



NEDUR

Núcleo de Estudos em Desenvolvimento
Urbano e Regional
Universidade Federal do Paraná

Análise Exploratória de Dados Espaciais

Alexandre Porsse^Φ • Vinícius Vale^Φ

^Φ Professor do Departamento de Economia e do Programa de Pós-Graduação em Desenvolvimento Econômico (PPGDE) da Universidade Federal do Paraná (UFPR) e Pesquisador do Núcleo de Estudos em Desenvolvimento Urbano e Regional (NEDUR)

Material desenvolvido para a disciplina Economia Regional e Urbana do Curso de Ciências Econômicas da Universidade Federal do Paraná (UFPR). Os professores autorizam o uso desse material em outros cursos desde que devidamente citados os créditos.

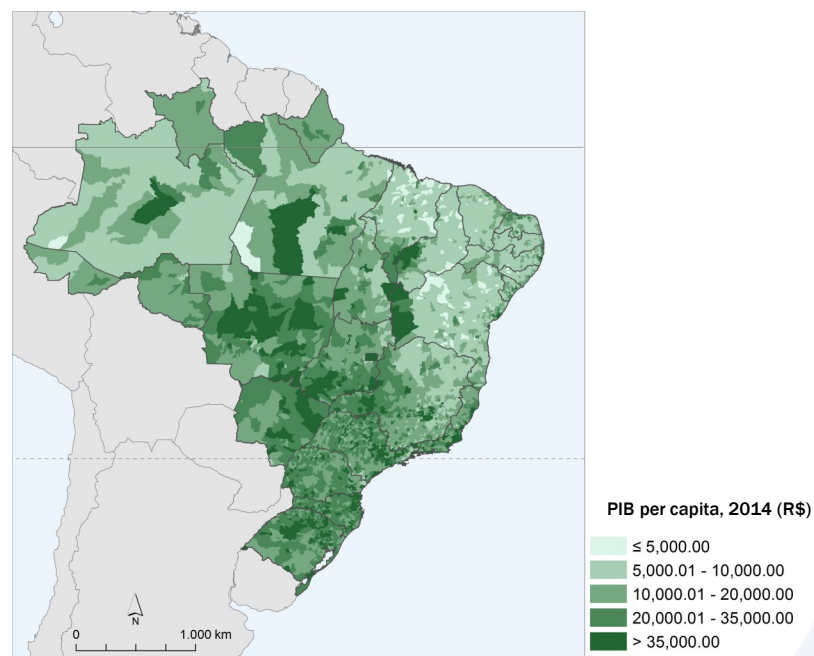
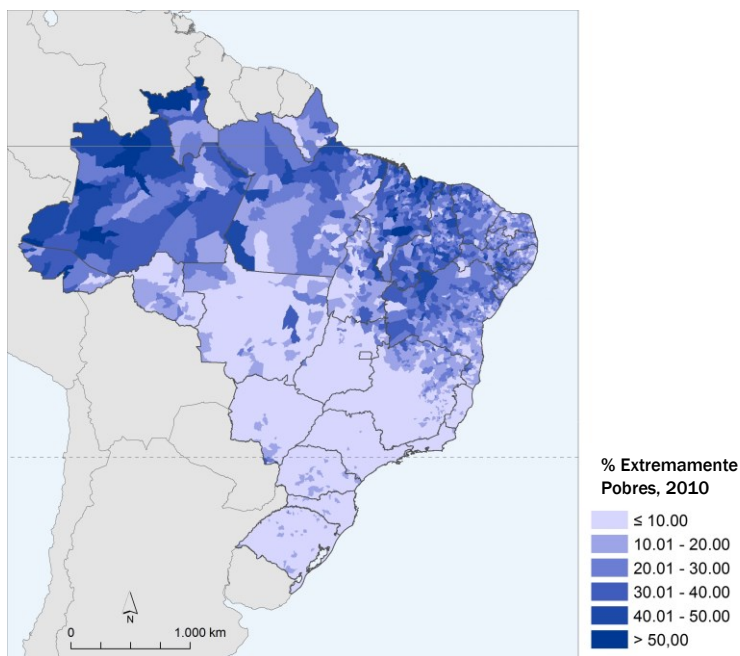
Agosto/2020

- Introdução
- Dados espaciais
- Matrizes de ponderação espacial
- Análise Exploratória de Dados Espaciais (AEDE)
- Autocorrelação espacial global (univariada e bi-variada)
- Autocorrelação espacial local (univariada e bi-variada)
- Aplicações usando o GeoDa:
 - <https://geodacenter.github.io>

Introdução

- Primeira Lei da Geografia (Waldo Tobler, 1970):

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



Fonte: Adaptado de Haddad et al. (2017).

- **O que é Análise Espacial?**

Consiste em métodos e técnicas que buscam identificar **padrões de associação espacial (dependência e heterogeneidade espaciais)** nos fenômenos socioeconômicos com o objetivo de compreender suas estruturas e dinâmicas no espaço, contribuindo para a formulação e avaliação de políticas públicas e privadas.

- **Efeitos espaciais:**

- **Dependência espacial:** interação dos agentes através do espaço.
- **Heterogeneidade espacial:** instabilidade estrutural através do espaço.

- **Fontes de dependência espacial:**
 - **Processos de difusão:**
 - Difusão tecnológica (transbordamento tecnológico)
 - Avanço de uma epidemia
 - **Processos de trocas e transferências:**
 - Relações comerciais (bens e serviços)
 - Multiplicador regional da renda (transferência de renda)
 - **Interações estratégicas:**
 - Competição e cooperação
 - **Dispersão ou espraiamento:**
 - Fronteira agrícola, migração, etc.

Dados espaciais

- O uso de informações georreferenciadas no processo decisório tem aumentado sistematicamente devido ao avanço da tecnologia:
 - Hardware:
 - Aumento da capacidade computacional
 - Programas SIG (Sistema de Informações Geográficas):
 - **QGIS**, ArcGis, ArcView, MapInfo, TerraView, etc.
 - Softwares voltados para a estatística espacial:
 - **Geoda**, SpaceStat, Matlab, R, etc.

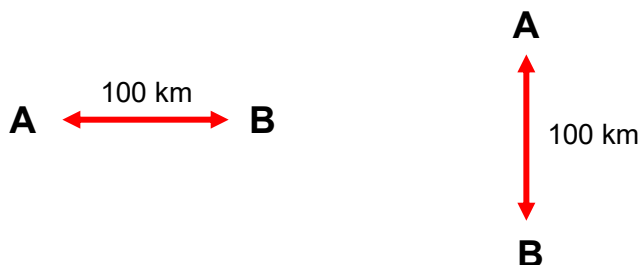
Dados espaciais

- **Definição:**

- Dados espaciais denotam a variação de algum fenômeno tendo a preocupação em determinar **onde** ocorre tal variação (localização).
- Os dados espaciais mostram a magnitude da variação do atributo do fenômeno em estudo e fornece a referência explícita em termos de **localização geográfica** deste atributo.

- **Conceito de isotopia:**

- Apenas a **distância** é importante na avaliação da interação espacial do fenômeno e não a direção dessa interação espacial.



