NEREUS

Núcleo de Economia Regional e Urbana
da Universidade de São Paulo

The University of São Paulo
Regional and Urban Economics Lab

Aulas 2-4: Análise Exploratória de Dados Espaciais (AEDE)

Prof. Eduardo A. Haddad

Principal mensagem

Dependência espacial

Primeira Lei da Geografia (Waldo Tobler):

"Everything is related to everything else, but near things are more related than distant things"

Análise espacial

Invariância locacional

- Análise espacial **não** é invariante locacionalmente
- Os resultados mudam quando as localizações do objeto de estudo mudam
- "Onde" importa!

Mapeamento e Geovisualização

- Apresentando padrões interessantes

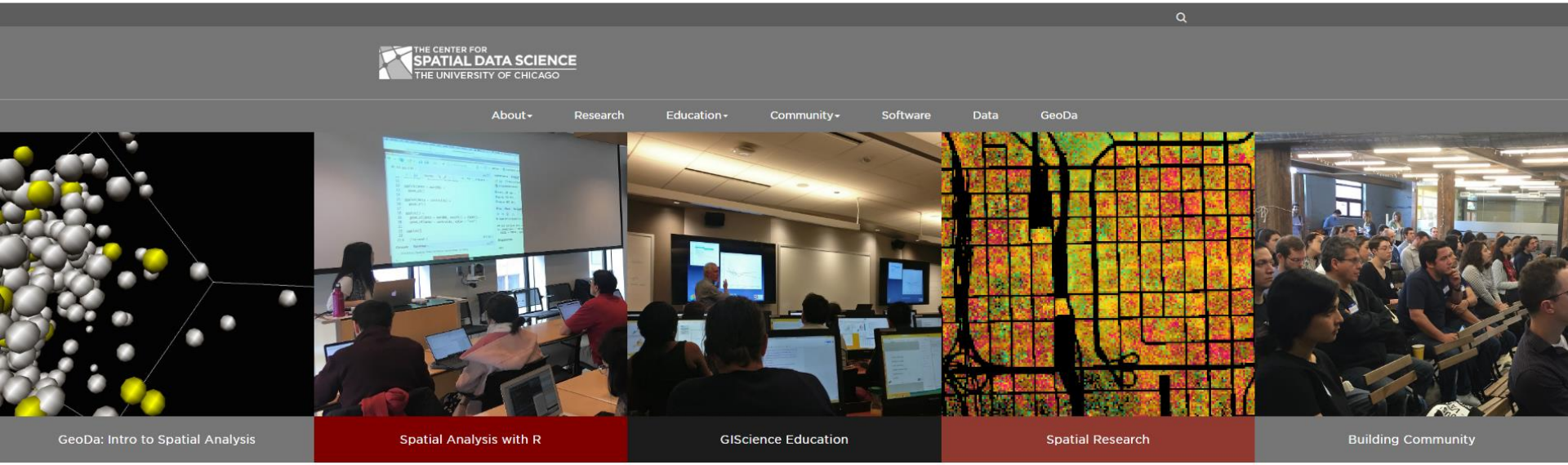
Análise Exploratória de Dados Espaciais

- Descobrindo padrões interessantes

Modelagem Espacial

- Explicando padrões interessantes

GeoDa – https://spatial.uchicago.edu/



About the Center

At the Center for Spatial Data Science (CSDS), we think spatially about research problems: We develop state-of-the-art methods for geospatial analysis; implement them through open source software tools; apply them to policy-relevant research in the social sciences; and disseminate them through education and support to a growing worldwide community of over 270,000 spatial analysts. Locally, we are building a spatial community at the University of Chicago. CSDS is an initiative of the Division of Social Sciences and part of the UChicago's investment in computational social science. As of July 2016, CSDS succeeded the GeoDa Center for Geospatial Analysis and Computation at Arizona State University.

