Análise Exploratória, Modelação e Predição de Dados Geoestatísticos e de Dados Agregados por Área

Estatística Espacial

Rui Alves PG55577

7 de janeiro de 2026

Índice

Dados Geoestatísticos

wrc e soil250

Dados geoestatísticos — *wrc* e *soil250*

wrc (geoR) — Base de dados de retenção de água composta por 250 observações obtidas numa grelha regular de 10 por 25 pontos, espaçados por 5 metros, a uma profundidade de 25 cm no solo.
soil250 (geoR) — Propriedades químicas do solo nos mesmos pontos.

CoordX	Coordenada espacial no eixo X (m)
CoordY	Coordenada espacial no eixo Y (m)
Densidade	Densidade do solo (g/cm^3)
Pr5	Retenção de água no solo a 5mca
Pr10	Retenção de água no solo a 10mca
Pr100	Retenção de água no solo a 100mca
Pr15300	Retenção de água no solo a 15300mca
Areia	Percentagem de areia no solo (%)
Silte	Percentagem de silte no solo (%)
Argila	Percentagem de argila no solo (%)

Análise Exploratória Não Espacial

Variável Resposta - Quantidade de água retida

Pressão (mca)	Mínimo	1.º Quartil	Mediana	Média	3.º Quartil	Máximo	Desvio-padrão
5	0.2144	0.2971	0.3137	0.3154	0.3334	0.5384	0.0313
10	0.2295	0.2964	0.3130	0.3137	0.3324	0.5396	0.0307
100	0.2135	0.2539	0.2684	0.2674	0.2806	0.3150	0.0200
15300	0.1545	0.1842	0.1990	0.1987	0.2145	0.2398	0.0198

Tabela 1: Estatísticas descritivas da quantidade de água retida.

Análise Exploratória Não Espacial

Variável Resposta – Quantidade de água retida

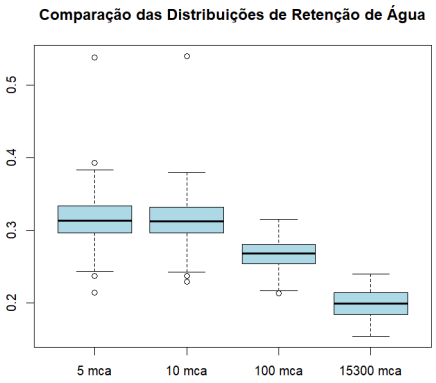


Fig.1 - Boxplot das quantidades de retenção de água.

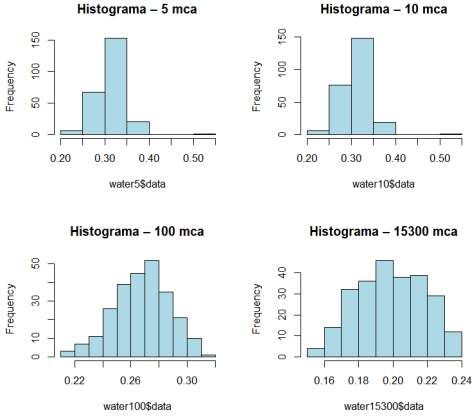


Fig.2 - Histograma das quantidades de retenção de água.

Variável	Mínimo	1.º Quartil	Mediana	Média	3.º Quartil	Máximo	Desvio-padrão
Areia (%)	6	8	9	8.78	9	11	0.894
Silte (%)	22	25	26	26.13	27	30	1.646
Argila (%)	35	40	43	42.65	45	51	3.206
Densidade (g/cm^3)	1.272	1.449	1.490	1.490	1.529	1.727	0.071

Tabela 2: Estatísticas descritivas das variáveis granulométricas do solo e da densidade.

Análise Exploratória Não Espacial

Variáveis Explicativas - Densidade e composição granulométrica do Solo

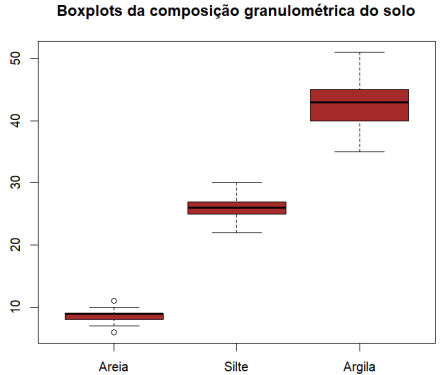


Fig.3 - Boxplots da composição granulométrica do solo.