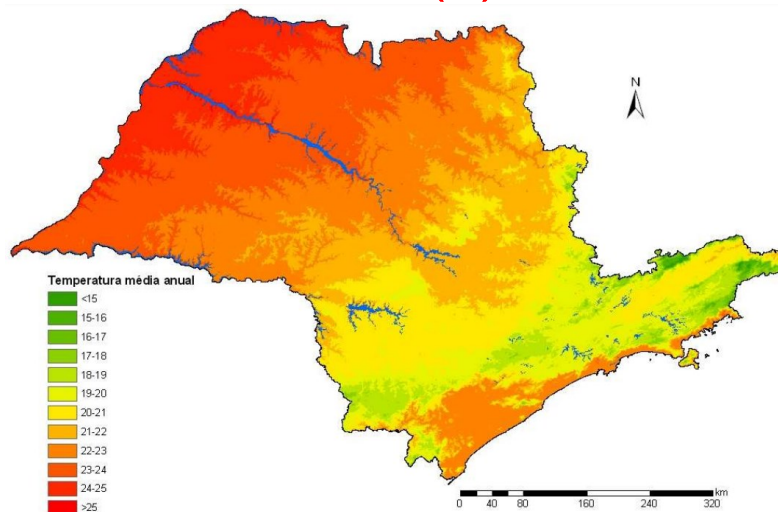
Dados espaciais

- **Tipos de dados espaciais:**
 - Dados geoestatísticos
 - Dados na forma de pontos
 - Dados na forma de áreas (polígonos)

Dados espaciais

- **Dados geoestatísticos:**
 - Distribuição contínua no espaço:
 - Exemplo: pressão atmosférica, temperatura.
 - Pouco uso para fenômenos socioeconômicos.

Temperatura média anual do ar (°C) no Estado de São Paulo

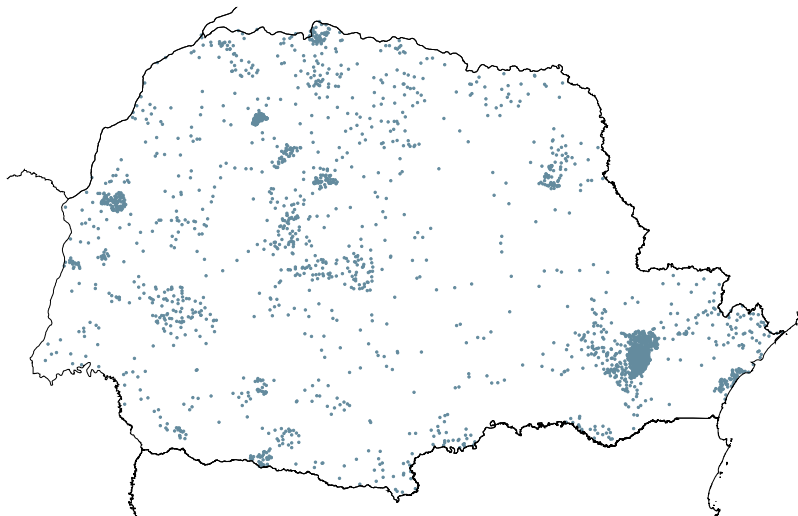


Fonte: Boletim de Pesquisa e Desenvolvimento – Embrapa (2014).

Dados espaciais

- **Dados na forma de pontos:**
 - Amostra de eventos no espaço com distribuição discreta.
 - Muito usado em áreas como epidemiologia, medicina e criminologia.

Homicídios Dolosos nos Municípios Paranaenses - 2015

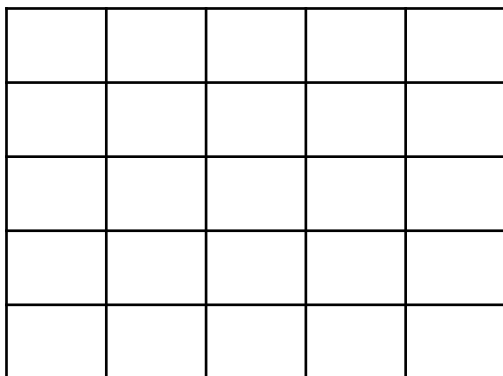


Dados espaciais

- **Dados na forma de áreas (polígonos):**
 - Distribuição discreta no espaço.
 - É o tipo de dado mais usado na Economia.

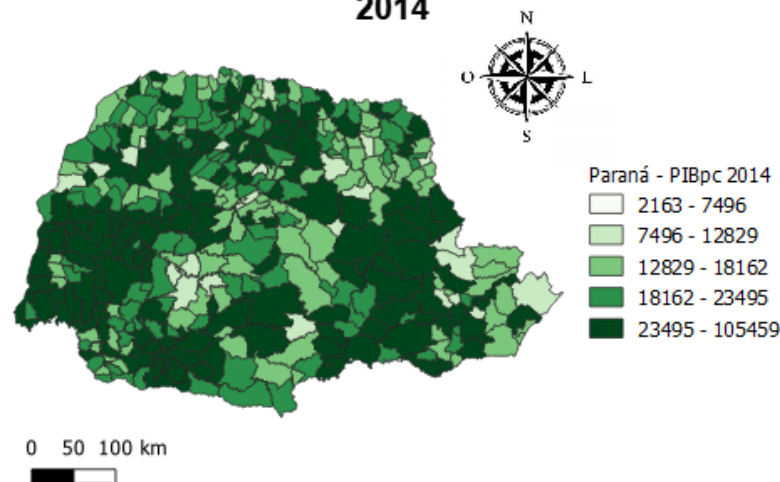
Polígonos regulares

(gride 5 por 5)



Polígonos irregulares

PIB municipal per capita - Paraná 2014



Dados espaciais

- A análise de dados espaciais deve ser realizada de forma cuidadosa, pois pode ser sensível a problemas muito específicos:
 - **Falácia ecológica** (inferência ecológica)
 - Problema da **unidade de área modificável**
 - **Efeito beirada**
 - Influência de **outliers espaciais**

- **Falácia ecológica (inferência ecológica):**
 - Inferência sobre o comportamento do indivíduo a partir da análise de dados agregados.
- **Problema da unidade de área modificável (MAUP):**
 - **Problema de escala:**
 - Os resultados de dependência espacial são influenciados pelo número de unidades espaciais (escala: municípios, microrregiões, etc.).
 - **Problema de zoneamento:**
 - Mesmo quando as escalas espaciais são constantes, os resultados são afetados pelas várias formas de fazer combinações das unidades espaciais.

Dados espaciais

- **Efeito beirada:**
 - Observações próximas da fronteira podem estar correlacionadas espacialmente com regiões que não fazem parte da área de estudo.
- **Influência de *outliers* espaciais:**
 - Observações espaciais podem apresentar uma dependência espacial distinta do restante das outras observações vizinhas.

