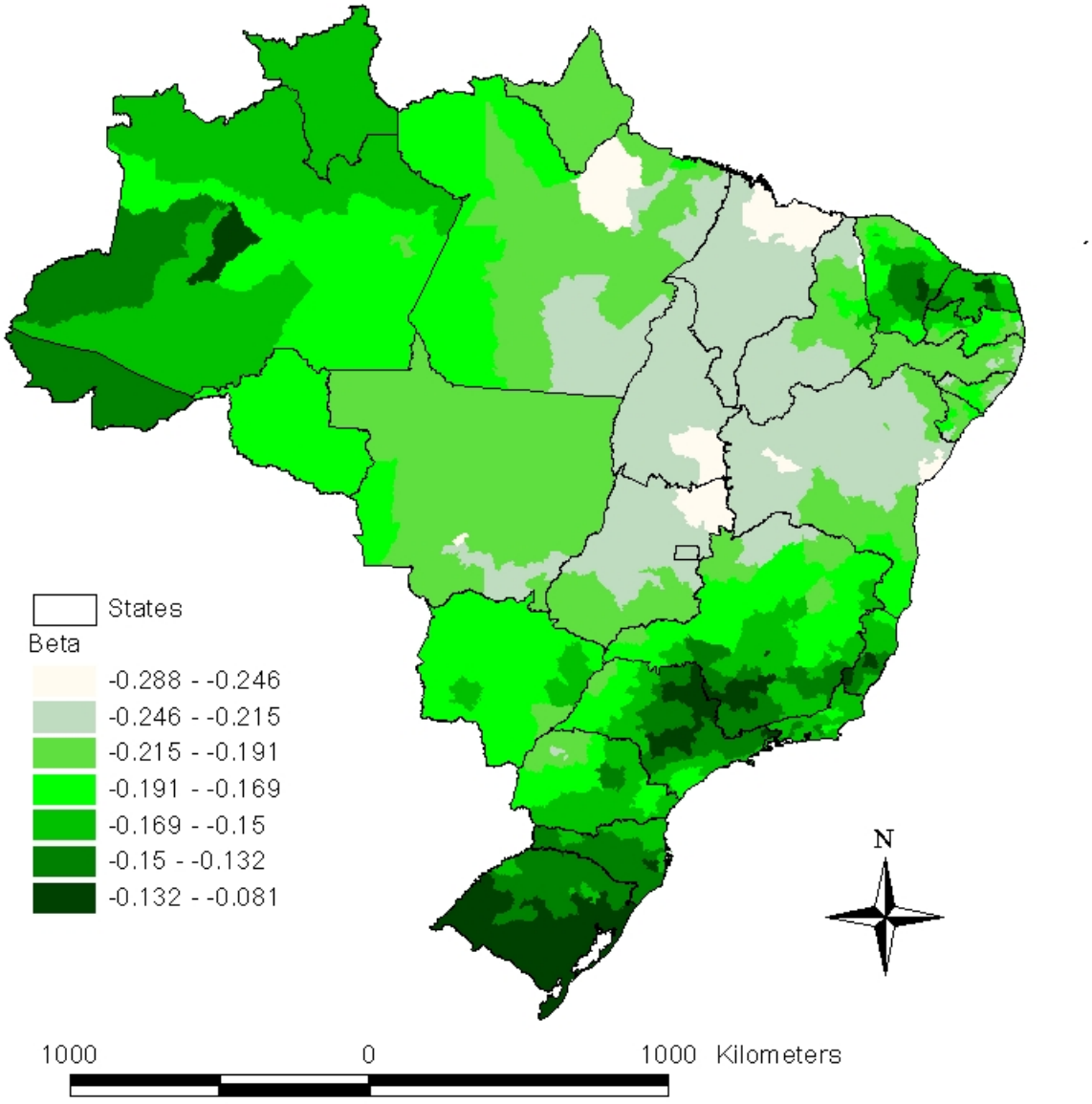
Minas Gerais
Crimes Violentos, 1997
(Taxa de Risco por 100.000 Habitantes)



