6 - Modelando dados de área.

Prof. Vinícius D. Mayrink
EST171 - Estatística Espacial

Sala: 4073 Email: vdm@est.ufmg.br

1^o semestre de 2024

Muitas vezes as observações de covariáveis relevantes não estão disponíveis e a detecção de autocorrelação espacial nos dados ou nos resíduos do modelo constitui de fato a única forma de modelar a variação restante.

Mostramos aqui como a estrutura espacial de dependência entre observações pode ser modelada para dados de área. As observações podem não ser independentes (pode haver correlação entre áreas vizinhas).

Do ponto de vista estatístico, é possível levar em conta observações correlacionadas ao assumir a seguinte estrutura de modelagem:

$$\boldsymbol{Y} = \boldsymbol{\mu} + \boldsymbol{\epsilon}$$

sendo $\boldsymbol{\mu}$ = vetor de médias e
$\boldsymbol{\epsilon}$ = vetor de erros aleatórios o qual assumimos com distribuição $N(\mathbf{0}, \boldsymbol{V})$.

Podemos supor que a média depende de covariáveis substituindo μ por $\boldsymbol{X}'\beta$.

A correlação entre áreas é introduzida através de $\boldsymbol{V}$.

Se não houver Normalidade, podemos tentar transformar $\boldsymbol{Y}$ para obtê-la.

Muitas estruturas de correlação podem ser utilizadas em $\boldsymbol{V}$; entretanto iremos focar em duas abordagens que são mais usadas na prática SAR (*Simultaneous Autoregressive*) e CAR (*Conditional Autoregressive*).

Exemplo (dados sobre a incidência da leucemia): Os dados são referentes a um censo realizado em 8 condados do estado de Nova York. O censo foi realizado com a coleta de dados em 281 regiões dentro destes 8 condados, incluindo: regiões rurais e regiões com diferentes níveis de urbanização. O número de casos de leucemia são registrados em cada região e agregados de acordos com grupos (blocos) de regiões definidos no censo. Visto que alguns casos não podiam ser alocados, eles foram adicionados proporcionalmente a outros blocos (levando assim a contagens não inteiras). As contagens são referentes a um períodos de 5 anos (1978-1982), enquanto que as outras variáveis do censo (medidas por região) são obtidas apenas em 1980.

- tamanho da população;
- % da população acima de 65 anos (`PCTAGE65P`);
- % da população com casa própria (`PCTOWNHOME`);
- Exposição a locais de coleta de resíduos (TCE) = log de 100 vezes a distância inversa do centroide da região ao local TCE mais próximo (`PEXPOSURE`);

Não temos normalidade para estas contagens de casos de leucemia. Uma transformação sugerida na literatura será:

$$Z_i = \log \frac{1000(Y_i + 1)}{n_i}$$

sendo Y_i a contagem de casos na região i e
n_i o tamanho da população da região i.

A transformação acima fornece três *outliers* (regiões com populações pequenas, mas inesperadamente com grandes contagens). Este *outliers* não serão removidos.

Como covariáveis iremos utilizar: `PCTAGE65P`, `PCTOWNHOME` e `PEXPOSURE`

Iniciaremos com um modelo linear expressando o relacionamento de Z_i com as covariáveis.

Carregando os dados, alguns gráficos e ajuste do modelo linear $\boldsymbol{Z} = \boldsymbol{X}'\boldsymbol{\beta} + \boldsymbol{\epsilon}$ com estimação via OLS (mínimos quadrados usual). Os dados estão disponíveis em "NYData.zip" (sistema Moodle). Salve os dados no diretório de trabalho do R.

```
library(spdep)
library(rgdal)

NY8 <- readOGR(".","NY8_utm18")
NYnb <- read.gal("NY_Nb.gal", region.id = row.names(NY8))

summary(NY_nb)

plot(NY8, border = "grey60")

plot(NY_nb, coordinates(NY8), pch = 19, cex = 0.6, add = TRUE)

nylm <- lm(Z ~ PEXPOSURE + PCTAGE65P + PCTOWNHOME, data = NY8)
summary(nylm)
NY8$lmresid <- residuals(nylm)
```

As variáveis "Idade" e "Casa própria" parecem contribuir para explicar a variância da variável resposta. A variável "TCE" não é significativa (p-valor do teste t = 0.1648).

Distribuição espacial dos valores dos resíduos do modelo linear.