Matrizes de ponderação espacial

- Uma matriz de ponderação espacial é uma matriz de dimensão $n \times n$.
- Reflete o arranjo das interações espaciais por meio de pesos espaciais w_{ij} *que* representam o **grau de conexão** entre as regiões segundo algum critério de proximidade.
- Princípio:
 - Regiões mais conectadas entre si interagem mais do que regiões menos conectadas.
- Critérios de proximidade:
 - Geográfica
 - Socioeconômica

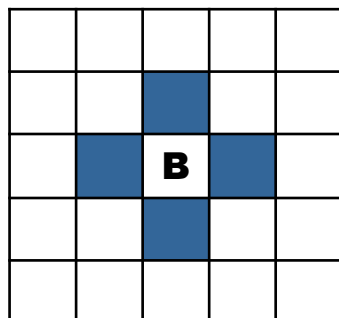
Matrizes de ponderação espacial

- **Tipologia das matrizes de ponderação espacial:**
- **Matrizes geográficas:**
 - Contiguidade
 - Distância geográfica
- **Matrizes socioeconômicas:**
 - Distância socioeconômica:
 - Baseada na similaridade
 - Baseada na dissimilaridade
 - Baseada nos fluxos

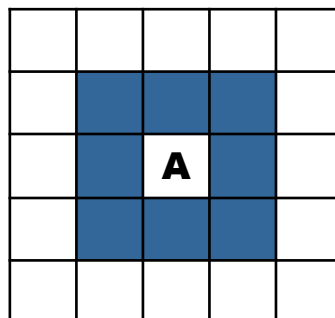
Matrizes geográficas de contiguidade

- Duas regiões são vizinhas caso elas partilhem uma fronteira física comum.
- **Matrizes binárias de contiguidade:**
 - Torre (Rook)
 - Rainha (Queen)
 - Bispo (Bishop)

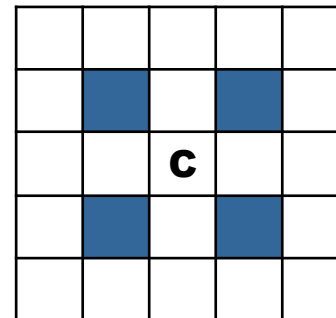
Torre



Rainha



Bispo



Matrizes geográficas de contiguidade

- Por convenção, assume-se que W representa a matriz de vizinhança ($n \times n$) tal que:

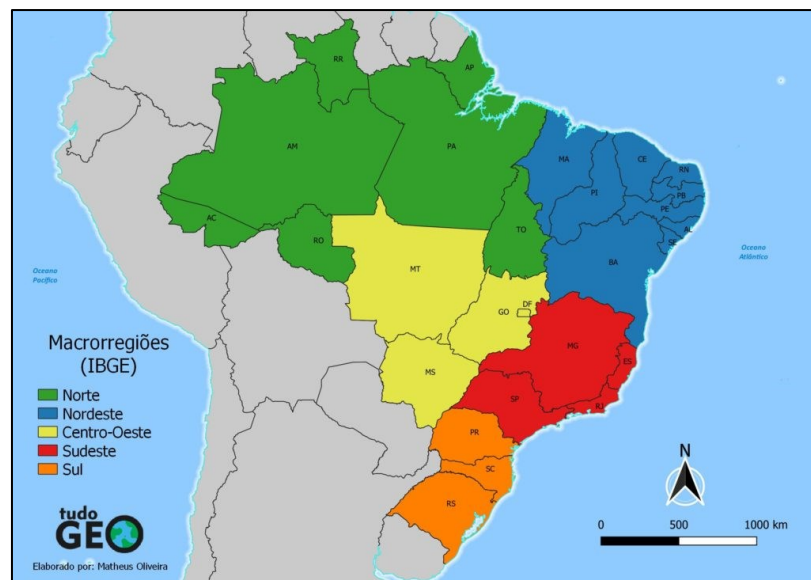
$$w_{ij} = \begin{cases} 1 & \text{se } i \text{ e } j \text{ são contíguos} \\ 0 & \text{se } i \text{ e } j \text{ não são contíguos} \end{cases}$$

- Desvantagens das matrizes de convenção “rainha”, “bispo” e “torre”:
 - **Conectividade desbalanceada:** pode haver regiões com muitos vizinhos e regiões com poucos vizinhos.
 - **Problema da “ilha”**

Matrizes geográficas de contiguidade

- Matriz binária para as macrorregiões brasileiras:

Reg.	N	NE	CO	SE	S
N	0	1	1	0	0
NE	1	0	1	1	0
CO	1	1	0	1	1
SE	0	1	1	0	1
S	0	0	1	1	0



Matrizes geográficas de contiguidade

- **Tipos de contiguidade:**
 - **Primeira ordem:**
 - Conjunto dos vizinhos diretos
 - **Segunda ordem:**
 - Conjunto dos vizinhos dos vizinhos das regiões, sem que esses vizinhos de segunda ordem sejam, eles próprios, vizinhos de primeira ordem.
 - **Ordem superior**

Matrizes de distância geográfica

- Os pesos espaciais podem ser definidos pela distância geográfica.
- **Matriz k -vizinhos mais próximos:**
 - Definição pela distância absoluta
 - Definição pela distância inversa
- **Métrica da distância:**
 - Distância euclidiana (na reta)
 - Distância no arco

Matrizes de distância geográfica

- **Matriz k-vizinhos mais próximos:**
 - Matriz binária baseada na **distância geográfica**:

$$w_{ij}(k) = \begin{cases} 1 & \text{se } d_{ij} \leq d_i(k) \\ 0 & \text{se } d_{ij} > d_i(k) \end{cases}$$

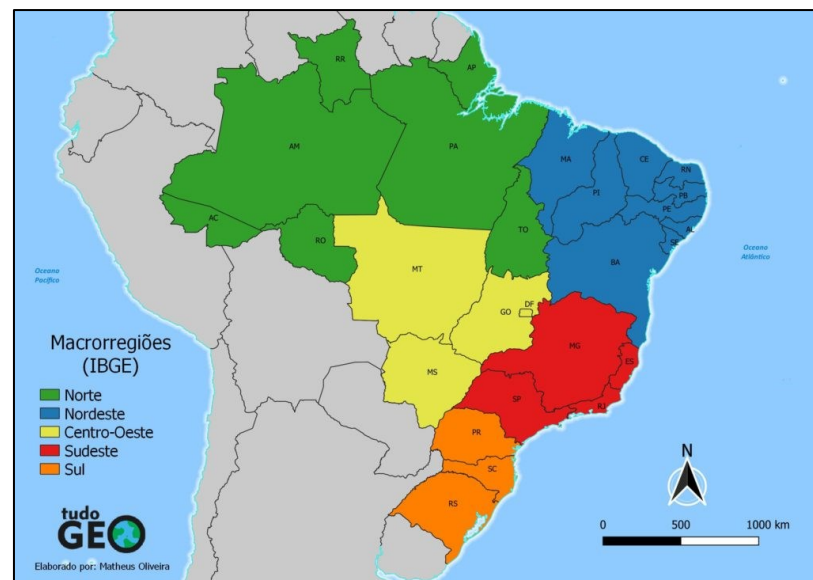
em que $d_i(k)$ é uma distância de corte tal que todas as regiões tenham o mesmo número de vizinhos (k).

- **Vantagem:** tem-se conectividade equilibrada (mesmo número de vizinhos para todas as unidades espaciais).

