Projeto Estatísitca Espacial - Análise Exploratória, Modelação e Predição de Dados Geoestatísticos e de Dados Agregados por Área

Rui Alves PG55577

2026-01-05

Contents

1 Introdução

A Estatística Espacial constitui uma área fundamental da Estatística dedicada ao estudo de fenómenos observados no espaço e à modelação da dependência espacial existente entre observações geograficamente localizadas. Em muitos contextos reais, nomeadamente nas ciências ambientais, na geologia e na epidemiologia, a suposição de independência entre observações é frequentemente falsa, uma vez que unidades espaciais próximas tendem a apresentar comportamentos semelhantes. Ignorar essa dependência pode conduzir a inferência enviesada, subestimação da incerteza e previsões pouco fiáveis.

De uma forma geral, os problemas de Estatística Espacial podem ser enquadrados em dois grandes paradigmas metodológicos. O primeiro corresponde à geoestatística, onde se assume que o fenómeno de interesse é uma realização de um processo aleatório contínuo no espaço, observado em localizações pontuais. Neste contexto, ferramentas como o variograma e a krigagem desempenham um papel central na caracterização da estrutura espacial e na interpolação em novos locais. O segundo paradigma diz respeito à modelação espacial de dados agregados por áreas, na qual as observações correspondem a regiões discretas (por exemplo distritos, concelhos, países,...), sendo a dependência espacial modelada através de estruturas de vizinhança.

O presente trabalho tem como principal objetivo aplicar e comparar estes dois paradigmas da Estatística Espacial, recorrendo a conjuntos de dados distintos e a metodologias apropriadas a cada tipo de problema. Numa primeira parte, é realizada uma análise geoestatística de dados espaciais contínuos, incluindo análise exploratória, modelação da estrutura de dependência espacial através do variograma e predição espacial por krigagem. Numa segunda parte, é abordada a modelação de dados agregados por áreas, no contexto da epidemiologia espacial, recorrendo a modelos de regressão Poisson, com e sem efeitos espaciais, para o estudo da incidência de cancro do lábio em condados da Escócia.

2 Dados Geoestatísticos - wrc e soil250

Para a modelação de dados geoestatísticos foi utilizada a base de dados wrc (Water Retention Curve), da biblioteca geoR, que contém 250 observações obtidas numa grelha retangular regular de 10 por 25 pontos, espaçados por 5 metros, a uma profundidade de 25 cm no solo. O conjunto de dados inclui medições da quantidade de água retida no solo a diferentes pressões, bem como informações relativas à densidade do solo.

O objetivo consiste em modelar a quantidade de água retida a diferentes pressões (5, 10, 100 e 15300 mca), recorrendo à informação da localização espacial (abcissa e ordenada) e à densidade do solo (em g/cm^3). Para complementar a análise da influência de covariáveis adicionais, foram ainda consideradas as variáveis AGrossa, Silte e Argila, provenientes do dataset soil250, medidas nas mesmas localizações e representativas da composição granulométrica do solo.

De seguida estão representadas as primeiras seis observações da base de dados inicial, com as medidas da localização das coordenadas X e Y, a densidade do solo, o resultado da quantidade de água medida a diferentes pressões e as quantidades de areia, silte e argila presente no solo.

```
##   CoordX CoordY Densidade    Pr5   Pr10   Pr60  Pr100  Pr306  Pr816 Pr3060
## 1      0      0    1.4605 0.3334 0.3223 0.2768 0.2702 0.2459 0.2391 0.1931
## 2      0      5    1.4487 0.3433 0.3303 0.2808 0.2726 0.2489 0.2451 0.2024
## 3      0     10    1.6127 0.2950 0.2830 0.2585 0.2545 0.2338 0.2202 0.1967
## 4      0     15    1.5847 0.2985 0.2951 0.2688 0.2634 0.2467 0.2190 0.1901
## 5      0     20    1.5249 0.3109 0.3114 0.2783 0.2728 0.2541 0.2292 0.1940
## 6      0     25    1.5081 0.3206 0.3170 0.2776 0.2700 0.2466 0.2180 0.1890
##   Pr15300 Areia Silte Argila
## 1  0.2054     9    26     43
## 2  0.2082     9    26     42
## 3  0.1999     9    25     41
## 4  0.2021    11    28     40
```

```
## 5  0.2113     9    27     41
## 6  0.2130     8    27     43
```