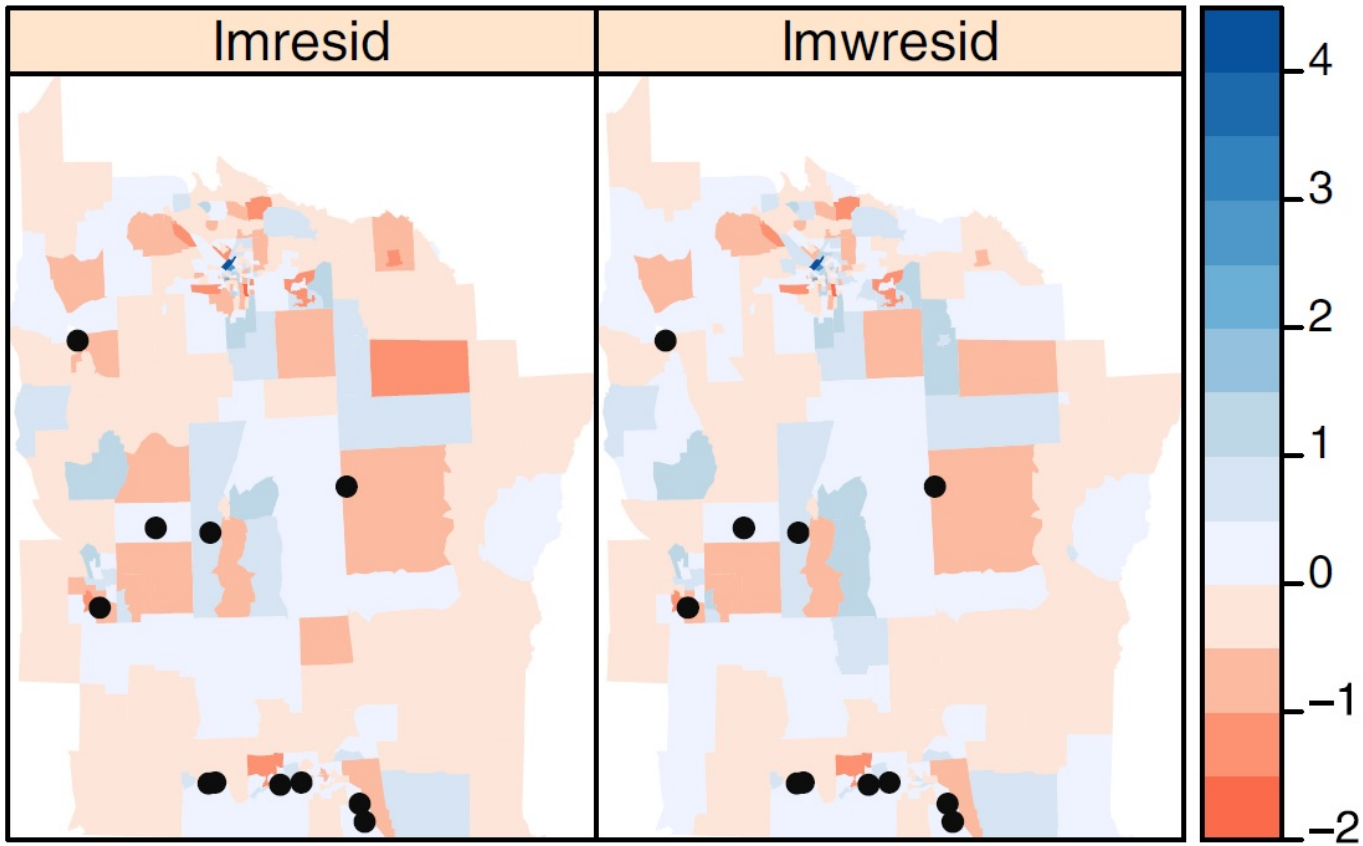

Fig.1 : Resíduos do modelo linear (esquerda). Resíduos de um modelo linear ponderado (direita). Locais TCE representados por pontos.

No modelo linear sem covariáveis ($Z_i = \beta_0 + \epsilon_i$) e com covariáveis ($Z_i = \boldsymbol{X}_i'\beta + \epsilon_i$) podemos calcular as estatísticas *I de Moran* e *C de Geary* para avaliar a autocorrelação espacial.

Embora tenhamos menos autocorrelação espacial nos resíduos do modelo "com covariáveis" do que no "sem covariáveis", existe ainda informação nos resíduos que deveríamos usar.

Visto que o teste de Moran tem o propósito de detectar autocorrelação espacial, podemos tentar ajustar um modelo que leve isso em consideração.

Sob a suposição "não há associação espacial" temos $E(I) = -1/(n-1)$ (cujo valor é próximo de zero para n grande).

Teste unilateral: $H_0 : I = E(I)$ vs. $H_1 : I > E(I)$, A estatística de teste, com distribuição $N(0,1)$, será: $Z_0 = [I_{obs} - E(I)]/\sqrt{\text{Var}(I)}$.
Rejeitamos H_0 se $Z_0 > z_\alpha$.

Teste bilateral: $H_0 : I = E(I)$ vs. $H_1 : I \neq E(I)$,

```
library(spdep)
NYlistw <- nb2listw(NY_nb, style = "B")
lm.morantest(nylm, NYlistw)
```

Especifica-se `style = "B"` para construir W usando indicador binário de vizinhança.

Testes globais: avaliam a associação espacial de toda a região de interesse.

Os testes baseados nas estatísticas "I de Moran" e "C de Geary" são ditos globais.

Os comandos acima aplicam o teste global de Moran diretamente para os resíduos da regressão (neste caso, `nylm`). Para um conjunto de dados qualquer utilize `moran.test()`.

Modelo SAR (*Simultaneous Autoregressive*)

Este modelo utiliza uma regressão baseada nos valores das demais áreas para introduzir a dependência espacial.