Matrizes de distância geográfica

- Matriz de dois vizinhos mais próximos (K=2) para as macrorregiões brasileiras:

Reg.	N	NE	CO	SE	S
N	0	1	1	0	0
NE	1	0	1	0	0
CO	0	0	0	1	1
SE	0	0	1	0	1
S	0	0	1	1	0



Matrizes de distância geográfica

- **Matriz k-vizinhos mais próximos:**
 - Matriz binária baseada na **distância inversa**:

$$w_{ij}(k) = \begin{cases} 1/d_{ij} & \text{se } d_{ij} \leq d_i(k) \\ 0 & \text{se } d_{ij} > d_i(k) \end{cases}$$

em que $d_i(k)$ é uma distância de corte tal que todas as regiões tenham o mesmo número de vizinhos (k).

Matrizes de distância geográfica

- **Métrica da distância:**

- Distância euclidiana na reta:

$$d_{ij} = \sqrt{(u_i - v_i)^2 + (u_j - v_j)^2}$$

em que u_i é a ordenada e v_i é a abscissa, projetadas no mapa.

- Distância euclidiana no arco:

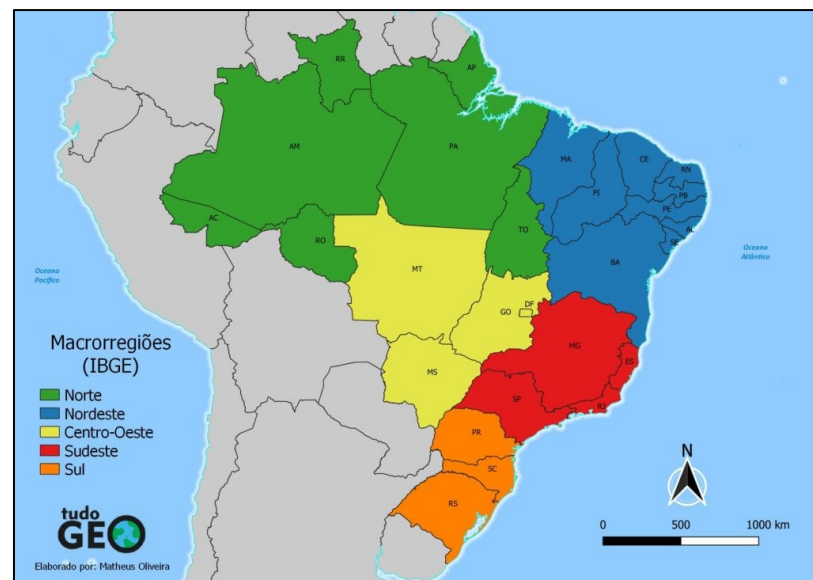
$$d_{ij} = R \cos^{-1} [\sin \phi_1 \sin \phi_2 + \cos \phi_1 \cos \phi_2 \cos(\kappa_1 - \kappa_2)]$$

em que R é o raio da Terra em torno do Equador (6.378km), ϕ é a latitude e κ é a longitude.

Matrizes de distância geográfica

- Matriz de **distância** (km) entre as macrorregiões brasileiras:

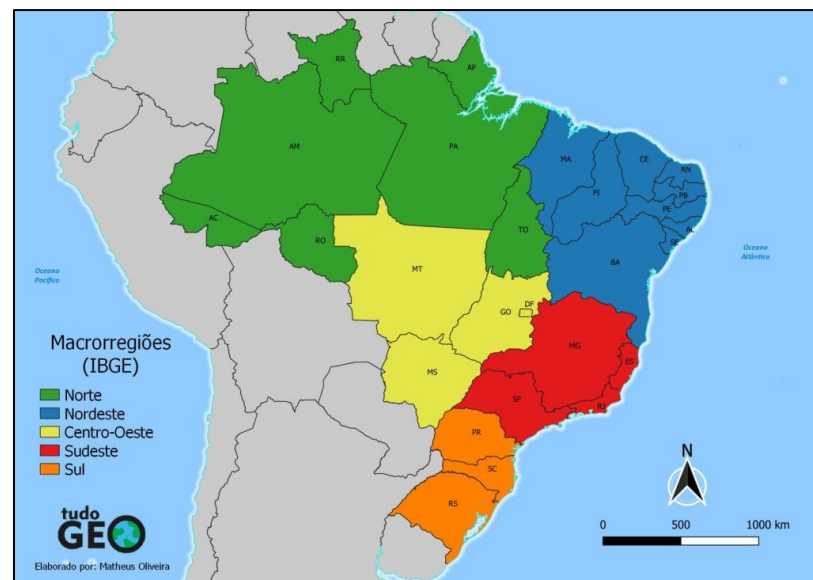
Reg.	N	NE	CO	SE	S
N	0	2074	2017	2933	3854
NE	2074	0	2332	2660	3779
CO	2017	2332	0	296	1847
SE	2933	2660	926	0	1109
S	3854	3779	1847	1109	0



Matrizes de distância geográfica

- Matriz de **distância inversa** entre as macrorregiões brasileiras:

Reg.	N	NE	CO	SE	S
N	0	0,00048	0,00050	0,00034	0,00026
NE	0,00048	0	0,00043	0,00038	0,00026
CO	0,00050	0,00043	0	0,00108	0,00054
SE	0,00034	0,00038	0,00108	0	0,00090
S	0,00026	0,00026	0,00054	0,00090	0



Matriz de pesos espaciais normalizada

- A normalização da matriz W gera a **matriz de pesos espaciais** utilizada na análise espacial.
- Essa normalização é formulada como:

$$w_{ij}^* = \frac{w_{ij}}{\sum_j w_{ij}} \quad \text{tal que} \quad \sum_j w_{ij}^* = 1$$

- **Observação:** a normalização na linha da matriz de pesos espaciais pode produzir assimetria se o número de vizinhos é diferente para cada localização:
 - A região i exerce uma influência distinta sobre a região j do que esta sobre a primeira região ($w_{ij}^* \neq w_{ji}^*$).

Matriz de pesos espaciais normalizada

Matriz original

Reg.	N	NE	CO	SE	S
N	0	1	1	0	0
NE	1	0	1	1	0
CO	1	1	0	1	1
SE	0	1	1	0	1
S	0	0	1	1	0

Matriz normalizada

Reg.	N	NE	CO	SE	S
N	0	0,500	0,500	0,000	0,000
NE	0,333	0	0,333	0,333	0,000
CO	0,250	0,250	0	0,250	0,250
SE	0,000	0,333	0,333	0	0,333
S	0,000	0,000	0,500	0,500	0

Variável espacialmente defasada

- Defasagem espacial do PIB macrorregional (R\$ Bilhões, 2000):

$$\begin{array}{ccccc}
 \text{NO} & \text{NE} & \text{CO} & \text{SE} & \text{S} \\
 \begin{pmatrix} 0,000 & 0,500 & 0,500 & 0,000 & 0,000 \\ 0,333 & 0,000 & 0,333 & 0,333 & 0,000 \\ 0,250 & 0,250 & 0,000 & 0,250 & 0,250 \\ 0,000 & 0,333 & 0,333 & 0,000 & 0,333 \\ 0,000 & 0,000 & 0,500 & 0,500 & 0,000 \end{pmatrix} & \begin{pmatrix} 50,6 \\ 144,1 \\ 76,5 \\ 636,4 \\ 193,5 \end{pmatrix} & = & \begin{pmatrix} 110,3 \\ 254,5 \\ 256,2 \\ 138,0 \\ 356,5 \end{pmatrix} \\
 \mathbf{W} & \mathbf{y} & & \mathbf{Wy}
 \end{array}$$