2.1 Análise Exploratória Não Espacial

Antes de proceder à modelação geoestatística, realizou-se uma análise exploratória não espacial com o objetivo de caracterizar a distribuição das variáveis em estudo e compreender as suas principais propriedades estatísticas. Esta etapa permite identificar tendências globais, níveis de variabilidade, assimetrias e valores extremos, fornecendo uma primeira compreensão do comportamento das variáveis, independentemente da sua localização no espaço.

Em primeiro lugar, foi realizada a análise exploratória não espacial das quatro variáveis de retenção de água no solo. Para cada profundidade foram calculadas as principais estatísticas descritivas (mínimo, máximos, quartis, mediana, média) e o respetivo desvio-padrão. Foram ainda visualizados o boxplot e o histograma, de modo a investigar a forma da distribuição dos valores observados.

Table 1: Estatísticas descritivas da quantidade de água retida no solo para diferentes níveis de pressão.

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

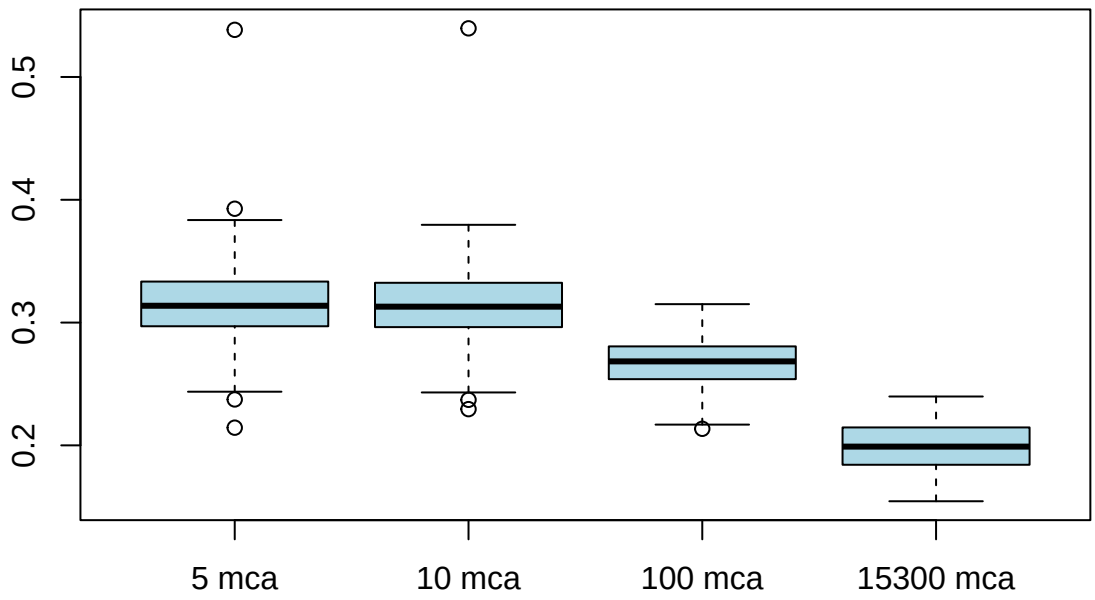
A retenção de água a 5 mca apresenta valores entre 0.214 e 0.538, com média de 0.315 e mediana de 0.314, praticamente coincidentes, o que indica uma distribuição essencialmente simétrica. A maior parte das observações encontra-se entre 0.297 (primeiro quartil) e 0.333 (terceiro quartil), e o desvio-padrão de 0.0313 revela uma variabilidade moderada.