Os resíduos e_i são modelados de forma a dependerem uns dos outros da seguinte forma:

$$e_i = \sum_{i=1}^{m} b_{ij}\, e_i \; + \; \epsilon_i$$

Aqui, ϵ_i's representam erros associados aos resíduos, os quais são assumidos independentes $\epsilon \sim N_m(\mathbf{0}, \Sigma_\epsilon)$ com matriz de covariâncias contendo na diagonal principal $\sigma^2_{\epsilon_i}$, $i = 1, \ldots, m$ (a mesma variância σ^2_ϵ, $\forall i$, pode ser considerada também).

Os valores b_{ij} representam a dependência espacial entre as áreas.
Assuma $b_{ii} = 0$ para não realizar regressão de uma área sobre ela mesmo.

Se expressarmos os termos de erro como $e = B(Y - X'\beta) + \epsilon$, o modelo poderá ser escrito como:

$$Y = X'\beta + B(Y - X'\beta) + \epsilon.$$

Portanto, este modelo pode ser formulado em formato matricial como segue:

$$(I - B)(Y - X'\beta) = \epsilon,$$

sendo B uma matriz que contém os parâmetros de dependência espacial b_{ij} e I uma matriz identidade com mesma dimensão.

Para o modelo SAR estar bem definido, a matriz $I - B$ precisa ser não-singular.

No SAR, $Y \sim$ Normal Multivariada com:

$$E[Y] = X'\beta \quad Var[Y] = (I - B)^{-1}\Sigma_\epsilon(I - B')^{-1}.$$

Em geral assumimos: $\Sigma_\epsilon = \sigma^2 I$ e assim $Var[Y] = \sigma^2(I - B)^{-1}(I - B')^{-1}$.

Uma reparametrização útil deste modelo será: $B = \lambda W$, sendo $\lambda =$ parâmetro de autocorrelação espacial (ρ, rever slides sobre dados de área) e W uma matriz de dependência espacial (em geral simétrica). Com esta especificação, a variância de Y se torna:

$$Var[Y] = \sigma^2(I - \lambda W)^{-1}(I - \lambda W')^{-1}$$

Estes modelos podem ser estimados eficientemente via máxima verossimilhança. No `R`, isto pode ser feito usando a função `spautolm` do pacote `spdep`. O modelo pode ser especificado aqui através de uma fórmula para o preditor linear, enquanto W deve ser inserido como um objeto `listw`.

Os comandos abaixo mostram como ajustar o SAR via OLS. Podemos considerar também o modelo ponderado (WLS) com pesos baseados no tamanho populacional em 1980; este caso será considerado adiante.

```
nysar <- spautolm(Z ~ PEXPOSURE + PCTAGE65P + PCTOWNHOME,
+                   data = NY8, listw = NYlistw)
summary(nysar)
```

Os resultados indicam correlação espacial significativa. O teste da razão de verossimilhanças fornece: $H_0 : \lambda = 0$ (sem associação espacial) e $H_1 : \lambda \neq 0$. Estatística $\lambda = 0.0405$ e p-valor $= 0.022$.

Se $\lambda = 0$, temos $B = 0W$ e $U = 0U + \epsilon$ com $\epsilon \sim N_m(\mathbf{0}, \hat{\sigma}^2 I_m)$.

A proximidade ao TCE parece não ser significante;
veja que PEXPOSURE tem p-valor $= 0.0913$.

Covariáveis significativas: PCTAGE65P (p-valor $= 1.86 \times 10^{-9}$) e PCTOWNHOME (p-valor $= 0.0282$). Conclui-se que regiões com maior ($\beta_2 = 3.75$) porcentagem de idosos e pequena ($\beta_3 = -0.42$) porcentagem de casa própria possuem incidências (transformadas) mais altas.

Recordando a definição do modelo: $Y = X'\beta + \lambda W(Y - X'\beta) + \epsilon$.
O primeiro termo $X'\beta$ é um componente de tendência espacial.
O segundo termo $\lambda W(Y - X'\beta)$ é um componente estocástico espacial.