Operador de defasagem espacial ←

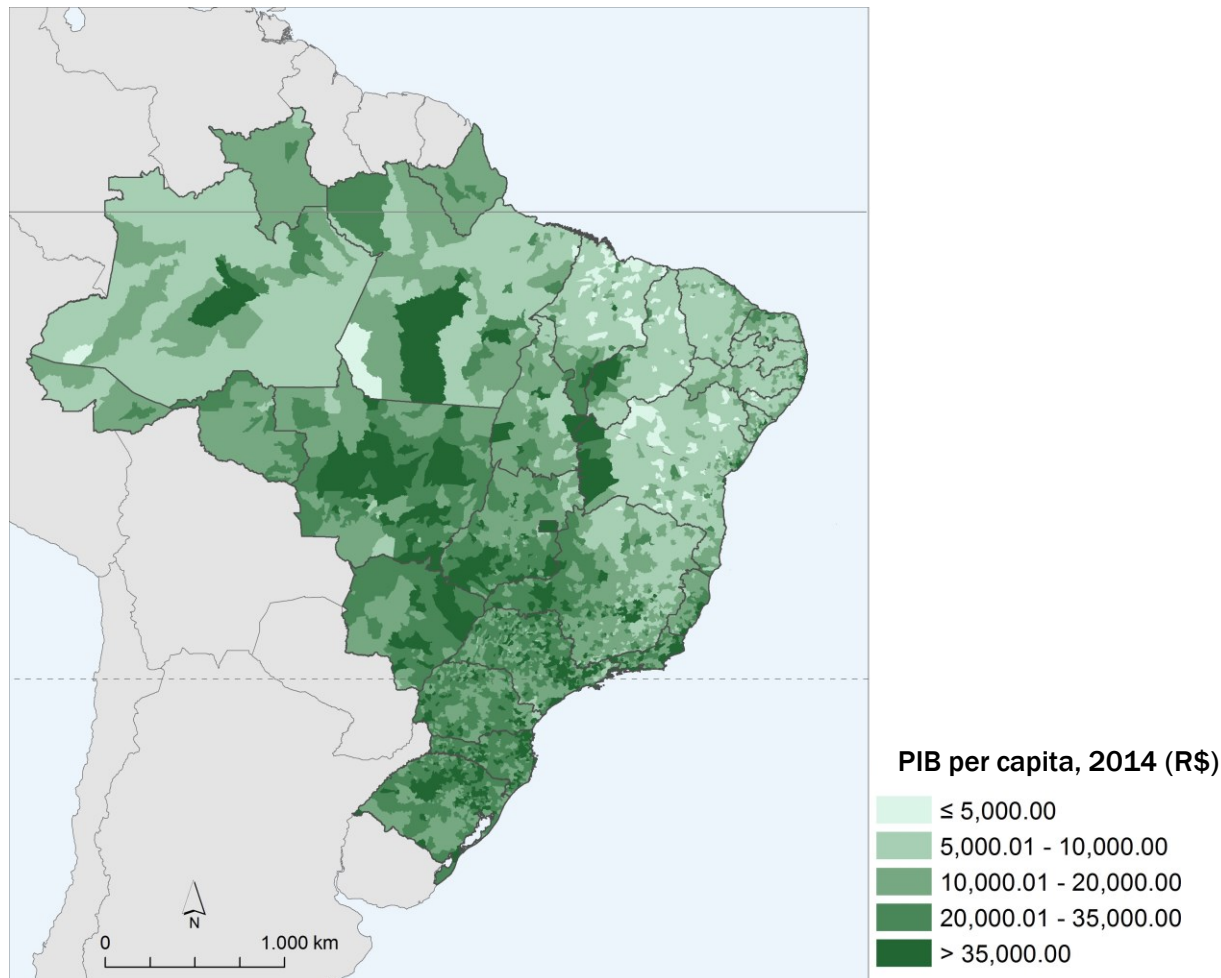
Média do valor da variável y nas regiões vizinhas

- A **AEDE** é uma coleção de técnicas para descrever e visualizar:
 - Distribuições espaciais
 - Autocorrelação espacial
 - Clusters (padrões de associação espacial)
 - *Outliers* (localidades atípicas)
 - Regimes espaciais
 - Outras formas de instabilidade espacial
- A técnica fornece subsídios para a especificação de modelos econométricos espaciais.

Variáveis usadas na AEDE

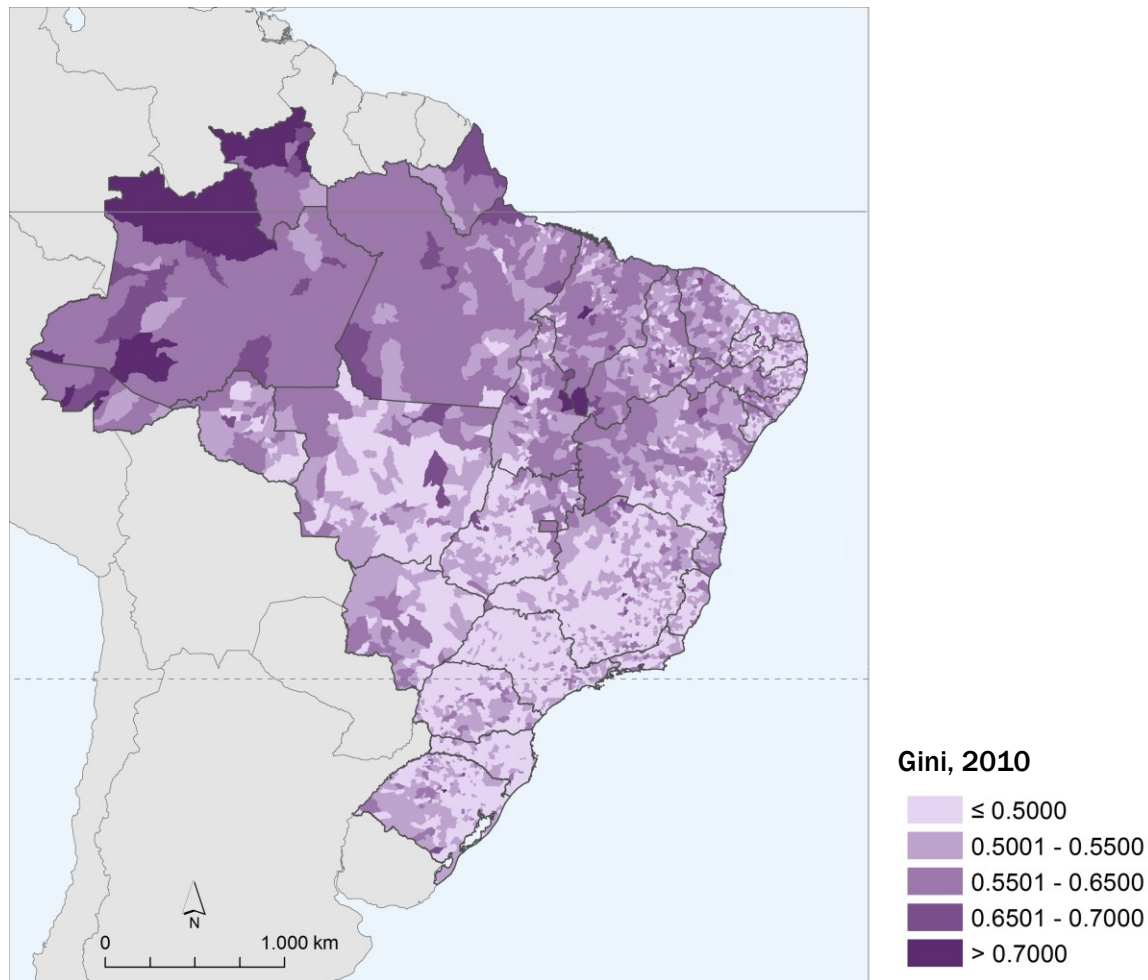
- **Variáveis extensivas** ou **absolutas** podem apresentar possível autocorrelação espacial espúria.
- **Variáveis intensivas**, normalizadas ou espacialmente densas são mais apropriadas na AEDE:
 - Variáveis per capita
 - Variáveis divididas por área, quantidade de trabalho, capital
 - Variáveis normalizadas em índices