A 10 mca, o comportamento mantém-se muito semelhante, com valores entre 0.229 e 0.540 e média e mediana igualmente próximas (0.314 e 0.313, respetivamente). A simetria da distribuição é elevada e a variabilidade é comparável à observada aos 5 mca, com desvio-padrão de 0.0307.

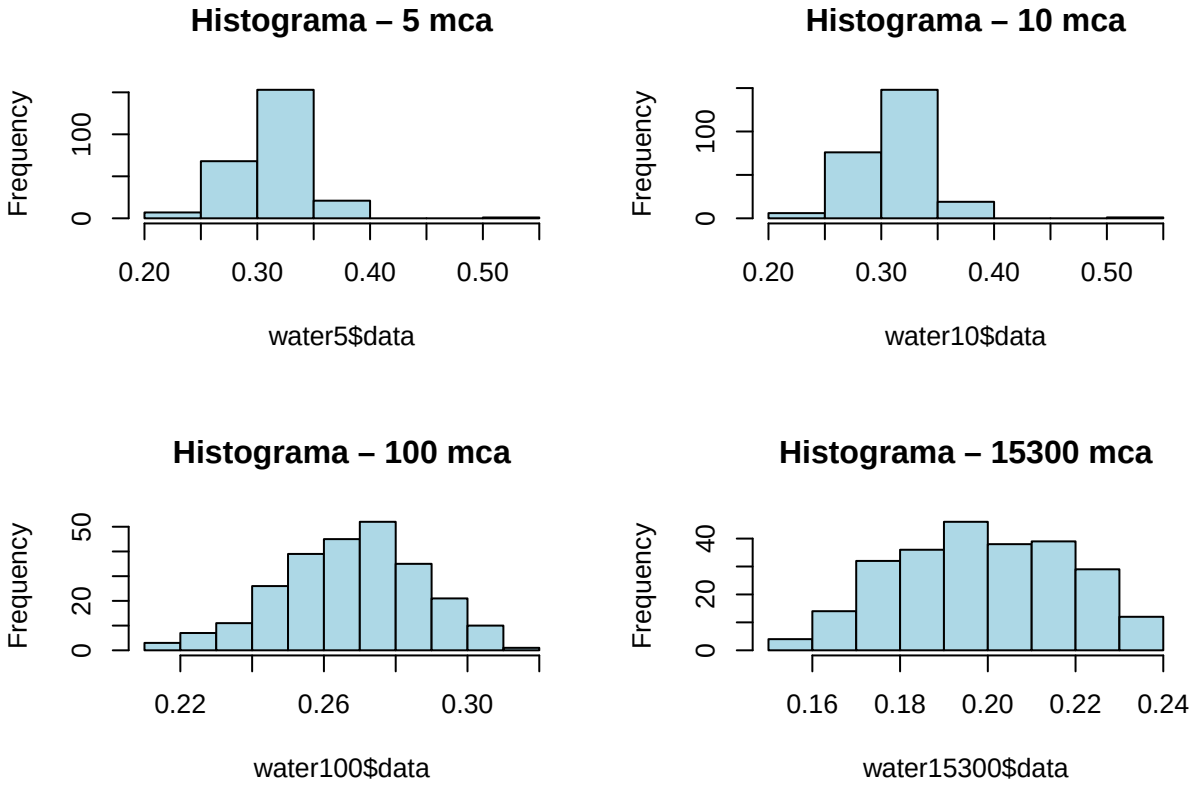
A 100 mca, os valores são globalmente mais baixos e mais concentrados, variando entre 0.214 e 0.315. A média (0.267) e a mediana (0.268) praticamente coincidem, sugerindo uma distribuição equilibrada. O desvio-padrão é menor (0.020), indicando uma redução da dispersão.

Finalmente, aos 15300 mca observa-se a menor retenção de água, com valores entre 0.154 e 0.240, média e mediana ambas iguais a 0.199 e variabilidade reduzida (desvio-padrão de 0.0198), evidenciando elevada homogeneidade.