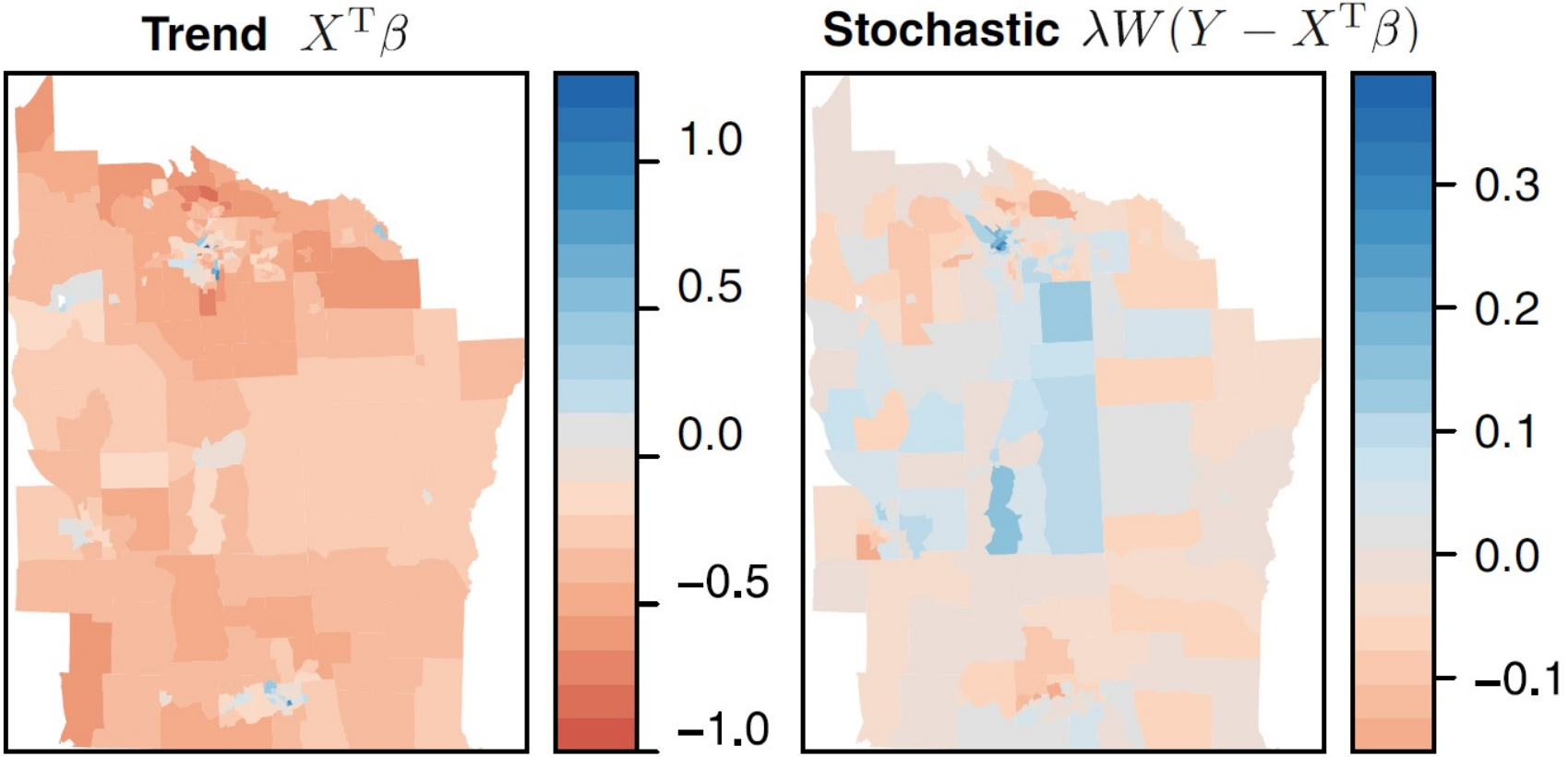

Fig.2 : Esquerda = componente de tendência dos valores ajustados via SAR. Direita = Componentes estocástico espacial dos valores ajustados via SAR.

Este modelo não leva em conta a heterogeneidade de distribuições populacionais entre as regiões. Uma versão do SAR pode ser ajustada de forma que as regiões sejam ponderadas proporcionalmente pelo inverso de seu tamanho populacional.

O ajuste via WLS do modelo sem tratamento espacial $Z = X\beta + \epsilon$ é obtido com os comandos abaixo:

```
nylmw <- lm(Z ~ PEXPOSURE + PCTAGE65P + PCTOWNHOME,
+                data = NY8, weights = POP8)
summary(nylm)
```

Estimador OLS: $\hat{\beta} = (X'X)^{-1}X'Y$;
Estimador WLS: $\hat{\beta} = (X'PX)^{-1}X'PY$, sendo P matriz $(n \times n)$ diagonal com os pesos.

A variável PEXPOSURE se tornou significativa (p-valor $= 0.0056$ e $\beta_1 = 0.0763$). Seu coeficiente positivo indica que quanto mais próxima de locais TCE, maior incidência transformada terá a região.

As outras duas covariáveis são mais significativa neste ajuste;
p-valor PCTAGE65P $= 8.6 \times 10^{-11}$ e p-valor PCTOWNHOME $= 0.0097$.