Legenda

- Menos de 25.70 (144)
- 25.70 a 44.99 (144)
- 45.00 a 66.24 (145)
- 66.25 a 95.16 (145)
- Mais de 95.16 (145)

Velocidade da Queda de Fecundidade: mais negativo, mais rapido



Dados de Área - 2

Região $\mathcal{R} = \bigcup_{i=1}^{n} \mathcal{A}_i$ com $\mathcal{A}_i \cap \mathcal{A}_j = \emptyset$ se $i \neq j$

Em cada área é feita uma observação aleatória Y_i

Via de regra, Y_i é uma agregação, uma soma ou é uma integral sobre a área $\mathcal{A}_i$

Medições referem-se a toda a área $\mathcal{A}_i$, não a um ponto particular dentro dela

Não faz sentido "interpolar" entre áreas

Problema típico: regressão de Y_i em variáveis explicativas $\boldsymbol{x}_i$ mas variáveis $\boldsymbol{Y}_i$ são correlacionadas

Estrutura de correlação do vetor $\boldsymbol{Y}$ é determinada pela topologia: posição no plano, indicadores binários de vizinhança espacial, ou distância entre todos os possíveis pares de áreas.

Exemplos com dados de área: preços hedônicos

Preços Hedônicos em mercado imobiliário e mercado de trabalho: ambos possuem um forte componente espacial

Existem três coisas que influenciam o preço de um imóvel: localização.

Salários e aluguéis ou valores imobiliários variam bastante dentro de uma cidade. Um mesmo tipo de imóvel (idade, tamanho, qualidade de acabamento, etc.) terá preços muito diferentes no Meyer, em Botafogo, no Leblon e na Barra. Tudo o mais igual, áreas vizinhas tendem a ter preços parecidos.

Controlando por fatores conhecidos, preços terão erros espacialmente correlacionados.

O uso de espaço ajuda a controlar variáveis não mensuradas que possuem uma estrutura espacial: crime, poluição do ar, acesso/transporte, e outras externalidades.

Ver, por exemplo, Basu e Thibodeau (1998) Analysis of Spatial Autocorrelation in House Prices. *Journal of Real Estate Finance and Economics*, 17, 61-85.

Econometria com dados de área: loteria

Coughlin, Garrett e Hernandez-Murillo (2003) Spatial probit and the geographic patterns of state lotteries *Working Papers from Federal Reserve Bank*

Alguns estados americanos correm uma loteria e outros não

O que diferencia uns dos outros ?

Modelo Probit espacial

algumas variáveis explicativas:

evidênvia de superdispersão (efeito misto): variabilidade extra binomial

Esta variação extra-binomial tem estrutura espacial: áreas próximas tendem a ter probabilidades desviando-se do preditor linear de forma similar

Regionalização

Em economia regional, um problema constante é agrupar áreas que sejam similares ou homogêneas com respeito a um certo número de variáveis