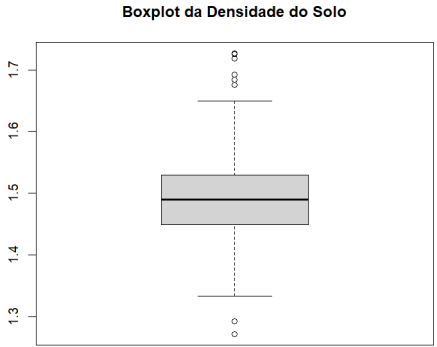
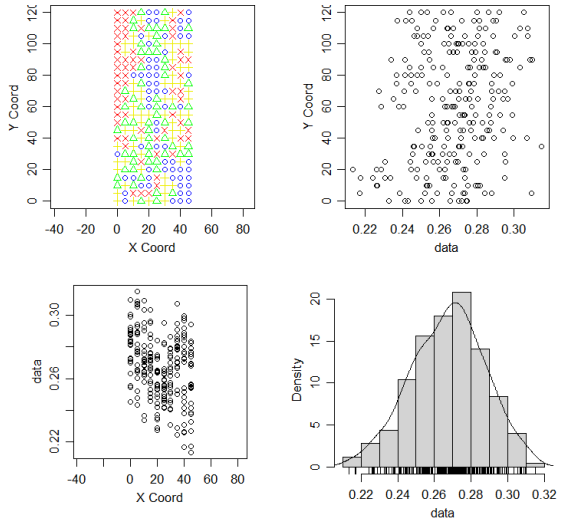


Fig.4 - Boxplot da densidade do solo.

Análise Exploratória Espacial



- Tendência negativa no aumento em X;
- Tendência positiva no aumento em Y;
- Aparente normalidade dos dados.

Fig. 5 — Distribuição espacial da retenção de água no solo a 100mca, incluindo gráficos de dispersão em função das coordenadas espaciais e histograma da variável resposta.

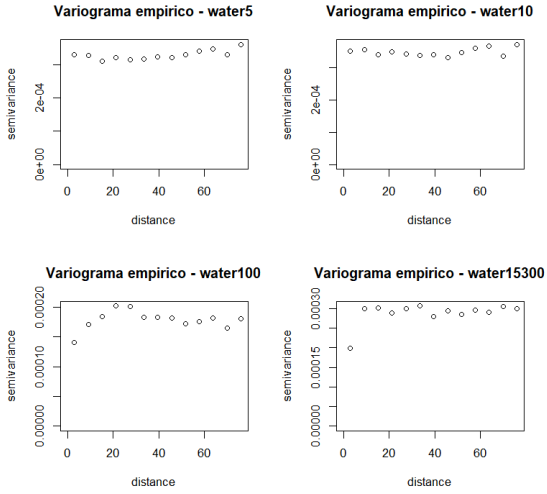
Análise de Tendências

- Foram ajustados **modelos de regressão linear** para cada variável resposta, com o objetivo de identificar e caracterizar a presença de **tendências espaciais** associadas às coordenadas e às propriedades do solo.

Variável	Estimativa	p-valor	Significância
CoordX	−0.00031	0.00001	***
CoordY	0.00006	0.03810	*
Densidade	−0.18560	0.00001	***
Areia	−0.00073	0.67960	
Silte	0.00132	0.22500	
Argila	0.00108	0.15999	

Tabela 3: Resultados da regressão linear para a retenção de água no solo a 100mca.

Variogramas Empíricos



- Para 5, 10 e 15300mca observa-se uma variação reduzida da variância;
- Sugere fraca dependência espacial residual;
- Para 100mca a estrutura espacial é mais pronunciada.

Fig. 6 — Variogramas empíricos após remoção das tendências para a retenção de água a 5, 10, 100 e 15300mca.

Variogramas Teóricos

- Foram ajustados modelos teóricos exponenciais e esféricos aos variogramas empíricos através de métodos de máxima verosimilhança e dos mínimos quadrados;
- A escolha do modelo foi realizada com base em loocv, comparando a média dos erros de predição e a médio do quadrado dos erros padrões.
- O modelo exponencial estimado por máxima verosimilhança apresentou melhor desempenho.