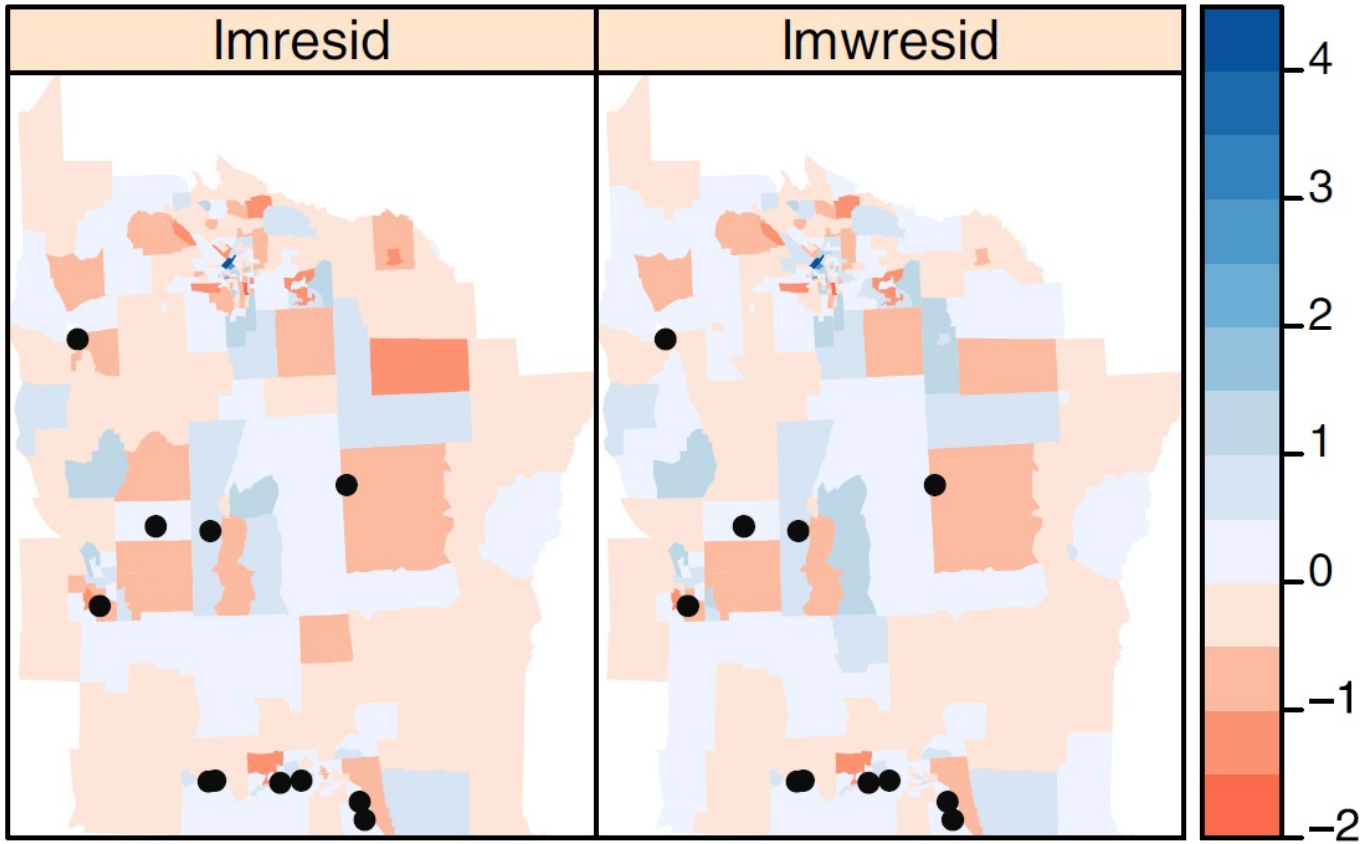

Fig.3 : Esquerda = resíduos do modelo linear para as proporções de incidências transformadas. Direita = Resíduos do modelo linear ponderado. Pontos = locais TCE.

O gráfico acima mostra que parte da informação foi removida dos resíduos e explicada pelo modelo. Pouca estrutura espacial visível resta nos mapas.

O teste de Moran para os resíduos da regressão também pode ser usado para um modelo linear ponderado:

```
> lm.morantest(nylmw, NYlistw)
```

Resultado: I de Moran observado $= 0.007533$.
H_0 : sem associação espacial.
p-valor $= 0.3166 > \alpha = 0.05$, então não rejeitamos H_0.

Este resultado sugere que o erro de especificação detectado pela estatística I de Moran está na verdade mais relacionado com a heterocedasticidade do que com a autocorrelação espacial.

Podemos verificar isso usando o modelo SAR (ajuste WLS); visto que a função `spautolm` aceita pesos como argumento.

```
> nysarw <- spautolm(Z ~ PEXPOSURE + PCTAGE65P + PCTOWNHOME, data = NY8,
+                    listw = NYlistw, weights = POP8)
> summary(nysarw)
```

Os coeficientes das covariáveis mudam pouco neste modelo SAR (WLS).
SAR (OLS): $\beta_0 = -0.6182$, $\beta_1 = 0.0710$, $\beta_2 = 3.7542$, $\beta_3 = -0.4199$;
SAR (WLS): $\beta_0 = -0.7971$, $\beta_1 = 0.0805$, $\beta_2 = 3.8167$, $\beta_3 = -0.3808$.

Os p-valores dos coeficientes diminuem substancialmente no SAR (WLS).
SAR (OLS): $4.7 \times 10^{-4}(\beta_0)$, $0.0912(\beta_1)$, $1.8 \times 10^{-9}(\beta_2)$, $0.0282(\beta_3)$;
SAR (WLS): $3.1 \times 10^{-8}(\beta_0)$, $0.0045(\beta_1)$, $3.4 \times 10^{-11}(\beta_2)$, $0.0149(\beta_3)$.

No SAR (WLS) a covariável PEXPOSURE (proximidade ao TCE) se tornou significativa.

Teste da razão de verossimilhanças:
$H_0 : \lambda = 0$ (sem associação espacial) e $H_1 : \lambda \neq 0$.
Estatística $\lambda = 0.009564$ e p-valor $= 0.56764$.
Conclusão: não rejeitamos H_0 e dizemos que não há indicação de autocorrelação espacial.