AEDE

Análise exploratória de dados (AED) utiliza um conjunto de técnicas para:

- maximizar *insights* sobre um banco de dados
- descobrir estruturas subjacentes
- extrair variáveis importantes
- detectar *outliers* e anomalias
- testar hipóteses subjacentes
- sugerir hipóteses
- desenvolver modelos parcimoniosos

AED**E** inclui os atributos espaciais dos dados

Geovisualização

Além de mapeamento:

- Combina mapas e métodos científicos de visualização
- Explora a capacidade humana de reconhecimento de padrões

Mapas estatísticos

- Formas inovadoras de mapeamento (*quantile map, percentile map, box map, standard deviation map, conditional map*)

Map movie

Exercício 1 (mapping)

Open and close a project

Load a shape file with the proper indicator (Key)

Select functions from the menu or toolbar

Make a simple choropleth map

Select items in the map

Dados: Angola (PIB_PC_DLR); Brazil_UF (Y00, G00, RG00); Brazil_MR (RENDAPC, ESGT, M1SM)

Variáveis estaduais, *Atlas de Desenvolvimento Humano*

Código	Descrição
D1	10% mais ricos / 40% mais pobres, 1991
D2	10% mais ricos / 40% mais pobres, 2000
D3	20% mais ricos / 40% mais pobres, 1991
D4	20% mais ricos / 40% mais pobres, 2000
G91	Índice de Gini, 1991
G00	Índice de Gini, 2000
T91	Índice de Theil, 1991
T00	Índice de Theil, 2000
Y91	Renda per Capita , 1991
Y00	Renda per Capita , 2000
RG91	% da renda proveniente de transferências governamentais, 1991
RG00	% da renda proveniente de transferências governamentais, 2000
RT91	% da renda proveniente de rendimentos do trabalho, 1991
RT00	% da renda proveniente de rendimentos do trabalho, 2000
PRG91	% de pessoas com mais de 50% da renda provenientes de transferências governamentais, 1991
PRG00	% de pessoas com mais de 50% da renda provenientes de transferências governamentais, 2000

Exercício 2 (data)

Open and navigate the data table

Select and sort items in the table

Create new variables in the table

<u>Data</u>: Brazil_UF (Y00)

Quais estados apresentaram maior/menor crescimento da renda per capita no período 1991-2000?

Table > Calculator

Exercício 3 (EDA)

Uso de gráficos, como *box plot*, diagramas de dispersão, histogramas, etc.

Visualização da distribuição **não-espacial** de uma variável

Método *linking and brushing*: seleção das unidades regionais é simultânea em todas as representações de dados, i.e., tabelas, gráficos e mapas

Dados: Brazil_UF (Y00, G00)

Exercício 3 (EDA)

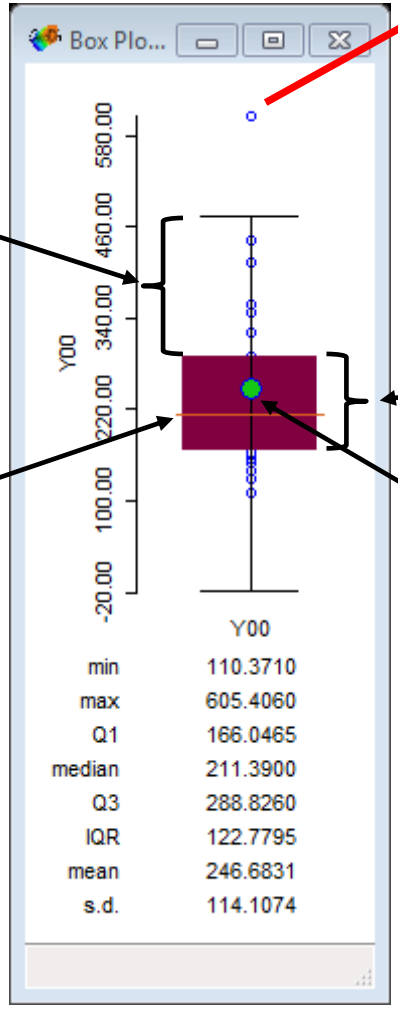
Histograma:

- Aproximação discreta da função densidade de uma variável aleatória

Box Plot:

- Mostra a mediana, primeiro e terceiro quartis de uma distribuição (pontos 50%, 25% e 75% na distribuição acumulada)
- Noção de ***outlier***: observação que se encontra a mais de um dado múltiplo (1.5 ou 3.0) do intervalo interquartílico, acima ou abaixo dos percentis 75% e 25%, respectivamente

Box plot



Outlier

Hinge = 1.5

Intervalo interquartílico (25% a 75%)

Mediana

Média

Exercício 3 (EDA)

<u>*Scatter plot*</u> (gráfico de dispersão):

- Relação entre duas variáveis

<u>*Correlation plot*</u> (gráfico de correlação):

- Variáveis padronizadas para unidades de desvio-padrão
- Valores > 2: *outliers*!