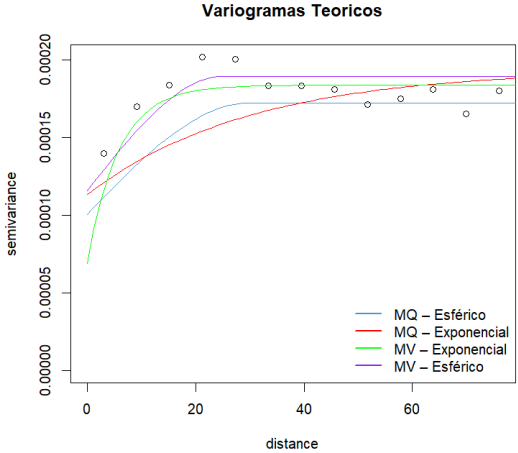


Fig. 7 — Variogramas teóricos ajustados aos variogramas empíricos, após remoção das tendências, para a retenção de água no solo a 100mca.

Modelo Matemático do Variograma Teórico

O modelo considerado para a variável resposta é dado por:

$$Y(x) = \mu(x) + S(x) + e(x),$$

onde $\mu(x)$ representa a tendência média, $S(x)$ o efeito espacial estruturado e $e(x)$ o erro aleatório não estruturado.

A tendência é modelada através de um preditor linear:

$$\mu(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3,$$

resultando no modelo completo:

$$Y(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + S(x) + e(x).$$

Substituindo pelos coeficientes estimados, obtém-se:

$$\hat{Y}(x) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_3 x_3 + S(x) + e(x),$$

$$\hat{Y}(x) = 0.5275 - 0.0004\, x_1 - 0.1695\, x_3 + S(x) + e(x)$$

onde x_1 e x_3 correspondem, respetivamente, às coordenadas espaciais e à densidade do solo.

Modelo Matemático do Variograma Teórico

Assume-se ainda que o erro aleatório e o efeito espacial obedecem às seguintes distribuições:

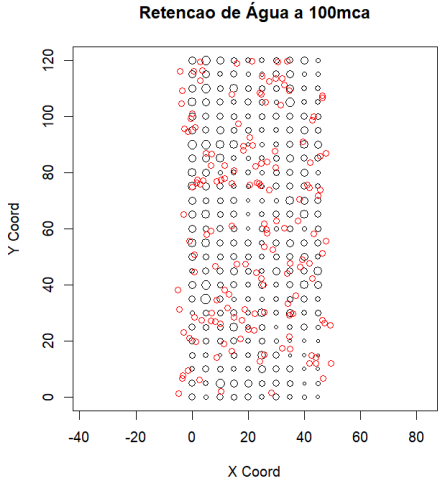
$$e(x) \sim \mathcal{N}(0, \hat{\tau}^2), \qquad \hat{\tau}^2 = 0.0001,$$

$$S(x) \sim \text{SGP}\big(0, \hat{\sigma}^2, p(\cdot \mid \hat{\phi})\big),$$

$$\hat{\sigma}^2 = 0.0001, \qquad \hat{\phi} = 5.8067,$$

onde $p(\cdot \mid \hat{\phi})$ corresponde a uma função de correlação exponencial.

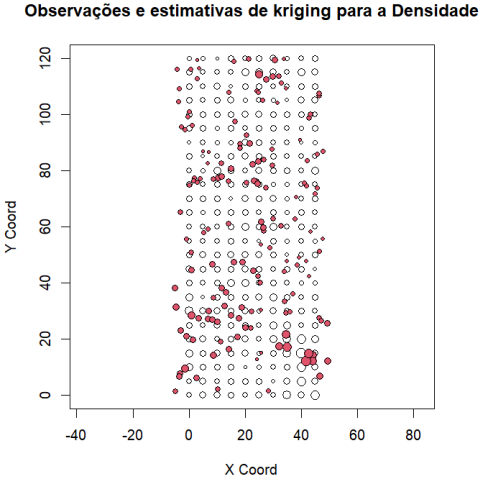
Predição Espacial - Novas Localizações



- Realizada em 150 pontos aleatórios dentro dos limites -5 e 50m em X e 0 a 120m em Y;

Fig. 8 — Distribuição espacial da retenção de água no solo a 100mca e das novas localizações onde irá ser realizada a predição.