No conjunto, verifica-se que a retenção de água diminui de forma sistemática com o aumento da pressão. Além disso, a variabilidade é mais elevada em camadas medidas a menor pressão (5 e 10 mca) e torna-se progressivamente menor a pressão mais elevadas (100 e 15300 mca), refletindo uma estrutura mais homogénea dos valores medidos em maior pressão.



Os boxplots mostram que os níveis de retenção de água diminuem à medida que a pressão aumenta. As distribuições correspondentes a 5 mca e 10 mca são muito semelhantes entre si e apresentam valores mais elevados, enquanto que a partir dos 100 mca a retenção torna-se mais baixa e substancialmente mais homogénea.

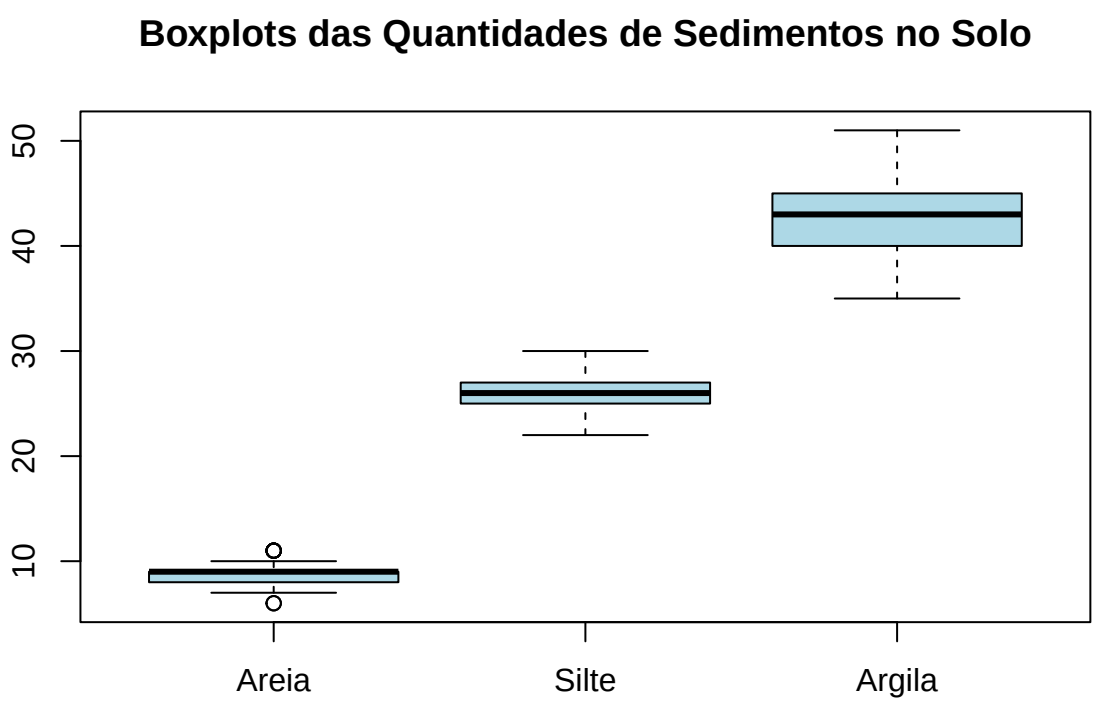


A análise dos histogramas revela que todas as variáveis apresentam distribuições globalmente simétricas, com a maior concentração de valores situada em torno das respetivas medianas. Observa-se uma dispersão mais elevada nas medições realizadas a pressões mais baixas (por exemplo, 5 e 10 mca), enquanto para pressões mais elevadas os valores são progressivamente mais concentrados, refletindo uma distribuição mais estreita e homogénea. Não existe evidência de caudas longas nem de outliers extremos, sugerindo um comportamento regular e bem condicionado das observações ao longo dos diferentes níveis de pressão aplicados.