REVISÃO

<u>*Dados*</u>*: Brazil_MR (RENDAPC)*

Exercício 3 (EDA)

Conditional plots:

Gráficos condicionais, também conhecidos como gráficos em grade ou gráficos de Trellis (Becker, Cleveland e Shyu 1996), fornecem um meio de avaliar as interações entre mais de duas variáveis. Permite avaliar se a relação entre duas variáveis depende de valores de outra(s) variável(is).

Conditional scatter plot

Dados: Brazil_MR (RENDA=f(POP), ESGT, M1SM); Angola (PIB=f(POP_2014), TX_DESEMP)

Exercício 4 (spatial scale and rate of density)

Inferência pode mudar com a escala espacial

Problemas de agregação espacial

- Estado *versus* microrregião

Mapas de intensidade

- Variável extensiva tende a ser correlacionada com tamanho (e.g. área ou população total)
- Taxa de densidade (variável intensiva) é uma métrica mais apropriada para um mapa coroplético.

<u>Dados</u>: Brazil_UF (Y00); Brazil_MR (RENDAPC, RENDA, POP); Angola (PIB, POP_2014)

Rates-calculated map (Raw rate; Excess risk)

Dependência espacial

O que ocorre em um lugar depende de eventos ocorridos em lugares próximos

"All things are related but nearby things are more related than distant things" (Tobler)

Categorização

- Tipo: substantivo *versus* ruído
- Direção: positiva *versus* negativa

Questões relevantes

- Tempo *versus* espaço
- Inferência

Conceitos

Pesos espaciais

Defasagem espacial

Autocorrelação espacial

Matriz de pesos espaciais

Definição:

- Matriz W, positiva, N x N, elementos w_{ij}
- w_{ij} diferente de zero para vizinhos, 0 para não-vizinhos
- $w_{ii} = 0$, não há auto-vizinhança

Pesos geográficos (contiguidade, distância, pesos baseados em relações gráficas)

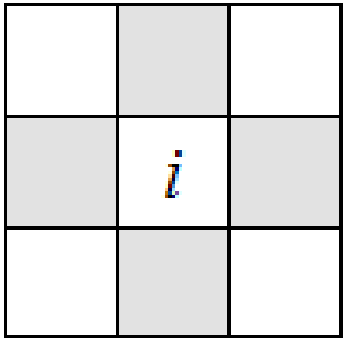
Pesos socioeconômicos

Pesos: contiguidade

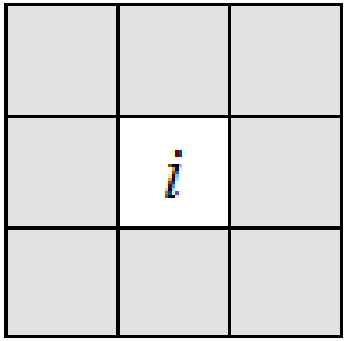
Contiguidade: unidades espaciais compartilham uma fronteira comum

Três visões de contiguidade:

- *rook*
- *bishop*
- *queen*



rook



queen

Exercício 5 (spatial weights)

Pesos espaciais baseados em contiguidade

Criar um arquivo de pesos espaciais baseados em contiguidade de primeira ordem a partir de um *shape file* de polígonos, usando os critérios *rook* e *queen* (.GAL)

Estrutura de conectividade dos pesos em um histograma

Contiguidade de ordem superior

Dados: Brazil_UF; Angola

Exercício 6 (spatial weights)

Pesos espaciais baseados na distância

Criar um arquivo de pesos espaciais baseados na distância a partir de um *shape file* de pontos, especificando uma distância crítica de referência

Ajustar a distância crítica

Criar um arquivo de pesos espaciais baseados no critério dos **k vizinhos mais próximos**

Dados: Brazil_UF; Angola (k=5)