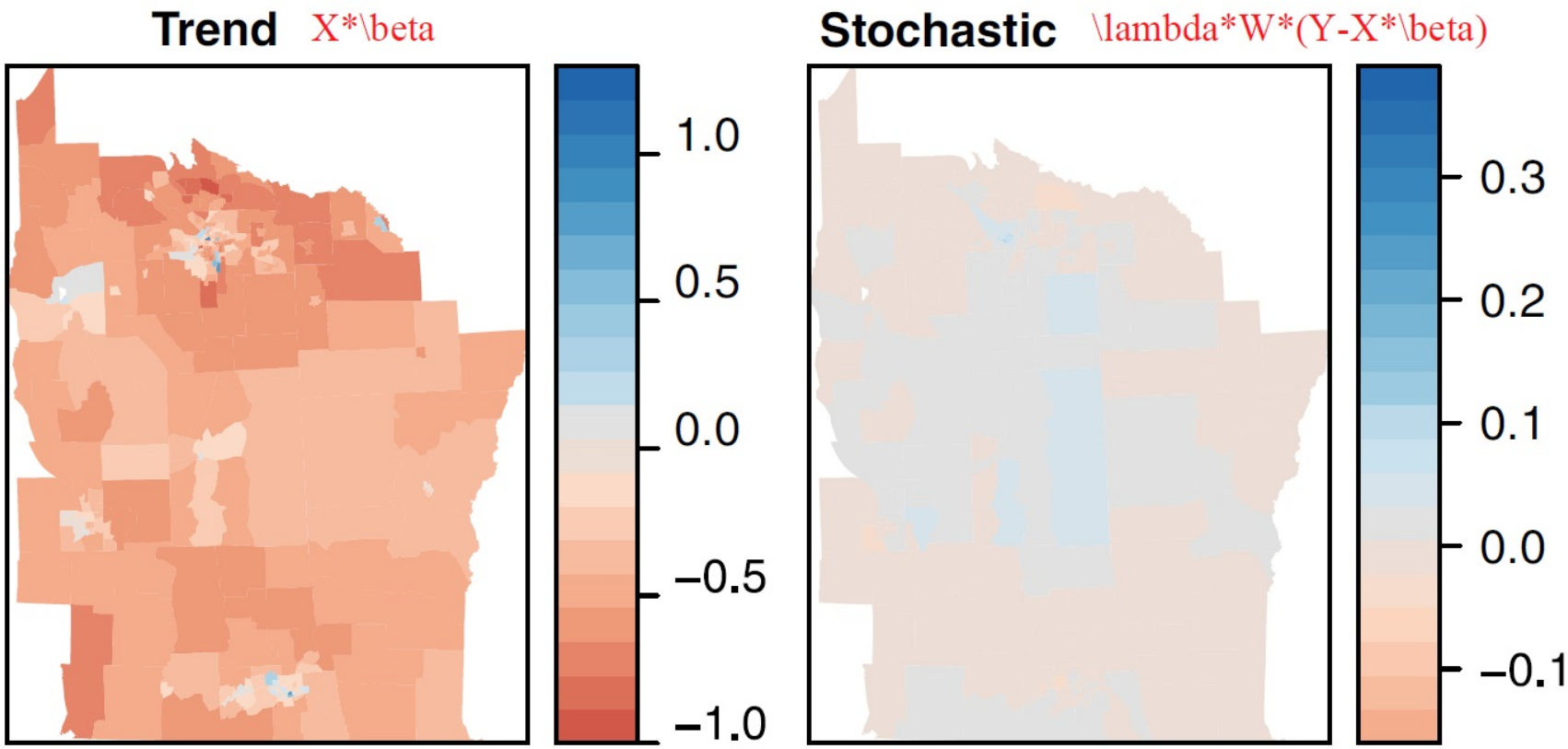

Fig.4 : Esquerda = componente de tendência dos valores ajustados via modelo SAR (WLS). Direita = componente estocástico espacial dos valores ajustados via SAR (WLS).

Esta figura sugere que os diferentes pesos (1/tamanho populacional) das regiões são responsáveis pela correlação espacial observada nos resíduos após ajustarmos com covariáveis.

Para comparar ambos os modelos, podemos usar o critério AIC (*Akaike's Information Criterion*) que é reportado no sumário do ajuste.

O AIC é uma soma ponderada envolvendo a log-verossimilhança do modelo e o número de coeficientes ajustados. De acordo com este critério, o melhor modelo possuirá o menor AIC.

A versão ponderada do SAR fornece um melhor ajuste ($AIC_{WLS} = 515.2 < 564.2 = AIC_{OLS}$).

Este resultado mostra a importância de levarmos em conta na análise as diferenças entre as populações das regiões.

Modelo CAR (*Conditional Autoregressive*)

Este modelo é baseado na distribuição condicional dos termos de erro. Neste caso, a distribuição condicional de $(e_i|e_{-i})$; sendo e_{-i} um vetor contendo todos os erros, exceto e_i.

Ao invés de todos os erros em e_{-i}, apenas aqueles referentes às regiões vizinhas de i serão usados. Este sub-vetor será representado por $e_{j\sim i}$.

Uma forma simples de expressar a distribuição condicional mencionada será:

$$(e_i|e_{j\sim i}) \sim N\left[\sum_{j\sim i}\frac{c_{ij}e_j}{\sum_{j\sim i}c_{ij}}, \frac{\sigma^2_{e_i}}{\sum_{j\sim i}c_{ij}}\right]$$

sendo c_{ij} parâmetros de dependência similares a b_{ij}. Veja que no IAR, teremos $c_{ij} = 1$ se i vizinho de j ($c_{ij} = 0$ c.c.); assim $\sum_{j\sim i} c_{ij} =$ número de vizinhos de i.