Distribuição espacial do PIB municipal per capita



Fonte: Adaptado de Haddad et al. (2017).

Distribuição espacial do Gini



Fonte: Adaptado de Haddad et al. (2017).

Autocorrelação espacial global univariada

- Os dados do Gini (ou PIB per capita) são distribuídos aleatoriamente ou seguem algum padrão espacial sistemático?
- O primeiro passo num estudo de AEDE é testar a hipótese de que os dados espaciais sejam distribuídos aleatoriamente!
 - Aleatoriedade espacial significa que os valores de um atributo numa região não dependem dos valores deste atributo nas regiões vizinhas.

Autocorrelação espacial global univariada

- Moran (1948) propôs a elaboração de um coeficiente de autocorrelação espacial, usando a medida de autocovariância na forma de produto cruzado.
- Coeficiente I de Moran (1948) para uma variável aleatória:

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

Autocorrelação espacial global univariada

- Uma forma de visualizar a autocorrelação espacial é baseada no diagrama de dispersão de Moran. Esse diagrama representa a relação linear entre Wz e z tal como em um modelo de regressão simples:

$$Wz = \alpha + \beta z + \varepsilon$$

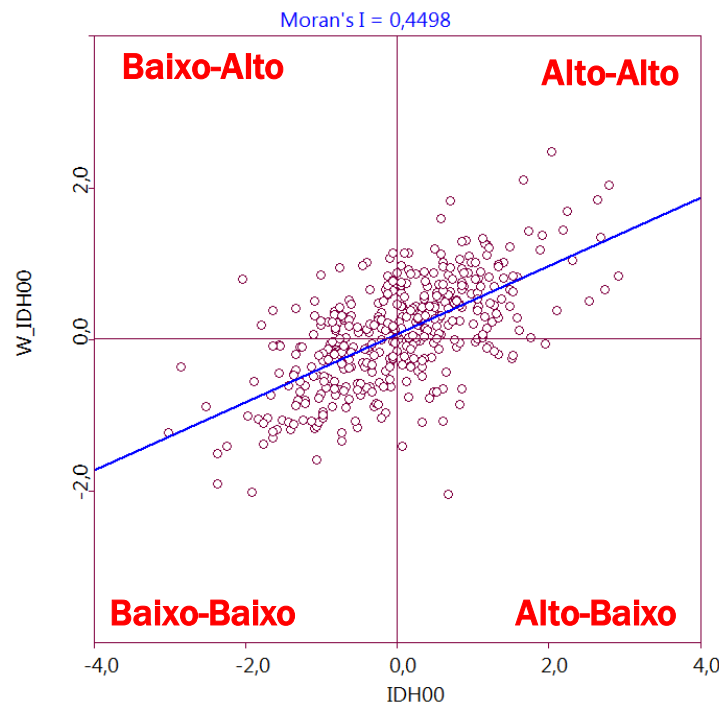
$$\hat{\beta} = I = \frac{z'Wz}{z'z}$$

- O coeficiente angular deste modelo é equivalente ao I-Moran.
- Se o coeficiente angular da reta for positivo (negativo), há evidências de que a autocorrelação espacial é positiva (negativa).

Autocorrelação espacial global univariada

- O **diagrama de dispersão de Moran** também permite classificar as unidades espaciais em 4 tipologias (quadrantes):

- Alto-Alto (AA) – *High-High* (HH)
- Baixo-Baixo (BB) – *Low-Low* (LL)
- Alto-Baixo (AB) – *High-Low* (HL)
- Baixo-Alto (BA) – *Low-High* (LH)



Autocorrelação espacial global bi-variada

- A autocorrelação espacial global pode ser averiguada também num contexto bivariado.
- O objetivo é verificar se os valores de uma variável observada numa dada região guardam uma associação com os valores de outra variável observada em regiões vizinhas.
- Coeficiente I de Moran para autocorrelação de duas variáveis aleatórias padronizadas:

$$I_{kl} = \frac{z'_k W z_l}{z'_k z_k}$$

em que $z_i = \frac{(y_i - \bar{y})}{\sigma_i}$ para todo $i = k, l$.

Autocorrelação espacial global bi-variada

- O diagrama de dispersão de Moran no caso bi-variado é baseado no modelo de regressão simples abaixo:

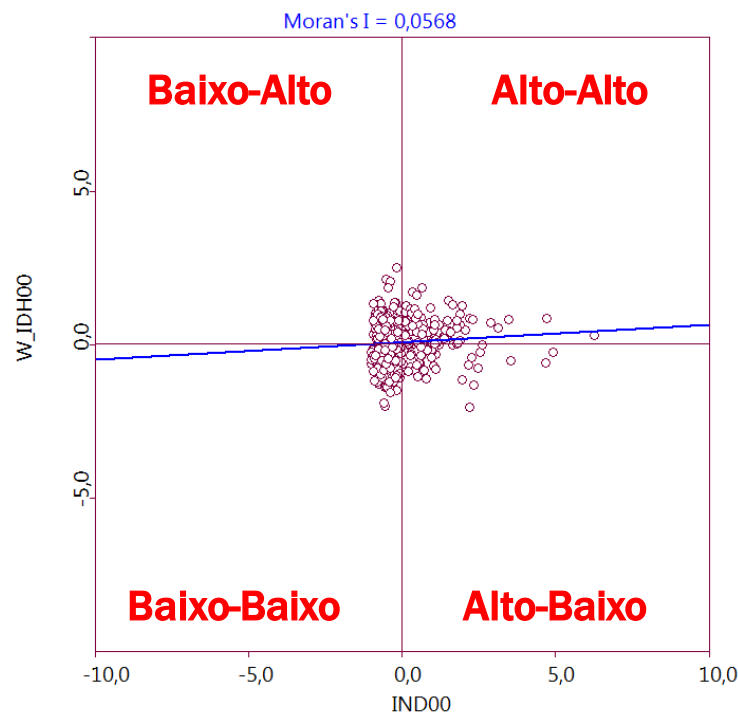
$$Wz_2 = \alpha + \beta z_1 + \varepsilon$$

$$\hat{\beta} = I^{z^1 z^2} = \frac{z_1' W z_2}{z_1' z_1}$$

- O coeficiente angular é equivalente ao I-Moran para o caso bi-variado.
- Se o coeficiente angular da reta for positivo (negativo), há evidências de que a autocorrelação espacial é positiva (negativa).