Defasagem espacial

Variáveis defasadas espacialmente são parte essencial no cálculo de testes de autocorrelação espacial e na especificação de modelos espaciais de regressão

Wy: é a média de uma variável nos lugares vizinhos (e.g. WY00 é a média da renda per captia dos vizinhos)

Exercício 7 (spatial lag)

Gráfico de dispersão de Y00 e WY00

Usar valores padronizados

Dados: Brazil_UF (Y00)

Clustering

Característica global

Propriedade de um padrão global = todas as observações com valores similares são agrupadas no espaço

Teste por meio de uma estatística de autocorrelação global

Não há determinação da localização de clusters

Clusters

Característica local

Onde valores similares são mais agrupados no espaço quando comparado a um processo aleatório?

Propriedade de um padrão local = específico de cada localização

Teste por meio de uma estatística de autocorrelação local

Clusters locais podem ser compatíveis com aleatoriedade espacial global

Estatística de autocorrelação global

Teste formal de *match* entre similaridade de valor e similaridade de localização

Estatística resume os dois aspectos

Significância: qual a probabilidade (*p-value*) da estatística calculada apresentar este valor (extremo) em um padrão de aleatoriedade espacial?

Similaridade de atributo

Resumo da similaridade ou dissimilaridade de uma variável em localidades distintas:

- Variável y na localidade i, j com $i \neq j$

Medidas de similaridade:

- Produto cruzado: $y_i y_j$

Medidas de dissimilaridade:

- Quadrado das diferenças: $(y_i - y_j)^2$
- Valor absoluto das diferenças: $|y_i - y_j|$

Autocorrelação espacial global (I de Moran)

$$I = \left(\frac{n}{S_0}\right) \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} z_i z_j}{\sum_{i=1}^{n} z_i^2}$$