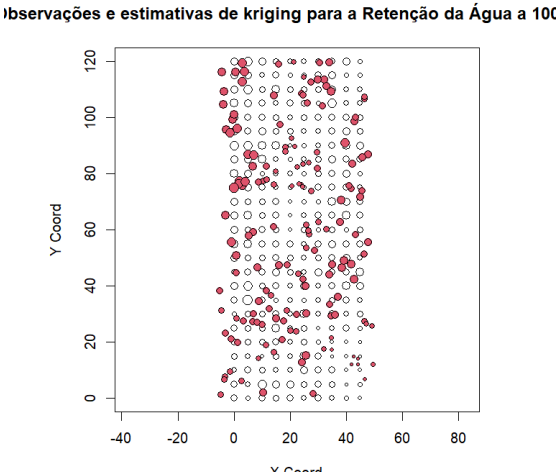
Predição Espacial - Densidade



- Para realizar kriging com tendência externa é necessário prever os valores de densidade nas novas localizações;
- Foi realizado o processo análogo: Análise Tendência (dependência Y) - Variograma empírico - Variograma Teórico (loocv) - Predição através kriging universal.

Fig. 9 — Distribuição espacial da densidade no solo e estimativas de kriging para as novas localizações.

Predição Espacial - Retenção de Água a 100mca



- Foi realizado kriging com tendência externa (densidade) para previsão da quantidade de água retida a 100mca;
- Pontos ligeiramente fora da grelha inicial tiveram variância de kriging superior.

Fig. 10 — Estimativas de kriging para a retenção de água a 100mca.

Conclusão — Dados Geoestatísticos

- A análise geoestatística permitiu descrever a variabilidade espacial da retenção de água no solo, destacando a importância da densidade do solo.
- Foram identificadas tendências espaciais, que justificaram a inclusão de uma componente determinística na modelação.
- Após remoção da tendência, a dependência espacial foi fraca para a maioria das pressões, sendo mais evidente para a retenção de água a 100 mca.
- O modelo exponencial por máxima verosimilhança mostrou bom desempenho, permitindo predições por krigagem coerentes.

Dados Agregados por Área

scotland_sf

Dados Agregados por Área — *scotland_sf*

scotland_sf — Dados agregados por condados da Escócia sobre a incidência de cancro do lábio em homens, no período 1975–1980.

geometry	Geometria dos condados (*multipolygon*)
county	Identificador do condado
cases	Número observado de casos
expected	Número esperado de casos
AFF	Percentagem da população empregada nos setores agrícola, florestal e da pesca

Análise Exploratória Não Espacial

Variável	Mínimo	1.º Quartil	Mediana	Média	3.º Quartil	Máximo	Desvio-padrão
Casos observados	0.00	4.75	8.00	9.57	11.00	39.00	7.91
Casos esperados	1.10	4.05	6.30	9.58	10.13	88.70	13.18
AFF	0.000	0.010	0.070	0.087	0.115	0.240	0.068

Tabela 4: Estatísticas descritivas das variáveis da base de dados scotland_sf para os 56 condados.

Análise Exploratória Não Espacial

Casos Observados e SMR (rácio padronizado de mortalidade)

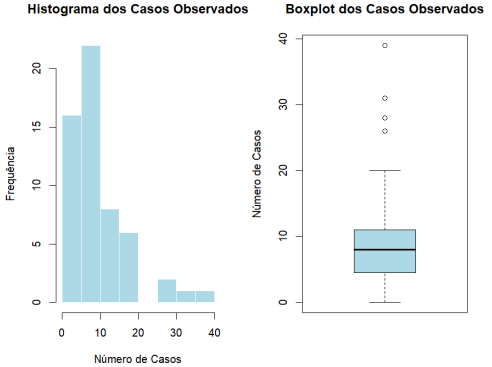


Fig.11 - Histograma e Boxplot do nº de casos observados de cancro no lábio.

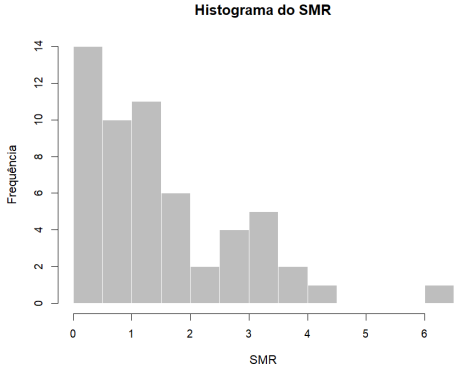


Fig.12 - Histograma do SMR (quociente entre nºcasos observados e nº casos esperados).