Entretanto, especificar as distribuições condicionais dos termos de erro não implica na existência da distribuição conjunta. Para termos uma distribuição própria, algumas restrições devem ser atendidas (incluir o parâmetro ρ).

Para ajustar o modelo CAR, podemos também usar a função `spautolm`. Neste caso, especificaremos o argumento `family = "CAR"`.

```
> nycar <- spautolm(Z ~ PEXPOSURE + PCTAGE65P + PCTOWNHOME, data = NY8,
+                       family = "CAR", listw = NYlistw)
> summary(nycar)
```

As estimativas dos coeficientes são bem similares às do modelo SAR:
SAR (OLS): $\beta_0 = -0.6182$, $\beta_1 = 0.0710$, $\beta_2 = 3.7542$, $\beta_3 = -0.4199$;
CAR (OLS): $\beta_0 = -0.6484$, $\beta_1 = 0.0779$, $\beta_2 = 3.7038$, $\beta_3 = -0.3829$;

As covariáveis PEXPOSURE e PCTOWNHOME não são significativas (p-valores um pouco acima de 0.05).
CAR (OLS): $0.00034(\beta_0)$, $0.0746(\beta_1)$, $3.5 \times 10^{-9}(\beta_2)$, $0.0503(\beta_3)$;

Teste da razão de verossimilhanças:
$H_0 : \lambda = 0$ (sem associação espacial) e $H_1 : \lambda \neq 0$.
Estatística $\lambda = 0.0841$ e p-valor $= 0.016$.
Conclusão: rejeitamos H_0; existe indicação de autocorrelação espacial.

A seguir, o modelo CAR será ajustado na versão ponderada (WLS com pesos = inversa do tamanho populacional).

```
> nycarw <- spautolm(Z ~ PEXPOSURE + PCTAGE65P + PCTOWNHOME, data = NY8,
+                    family = "CAR", listw = NYlistw, weights = POP8)
> summary(nycarw)
```

As estimativas dos coeficientes não mudam muito em relação ao SAR:
SAR (WLS): $\beta_0 = -0.7971$, $\beta_1 = 0.0805$, $\beta_2 = 3.8167$, $\beta_3 = -0.3808$.
CAR (WLS): $\beta_0 = -0.7902$, $\beta_1 = 0.0819$, $\beta_2 = 3.8259$, $\beta_3 = -0.3868$;

Todas as covariáveis são significativas. Seus p-valores:
CAR (WLS): $4.9 \times 10^{-8}(\beta_0)$, $0.0042(\beta_1)$, $3.5 \times 10^{-11}(\beta_2)$, $0.014(\beta_3)$;

O uso de pesos remove completamente a autocorrelação espacial nos dados.
Teste da razão de verossimilhanças:
$H_0 : \lambda = 0$ (sem associação espacial) e $H_1 : \lambda \neq 0$.
Estatística $\lambda = 0.02242$ e p-valor $= 0.5334$.
Conclusão: não rejeitamos H_0; não há indicação de autocorrelação espacial.

Modelar estes dados através do SAR ou CAR leva a resultados similares.

Modelo de regressão espacial

A função `spautolm` também ajusta modelos de regressão via máxima verossimilhança. Ela segue os passos:

1. Encontrar o valor do coeficiente autoregressivo espacial que maximiza a log-verossimilhança do modelo.
2. Ajustar os outros coeficientes (β's) via mínimos quadrados generalizados naquele ponto.

A parte mais trabalhosa das contas está no cálculo do Jacobiano $\log(|I - B|)$ dentro da otimização; considere $|I - B|$ = determinante da matriz $(n \times n)$ $I - B$ (ou $I - \lambda W$).

$$\log(|I - \lambda W|) = \log\left(\prod_{i=1}^{n}(1 - \lambda\zeta_i)\right)$$

sendo ζ_i's os autovalores de W. Encontrar os autovalores de W fica mais difícil quando n é grande.

O `spautolm` considera o método baseado nos autovalores (`method = "eigen"`) como default. Podemos também indicar os limites inferior e superior $[1/\min_i(\zeta_i), 1/\max_i(\zeta_i)]$ para a busca de λ. Aqui, λ equivale ao parâmetro ρ do CAR próprio. Se calcularmos os autovalores de W, teremos os limites de ρ (que garantem a inversão de $I - \lambda W$). Entretanto, encontrar os autovalores será problemático quando n for grande.

Abordagens alternativas envolvem a busca pelo log-determinante da decomposição de Cholesky da matriz esparsa $(I - \lambda W)$. Não é possível pré-calcular os autovalores, então o log-determinante será computado para cada valor de λ usado. O número de $\lambda's$ necessários não é grande em geral; desta forma, problemas com n grande se tornam possíveis de trabalhar.

Diversas abordagens baseadas na matriz esparsa foram avaliadas; o método "Matrix" (argumento `method = "Matrix"`) será aquele que utilizaremos.