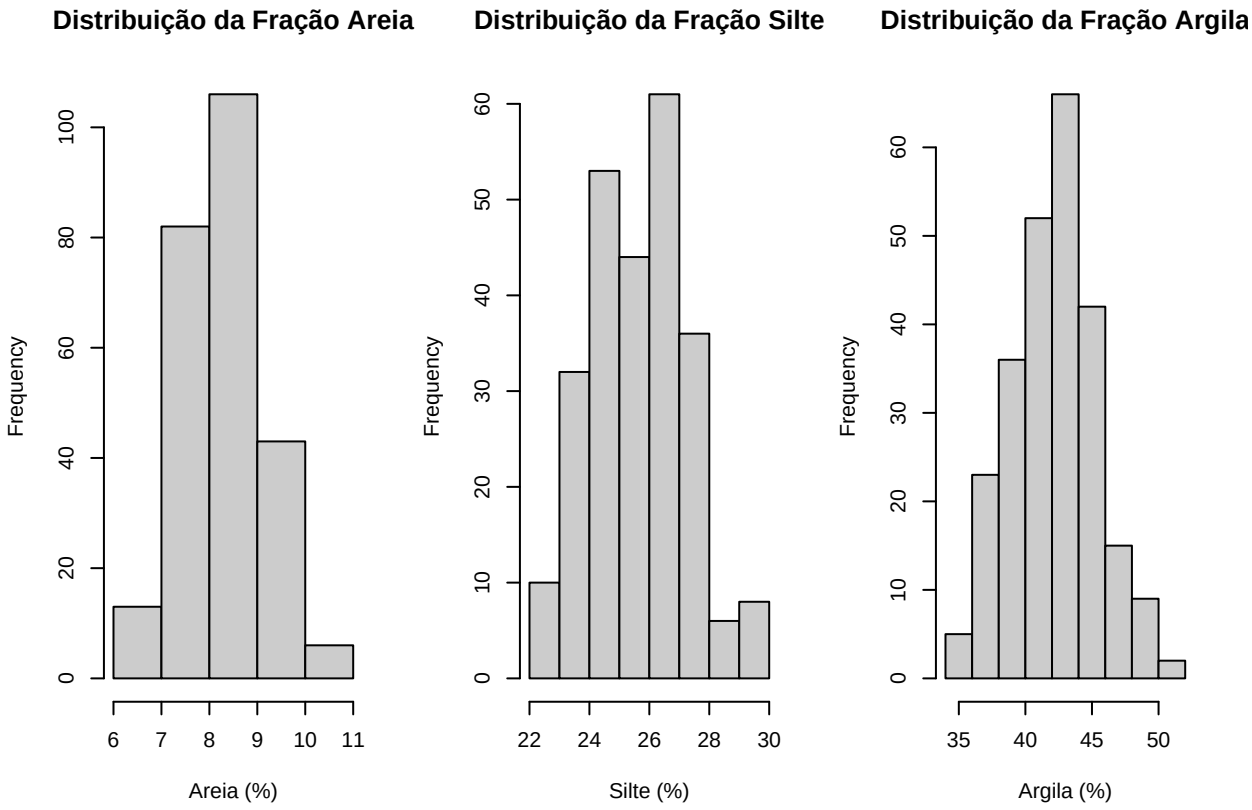
De seguida foi realizada a análise exploratória não espacial das covariáveis em estudo. As três covariáveis sedimentares analisadas: Areia, Silte e Argila, apresentam distribuições distintas que refletem diferentes composições do solo na área de estudo. As estatísticas descritivas mostram que a fração de Areia é claramente a mais baixa (valores entre 6 e 11%), seguida do Silte (entre 22 e 30%), enquanto a Argila é a componente mais abundante (entre 35 e 51%).

Table 2: Estatísticas descritivas das componentes granulométricas do solo.

Componente do solo	Mínimo	1.º Quartil	Mediana	Média	3.º Quartil	Máximo
Areia	6	8	9	8.78	9	11
Silte	22	25	26	26.13	27	30
Argila	35	40	43	42.65	45	51



Os boxplots mostram diferenças claras entre as três componentes do solo. A fração de Areia apresenta valores mais baixos e menor variabilidade, concentrando-se sobretudo entre 8 e 9%. O Silte revela uma distribuição centrada nos 25–27%, com amplitude moderada, enquanto a Argila surge como a componente dominante, com valores consistentemente mais elevados e maior dispersão. Não se observam outliers marcados, o que indica medições estáveis e coerentes ao longo da área amostrada.