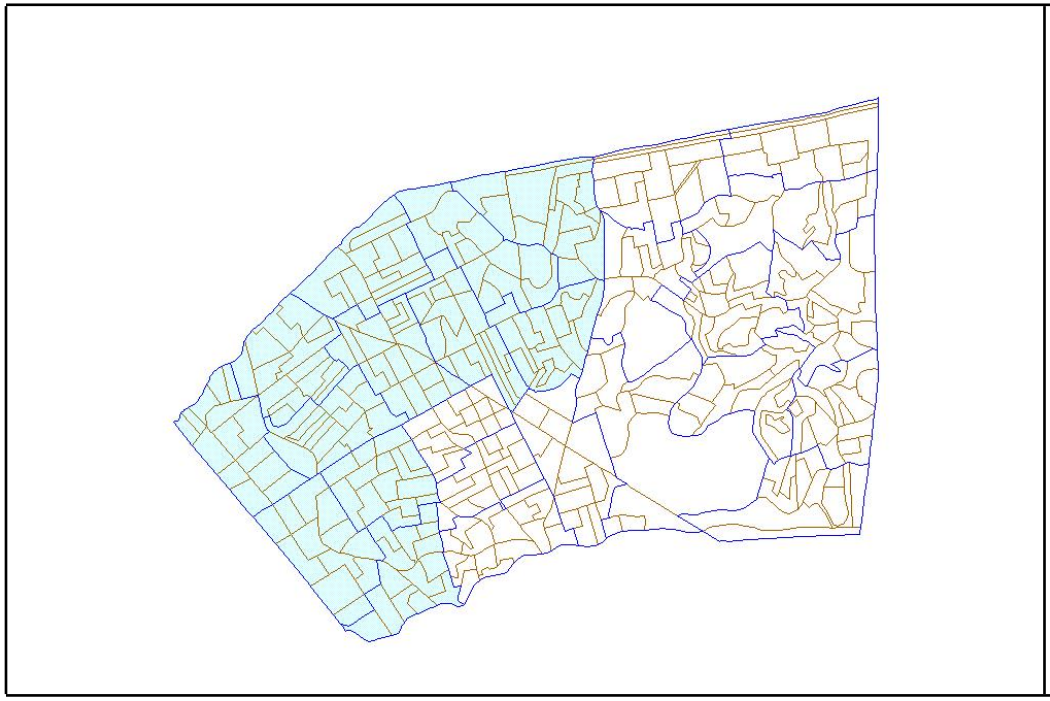
Problema dual: idenificar fronteiras entre regiões homogêneas

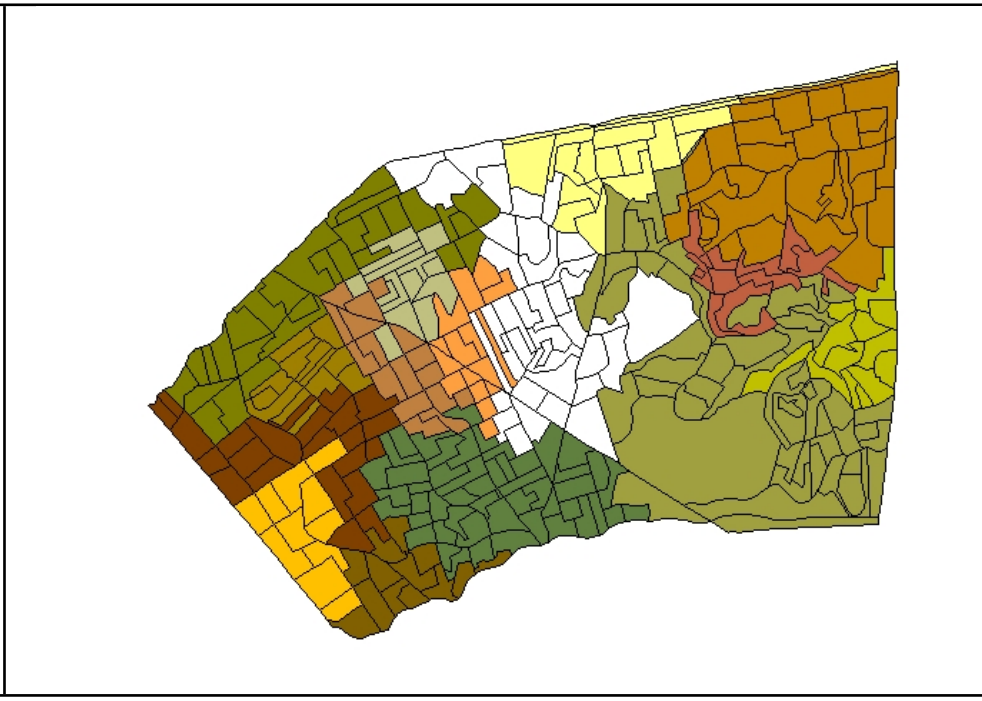
Assunção *et al.* (2001) propuseram método baseado em teoria de grafos.

Implementado no software SKATER: Spatial K-lustering Analysis Through Edge Removal

Disponível em www.est.ufmg.br/leste

Exemplo de Regionalização





São João do Meriti - RJ. Setores Censitários agregados com base em 15 variáveis sociais e econômicas do Censo Demográfico, 1991