Todas as abordagens baseadas na decomposição de Cholesky, para calcular o Jacobiano, requerem que W seja simétrica ou pelo menos similar a uma matriz simétrica. Matrizes que são similares a simétricas possuem os mesmos autovalores; ou seja, os autovalores da matriz simétrica $W^* = D^{1/2}\tilde{W}D^{1/2}$ e da matriz padronizada nas linhas $\tilde{W} = DB$ são os mesmos, se B for matriz de pesos (binária ou geral).

O ajuste do modelo SAR (WLS) com estimação de λ via método "Matrix" é obtido com os comandos a seguir:

```
> nysarwM <- spautolm(Z ~ PEXPOSURE + PCTAGE65P + PCTOWNHOME, data = NY8,
+                 family = "SAR", listw = NYlistw, weights = POP8, method = "Matrix")
> summary(nysarwM)
```

O resultado é idêntico àquele obtido com o SAR (WLS) via autovalores de W.

Coeficientes: $\beta_0 = -0.7971$, $\beta_1 = 0.0805$, $\beta_2 = 3.8167$, $\beta_3 = -0.3808$.

p-valores: $3.1 \times 10^{-8}(\beta_0)$, $0.0045(\beta_1)$, $3.4 \times 10^{-11}(\beta_2)$, $0.0149(\beta_3)$.

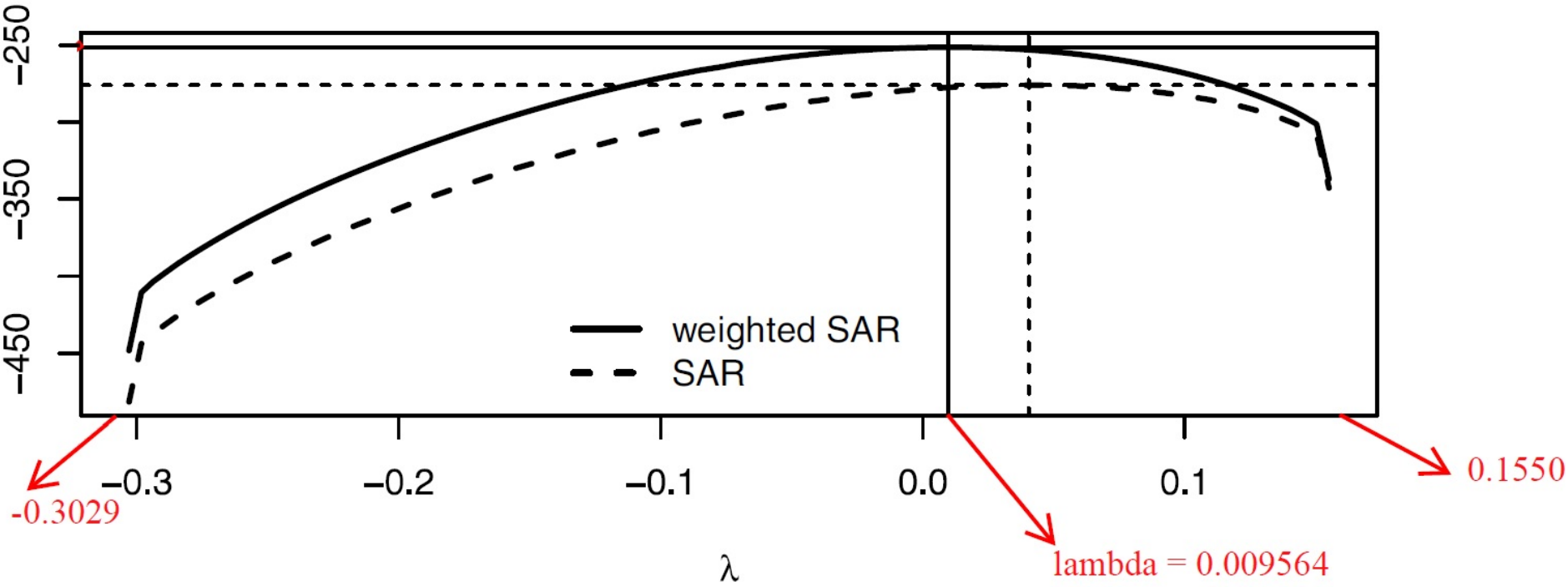

Fig.5 : Valores da log-verossimilhança para um intervalo de valores de λ (modelo SAR ponderado e não ponderado). Valores ajustados do coeficiente espacial e o máximo da log-verossimilhança são indicados no gráfico.

Para obter os limites de λ:

```
> 1/range(eigenw(NYlistw))
```

Se for de interesse examinar os valores da função log-verossimilhança para uma sequência de valores λ, o argumento `llprof` pode ser usado para fornecer o número de valores λ igualmente espaçados a serem escolhidos entre o inverso do menor e do maior autovalor (usando `method = "eigen"`).

```
> nysar_ll <- spautolm(Z ~ PEXPOSURE + PCTAGE65P + PCTOWNHOME, data = NY8,
+                      family = "SAR", listw = NYlistw, llprof = 100)

> nysarw_ll <- spautolm(Z ~ PEXPOSURE + PCTAGE65P + PCTOWNHOME, data = NY8,
+                      family = "SAR", listw = NYlistw, weights = POP8,
+                      llprof = 100)
```