$z_i = x_i - x$

Estatística de produto cruzado

Similar ao coeficiente de correlação

Valor depende de W

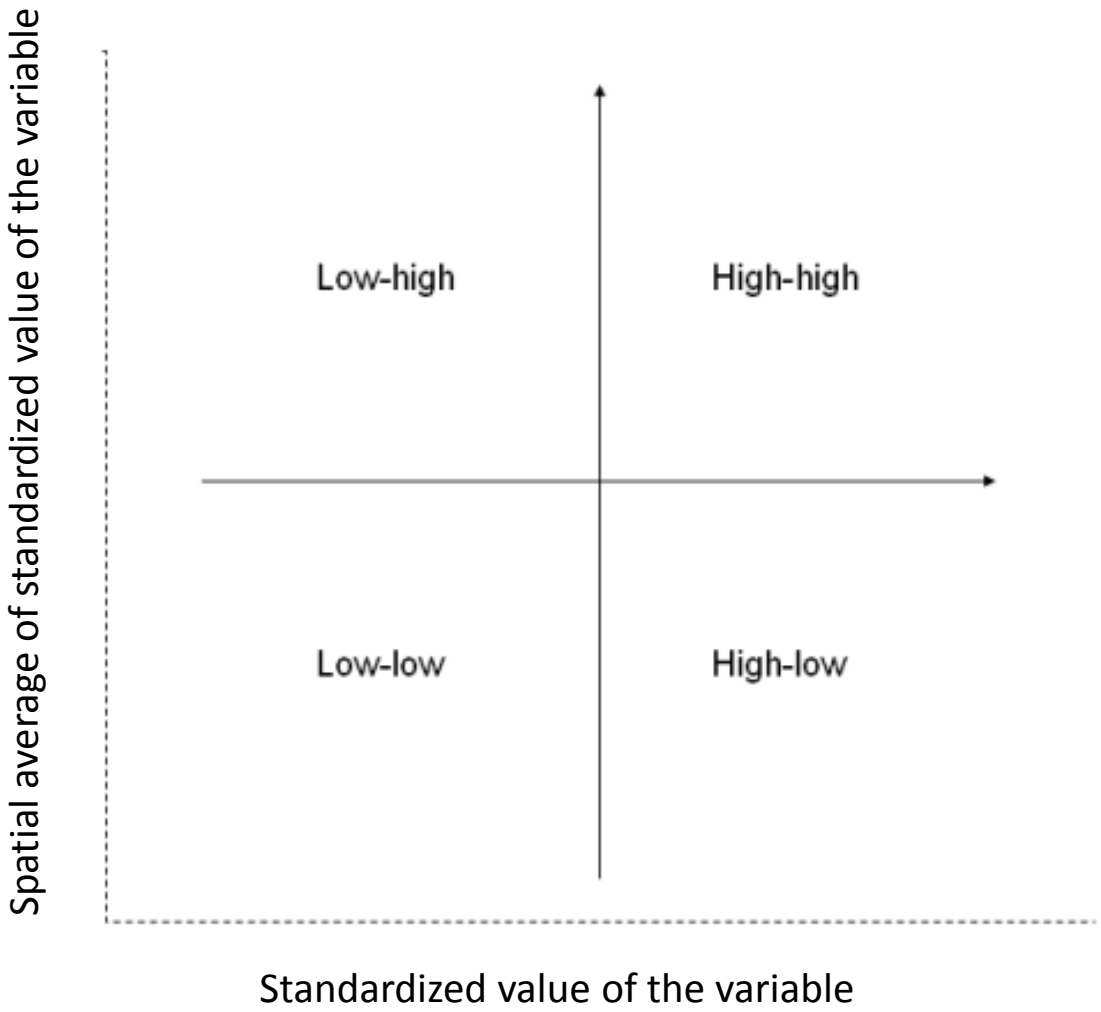
Gráfico de dispersão de Moran

I de Moran como o coeficiente de uma regressão de variáveis padronizadas

Gráfico de dispersão de Moran : associação linear entre Wz (no eixo y) e z (no eixo x)

Cada ponto é um par (zi, Wzi); inclinação da reta de regressão é a estatística I de Moran

Quadro esquemático do diagrama de dispersão de Moran



Exercício 8 (Moran scatter plot)

Exercício 7

Detectar a influência de *outliers*

- ± dois desvios-padrão

Como os *outliers* detectados influenciam a estatística *I* de Moran?

Usar o menu do GeoDa:

Space > Univariate Moran's I

<u>Dados</u>: Brazil_UF (Y00); Angola (PIB_PC_DLR)

Inferência

Hipótese nula: aleatoriedade espacial

- O padrão espacial observado é igualmente provável a qualquer outro padrão espacial
- Valores em uma localização não dependem dos valores em outras localizações (vizinhas)
- Sob a hipótese de aleatoriedade espacial, a localização dos valores pode ser alterada sem afetar a informação de conteúdo dos dados

Inferência computacional

Permutações

- Ajuda a determinar a probabilidade de se observar o valor da estatística I de Moran da distribuição em questão sob condições de aleatoriedade espacial

- n permutações dos valores da distribuição distribuídos aleatoriamente pelo espaço

Inferência computacional

Permutações (cont.)

- Distribuição de referência do I de Moran
- Aproximação normal: $E(I) = -\dfrac{1}{n-1}$
- Diferente de zero, mas tende a zero quando $n \to \infty$
- Pseudo-significância

Exercício 9 (inference)

Ilustrar o conceito de autocorrelação espacial global contrastando dados reais altamente correlacionados no espaço *versus* mesmas observações distribuídas aleatoriamente no espaço

Box Map, gráfico de dispersão de Moran

Dados: Grid100 (ZAR09, RANZAR09)

Análise global *versus* local

Análise global

- Uma estatística resume o padrão
- *Clustering*
- Homogeneidade

Análise local

- Uma estatística para cada localização
- *Clusters*
- Heterogeneidade

LISA: definição

Local **I**ndicator of **S**patial **A**utocorrelation

Anselin (1995)

- Estatística espacial local
- Indica autocorrelação espacial significante para cada localização

Relação local-global

- Soma do LISA é proporcional ao indicador de autocorrelação espacial global correspondente