Análise Exploratória Espacial

Casos Observados e SMR (rácio padronizado de mortalidade)

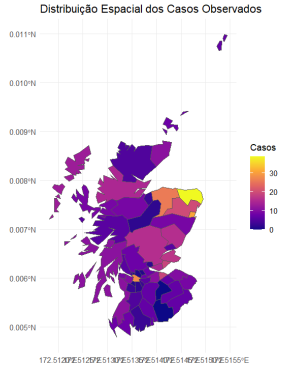


Fig.13 - Distribuição espacial dos casos de cancro no lábio nos 56 condados.

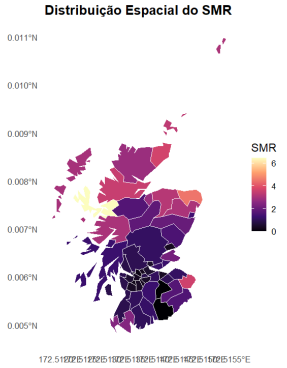


Fig.14 - Distribuição espacial do SMR nos 56 condados.

Coeficiente de Moran I

Em dados agregados por área, é essencial avaliar a existência de dependência espacial entre regiões vizinhas. Para esse efeito, utiliza-se o coeficiente de Moran I, que quantifica o grau de autocorrelação espacial com base numa matriz de vizinhança (definida por *queen contiguity*).

Hipóteses:

- H_0 : Ausência de autocorrelação espacial entre áreas vizinhas
- H_1 : Existência de autocorrelação espacial

Teste: *Moran global*

Estatística	Valor
Moran's I	0.2456
Valor esperado	-0.0192
Variância	0.0083
Z-value	2.9030
p-value	0.00185

Tabela 5: Resultados do teste de Moran global.

Enquadramento da Modelação

- A variável resposta corresponde ao **número de casos observados de cancro do lábio** em cada condado, tratando-se de dados de contagem.
- Consequentemente, em todos os modelos foi assumida a **distribuição de Poisson** para a variável resposta.
- Foi utilizada a **função de ligação logarítmica**, apropriada para modelos de contagem.
- Em todos os modelos foi incluído um **termo de offset** correspondente ao logaritmo do número esperado de casos, permitindo modelar o risco relativo.
- A covariável socioeconómica **AFF** foi incluída como **efeito fixo em todos os modelos**, de forma a avaliar a sua associação com a incidência da doença.

Modelo de Poisson simples (GLM) (1)

O modelo de referência considerado foi um modelo de Poisson simples (GLM):

$$Y_i \mid \mu_i \sim \text{Poisson}(\mu_i),$$

$$\log(\mu_i) = \log(e_i) + \beta_0 + \beta_1 \, \text{AFF}_i,$$

onde e_i representa o número esperado de casos (offset).

Parâmetro	Estimativa	Std. Error	z-value	p-value
Intercept	-0.5423	0.0695	-7.80	< 0.001
AFF	7.3732	0.5956	12.38	< 0.001

Tabela 6: Estimativas dos coeficientes do modelo Poisson não espacial com offset.

Modelo Poisson Espacial (GLMM) — Efeito CAR (2)

O modelo Poisson espacial (GLMM) com efeito estruturado espacialmente é dado por:

$$Y_i \mid \mu_i \sim \text{Poisson}(\mu_i),$$

$$\log(\mu_i) = \log(e_i) + \beta_0 + \beta_1 \,\text{AFF}_i + \nu_i, \qquad \nu_i \sim \text{CAR}(\sigma_\nu^2),$$

onde ν_i representa o efeito aleatório espacialmente estruturado.

Parâmetro	Estimativa	Std. Error	z-value	p-value
Intercept	2.0955	0.2000	10.48	< 0.001
AFF	-0.2907	1.6323	-0.18	0.859

Tabela 7: Estimativas dos efeitos fixos do modelo Poisson espacial (CAR).

Parâmetro CAR	Estimativa
Variância dependente (de)	2.1776
Range	0.5622
Variância extra	0.0721

Tabela 8: Parâmetros da componente espacial CAR.