Autocorrelação espacial global bi-variada

Diagrama de Dispersão de Moran Bivariado:
W.IDH versus Grau de Industrialização - Paraná - 2000



Autocorrelação espacial local univariada

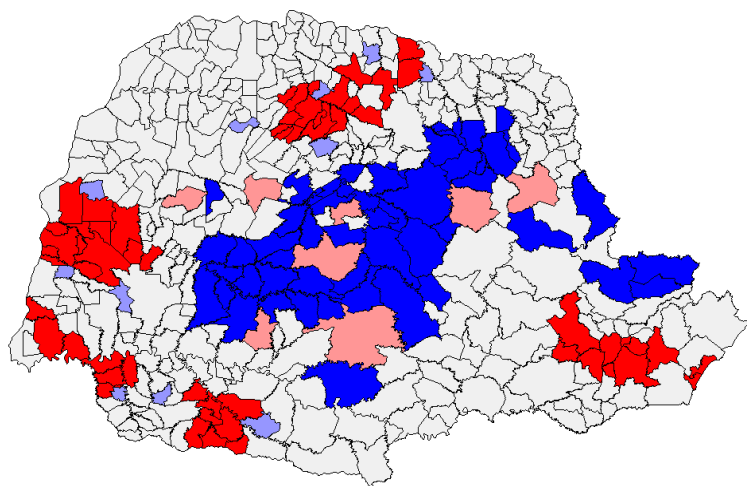
- Para observar a **autocorrelação espacial local**, podemos decompor o I-Moran global e identificar a relevância estatística das observações nos quadrantes do diagrama de Moran (AA, BB, AB, BA).
- O indicador LISA (*Local Indicator of Spatial Association*) é conhecido como **I-Moran local** para uma variável padronizada, sendo definido como:

$$I_i = z_i \sum_{j=1}^J w_{ij} z_j$$

- A partir da significância estatística dos índices LISA pode-se identificar padrões locais de homogeneidade ou heterogeneidade espacial (clusters espaciais).

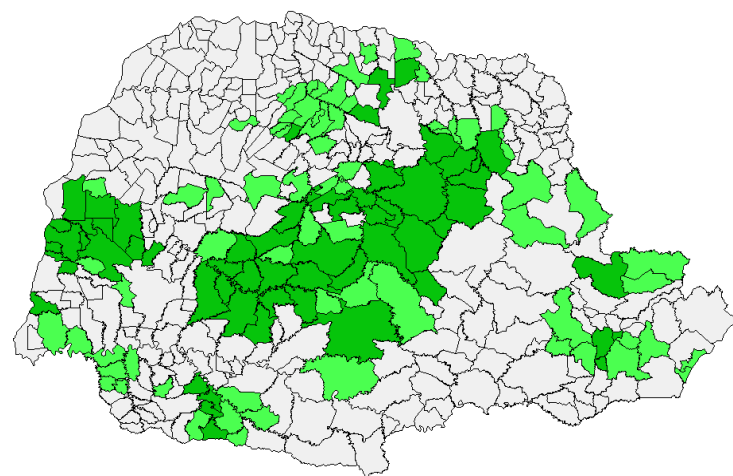
Cluster LISA univariado

IDH do Paraná – 2000



LISA Cluster: I_IDH00

- Não Significativo
- Alto-Alto
- Baixo-Baixo
- Baixo-Alto
- Alto-Baixo



Mapa de Significância
I_IDH00

- Não Significativo
- $p = 0.05$
- $p = 0.01$
- $p = 0.001$

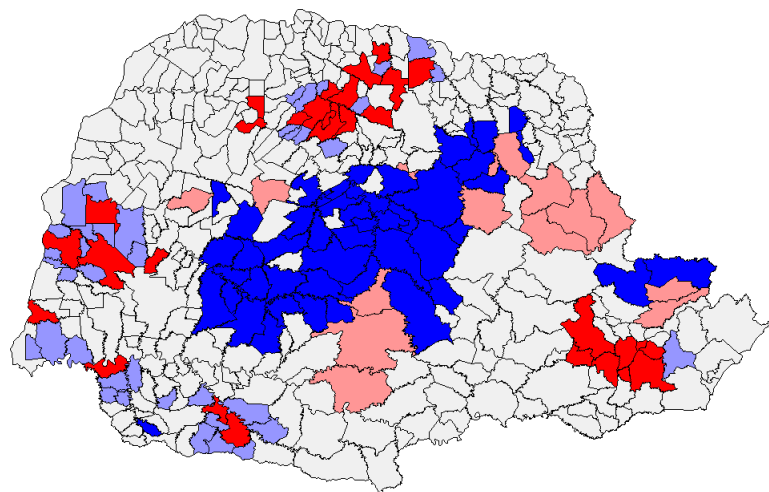
Autocorrelação espacial local bi-variada

- O coeficiente I de Moran local para um contexto bi-variado, no qual as variáveis aleatórias são padronizadas, é definido como:

$$I_i^{z_1 z_2} = z_{1i} W z_{2i}$$

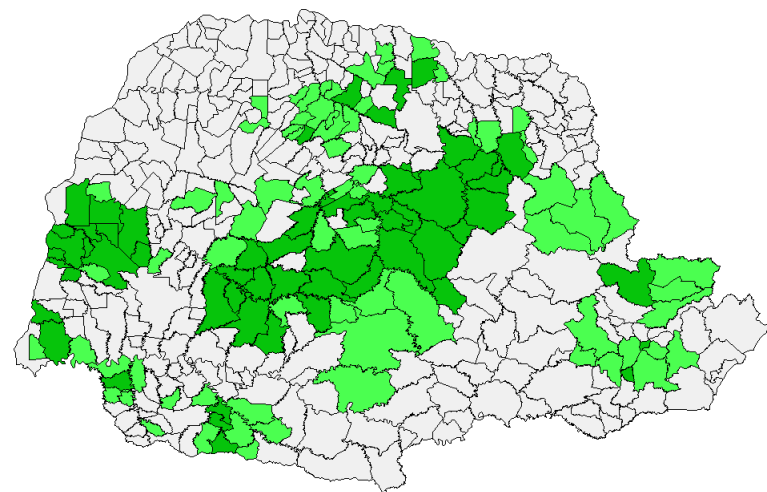
Clusters LISA bi-variado

IDH versus Grau de Industrialização no Paraná – 2000



W_IDH00

- Não Significativo
- Alto-Alto
- Baixo-Baixo
- Baixo-Alto
- Alto-Baixo



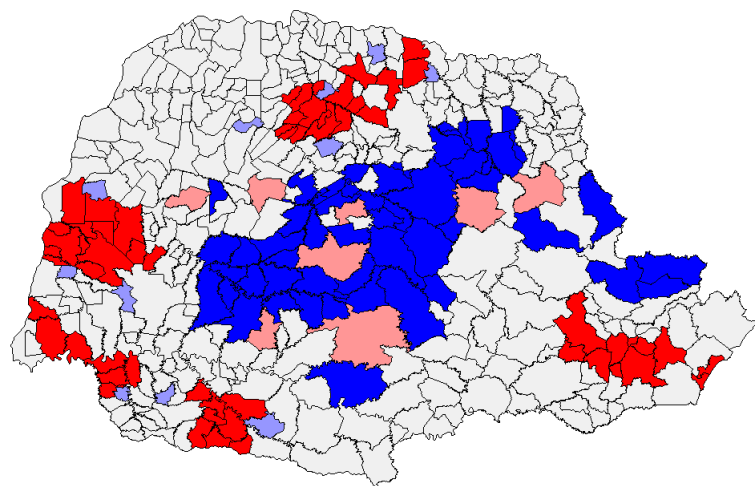
Mapa de Significância

W_IDH00

- Não Significativo
- $p = 0.05$
- $p = 0.01$
- $p = 0.001$

Clusters LISA comparados

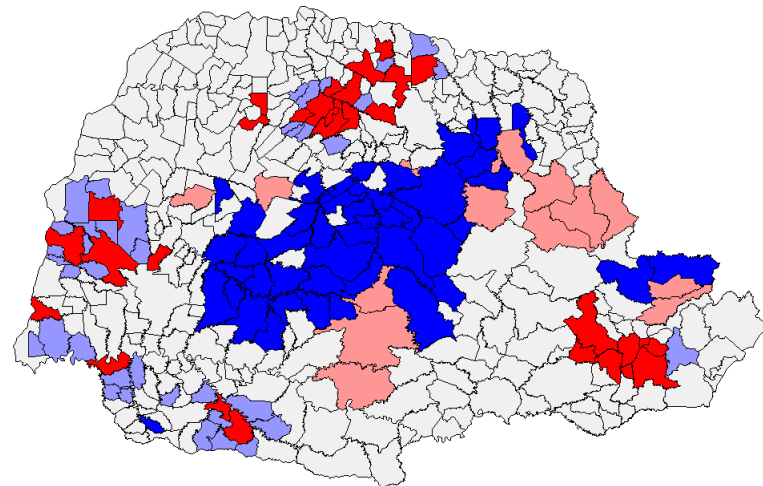
Univariado



LISA Cluster: I_IDH00



Bi-variado



W_IDH00



Outliers espaciais

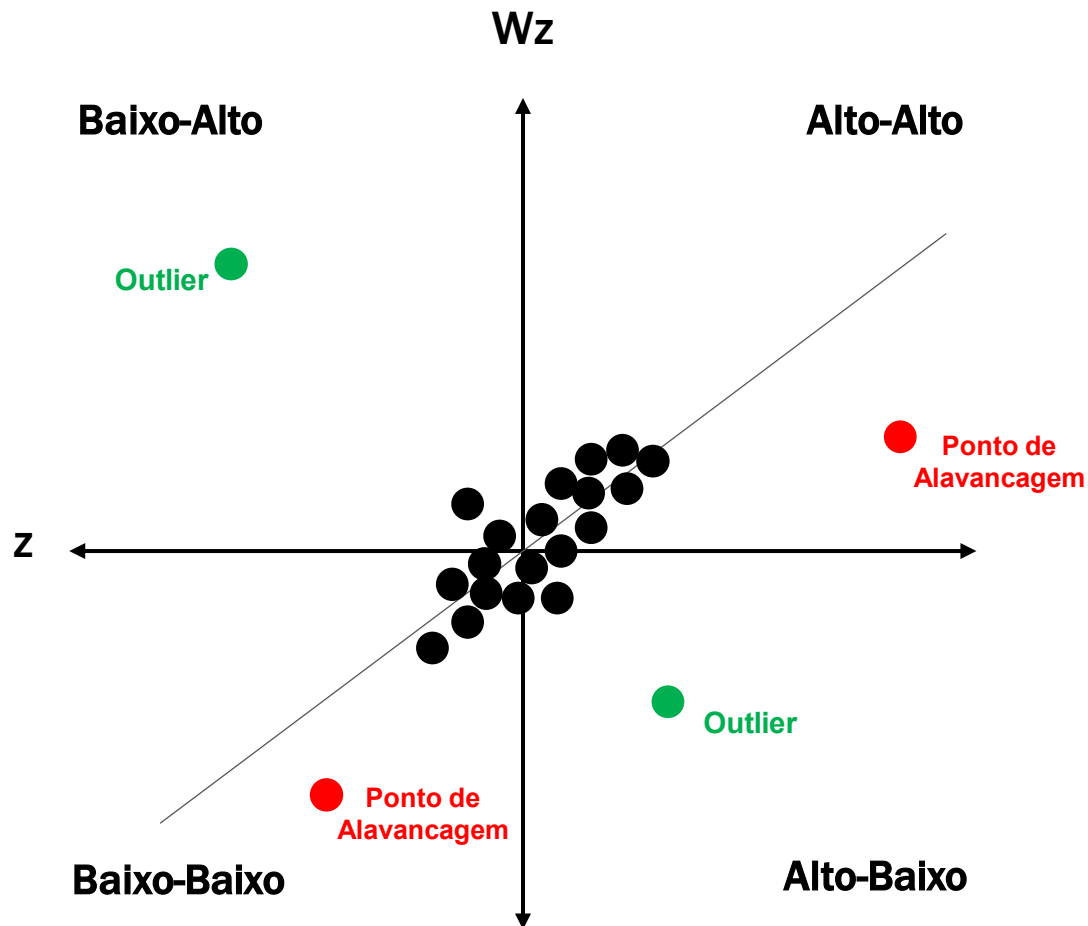
- **Pontos de alavancagem:**

Observações que, embora seguindo a mesma associação espacial do restante dos dados, exercem uma influência exagerada no grau de associação espacial.

- **Outliers espaciais:**

Observações que não seguem o mesmo processo de dependência espacial que a maioria dos dados e exercem elevada influência no grau de associação espacial

Outliers espaciais



Básica:

- ALMEIDA, E. S. **Econometria Espacial Aplicada**. Campinas: Editora Alínea, 2012.
 - Capítulo 1: O que é Econometria Espacial?
 - Capítulo 2: Dados Espaciais
 - Capítulo 3: Matrizes de Ponderação Espacial
 - Capítulo 4: Análise Exploratória de Dados Espaciais

Complementar:

- HADDAD, E. A.; HADDAD, P. R.; VALE, A. V. **Desenvolvimento territorial no Maranhão**: um diagnóstico espacial. Banco Mundial, 2017, mimeo.

- Professores:

Prof. Alexandre Alves Porsse:

porsse@gmail.com

Prof. Vinícius de Almeida Vale:

vinicius.a.vale@gmail.com



NEDUR

Núcleo de Estudos em Desenvolvimento
Urbano e Regional

Universidade Federal do Paraná



Av. Prefeito Lothário Meissner, nº 632 – Setor de Ciências Sociais | UFPR



www.nedur.ufpr.br



nedur.ufpr@gmail.com