Estatística de Moran local

$$l_i = (z_i / m_2) \sum_j w_{ij}\, z_j$$

$$m_2 = \sum_i z_i^2, \qquad \sum_i l_i = nl, \qquad l = \sum_i l_i / n$$

Link local-global

Global é a média dos locais

Inferência

Computacional

Permutação condicional

- Mantém fixo o valor em *i*, permuta os demais

Mapa de significância LISA

Localizações com estatísticas locais significantes

Análise de sensibilidade para o p-valor

Mapa coroplético

Unidades espaciais destacadas pelo nível de significância

Localizações não-significantes não são destacadas

Mapa de cluster LISA

Apenas as localizações significantes

- Mesmas do mapa de significância

Tipos de autocorrelação espacial

- *Clusters* espaciais **(+)**
  - *high-high* (vermelho), *low-low* (azul)
- *Outliers* espaciais **(-)**
  - *high-low* (rosa), *low-high* (azul claro)

Clusters espaciais e *outliers* espaciais

Outliers espaciais

- Localizações individuais

Clusters espaciais

- Núcleo do *cluster* no mapa LISA
- *Cluster* também inclui os vizinhos
- Usar p-valor < 0.001 para identificar núcleos de *clusters* relevantes e seus vizinhos

Limitações

Clusters LISA e *hot spots*

- Sugere localizações interessantes
- Sugere estruturas espaciais significantes
- Não explica

Necessidade de se considerar relações multivariadas

- Autocorrelação espacial univariada devido a outras covariadas
- Econometria espacial

Exercício 10 (LISA)

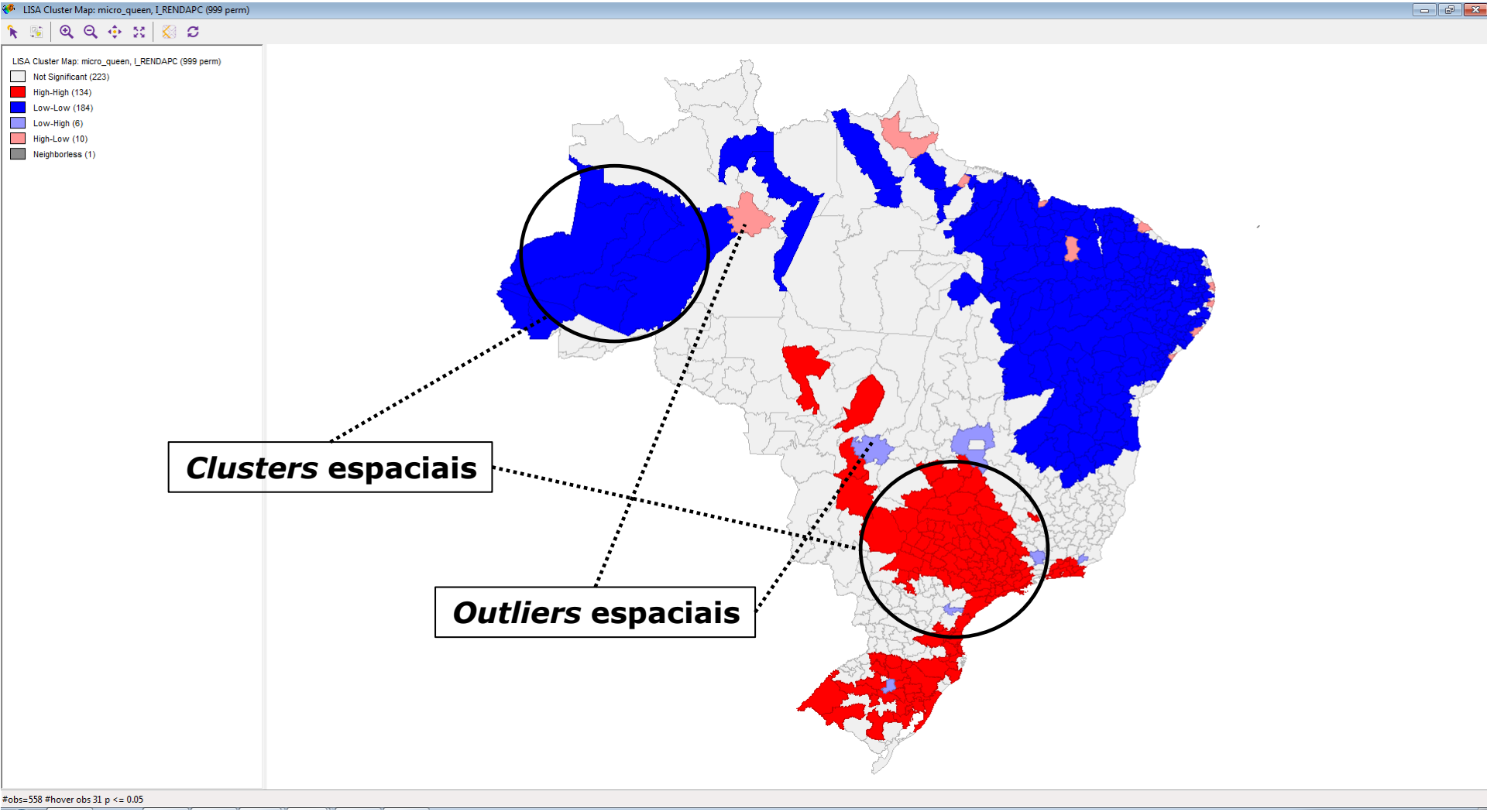
Computar a estatística de Moran local, o mapa de significância e o mapa de *cluster*