Modelo Poisson Espacial Completo (3)

O modelo de Poisson completo considera:

$$Y_i \mid \mu_i \sim \text{Poisson}(\mu_i),$$
$$\log(\mu_i) = \log(e_i) + \beta_0 + \beta_1 \,\text{AFF}_i + \phi_i + \nu_i,$$
$$\phi_i \sim \mathcal{N}(0, \sigma_\phi^2), \qquad \nu_i \sim \text{CAR}(\sigma_\nu^2),$$

onde ϕ_i representa heterogeneidade não estruturada e ν_i dependência espacial.

Parâmetro	Estimativa	Std. Error	z-value	p-value
Intercept	2.0867	0.2890	7.22	< 0.001
AFF	-0.1167	1.6897	-0.07	0.945

Tabela 9: Estimativas dos efeitos fixos do modelo Poisson espacial completo.

Parâmetro CAR	Estimativa
Variância dependente (de)	0.3861
Range	0.9791
Variância extra	0.0006

Componente	Variância
Efeito não estruturado (iid)	0.3492

Tabela 10: Parâmetros das componentes espacial estruturada (CAR) e não estruturada (iid).

Comparação dos Modelos

Modelo	AIC	Deviance	RMSE	MAE	MAE_{RR}
Poisson não espacial	450.60	238.62	7.48	5.09	0.823
Poisson espacial (CAR)	473.98	22.70	10.44	7.59	1.144
Poisson espacial Completo	471.07	24.90	10.44	7.59	1.140

Tabela 11: Comparação das métricas de ajuste e desempenho preditivo entre os modelos.

- **AIC:** o modelo Poisson não espacial apresenta o menor valor, mas não capta a dependência espacial presente nos dados.
- **Deviance:** os modelos espaciais apresentam valores substancialmente mais baixos, indicando um melhor ajustamento estrutural.
- **Efeitos espaciais:** a introdução de efeitos CAR reduz a importância da covariável AFF, que deixa de ser estatisticamente significativa, sugerindo que a associação observada no modelo não espacial é explicada pela autocorrelação espacial.

Análise de Resíduos

Modelo	Variância dos resíduos	Moran's I	p-value
Poisson não espacial	0.775	0.373	< 0.001
Poisson espacial (CAR)	0.430	-0.034	0.563
Poisson espacial Completo	0.389	-0.001	0.423

Tabela 12: Variância dos resíduos e teste de Moran aplicado aos resíduos dos modelos ajustados.

- **Variância dos resíduos:** observa-se uma redução clara da variabilidade residual à medida que são introduzidos efeitos espaciais.
- **Autocorrelação espacial:** os resíduos do modelo Poisson não espacial apresentam dependência espacial significativa, enquanto nos modelos espaciais a hipótese de independência não é rejeitada.

Predição Espacial do SMR e comparação entre modelos

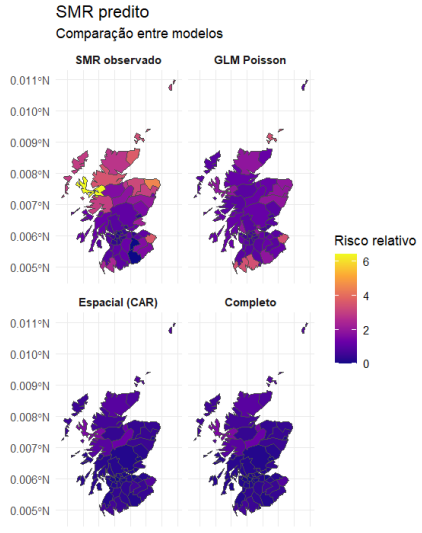


Figura 15 — Mapas de predição do SMR obtidos a partir dos diferentes modelos ajustados.

Conclusão — Dados Agregados por Área

- A análise confirmou a existência de autocorrelação espacial significativa entre condados, justificando o uso de modelos espaciais.
- O modelo Poisson não espacial revelou-se insuficiente, apresentando resíduos espacialmente correlacionados.
- A introdução de efeitos aleatórios (estruturados (CAR) e não estruturados espacialmente) permitiu captar a dependência espacial e reduzir a variabilidade residual.
- Após considerar a estrutura espacial, a covariável AFF deixou de ser estatisticamente significativa, indicando que o seu efeito estava associado a padrões espaciais não modelados.

Obrigado pela atenção.