Verificar a sensibilidade do mapa de *cluster* ao número de permutações e ao nível de significância

Interpretar a noção de *cluster* espacial e *outlier* espacial

Dados: Brazil_MR (RENDAPC); Angola (PIB_PC_DLR)

Exercício 10 (LISA)



Exercício 11 (Conditional local cluster map)

Os mapas de *cluster* podem ser visualizados como mapas condicionais.

Opção: **Show As Conditional Map**

Dados: Brazil_MR (RENDAPC, ESGT, M1SM); Angola (PIB_PC_DLR, TX_PRIMAR, S_AGUA)

Exercício 12 (crime em MG)

Almeida, Haddad e Hewings (2005)

Dados de crime da Secretaria de Segurança Pública do Estado de Minas Gerais

Taxa de homicídio e tentativas de homicídio por 100.000 habitantes (1994, 1995)

Data: Crime_MG

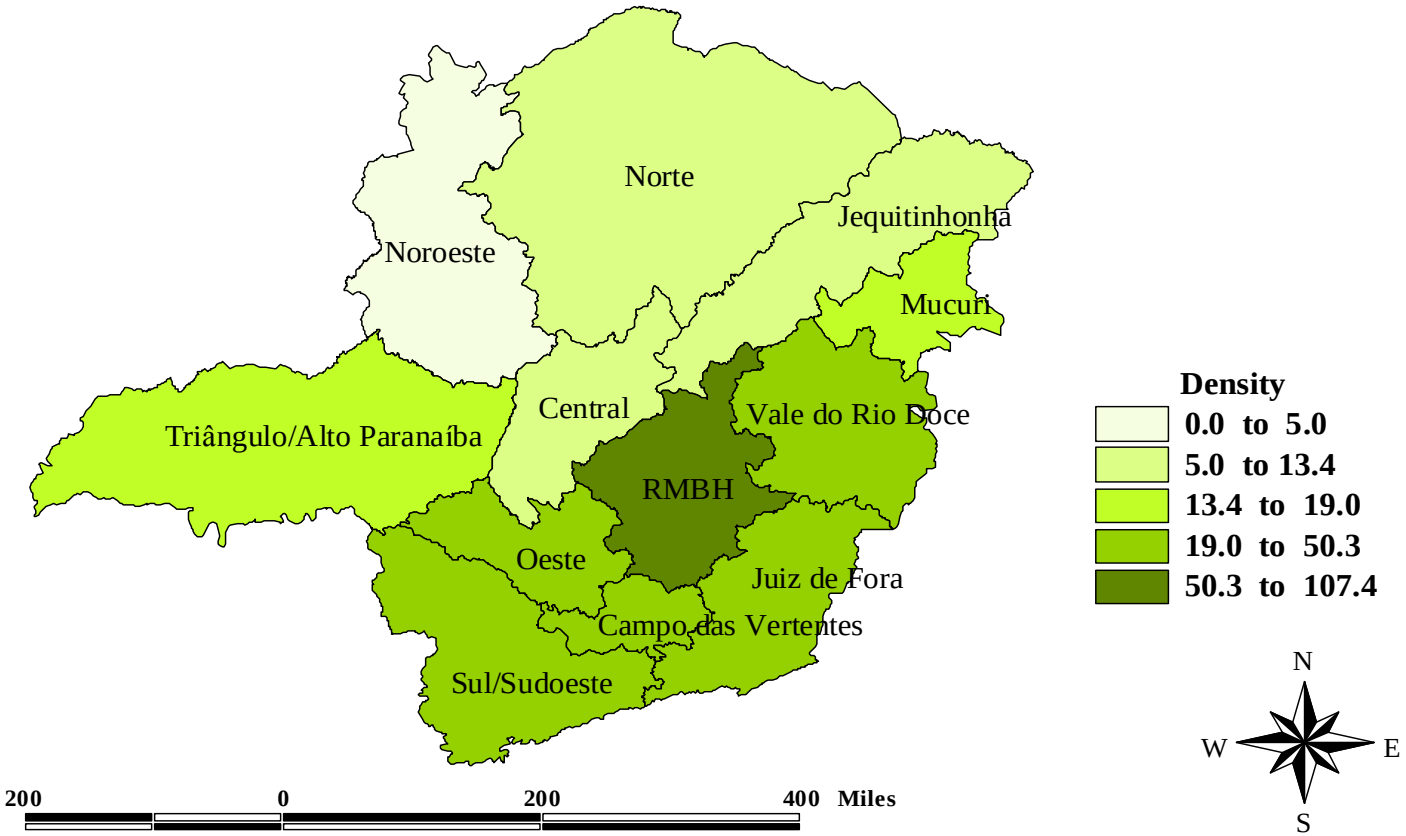
Exercício 12 (crime em MG) – cont.

Table 1: Area, Population and Production in Minas Gerais (1996)

Region	*Area*	*Population*	*Production*
Noroeste	10.7	1.9	1.6
Norte	21.7	8.7	4.4
Jequitinhonha	8.6	4.1	1.1
Vale do Mucuri	3.4	2.4	1.0
Triângulo/Alto Paranaíba	15.5	10.2	11.7
Central	5.4	2.2	1.6
RMBH	6.7	29.9	44.6
Vale do Rio Doce	7.2	9.0	9.1
Oeste	4.1	4.6	3.8
Sul/Sudoeste	8.5	12.5	10.9
Campo das Vertentes	2.1	2.9	1.9
Zona da Mata	6.1	11.6	8.3

Exercício 12 (crime em MG) – cont.

Map 1: Demographic Density of Minas Gerais (1995)



Exercício 12 (crime em MG) – cont.

Visualização das taxas (distribuição espacial)

- Sugestão de *clusters* espaciais

Teste para autocorrelação espacial global

- Estatística *I* de Moran (*queen* e *rook*)
- Gráfico de dispersão de Moran
- Influência de *outliers*

Teste para autocorrelação espacial local

- LISA
- Análise de *clusters* espaciais

Associação tempo-espaço

Autocorrelação espacial bivariada

Observações em dois pontos no tempo

- Defasagens espaciais

Processos de difusão para fora

- Do centro para a periferia

Processos de difusão para dentro

- Da periferia para o centro

Atividade: "Desenvolvimento Territorial de São Paulo: Diagnóstico Espacial"

Objetivo: utilizar técnicas de AEDE para analisar a distribuição espacial de indicadores socioeconômicos do Estado de São Paulo

1. Identifique **padrões espaciais** de desenvolvimento territorial

2. Identifique **regiões** relativamente mais/menos desenvolvidas

3. Quais os elementos relevantes para entendermos as configurações espaciais encontradas?

Atividade: “Desenvolvimento Territorial de São Paulo: Diagnóstico Espacial” (cont.)

Dados municipais

- 48 variáveis (selecionem “sabiamente”!)

Apresentação e discussão em classe:

- PowerPoint (12-15 slides)

Grupos de até quatro alunos

Data de entrega: TBD

Referências

Estas notas de aula foram adaptadas do material elaborado pelo Prof. Sergio Rey para o curso "*Geographic Information Analysis* – GPH 483/598", ministrado no primeiro semestre de 2014 na *School of Geographical Sciences and Urban Planning* da Universidade Estadual do Arizona (ASU).

As notas também incluem material preparado pelo Prof. Eduardo Haddad para o curso "Economia Regional e Urbana", realizado anualmente no Departamento de Economia da Universidade de São Paulo (USP).