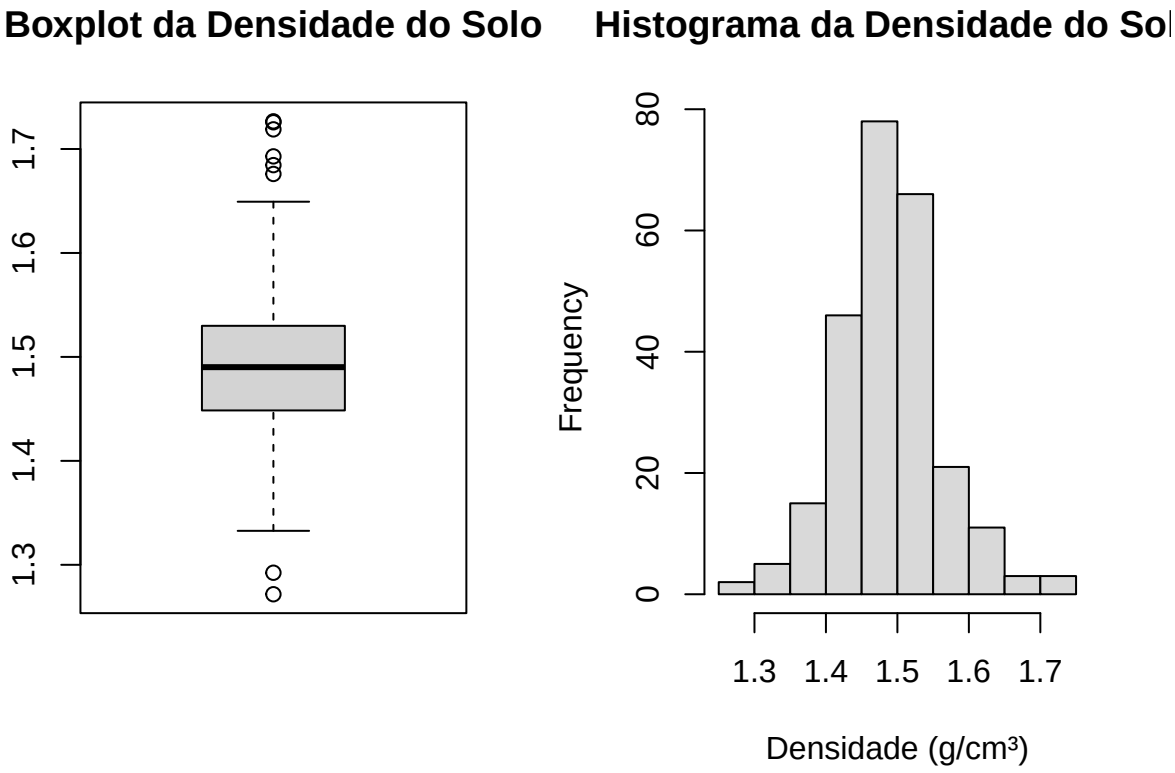


Os histogramas confirmam os padrões observados nos boxplots. A Areia apresenta uma distribuição relativamente estreita e concentrada, com assimetria ligeira devido à menor amplitude de valores. O Silte exibe uma distribuição unimodal centrada nos 25–27%, enquanto a Argila apresenta maior dispersão, ainda que também com formato unimodal e bem definido. No geral, as três distribuições mostram comportamentos regulares, sem valores extremos relevantes.

A variável Densidade do Solo apresenta valores entre 1.272 e 1.727 g/cm³, com média e mediana coincidentes em aproximadamente 1.49 g/cm³, o que sugere uma distribuição bastante equilibrada. O intervalo interquartílico situa-se entre 1.449 (Q1) e 1.529 (Q3), indicando que a maior parte das observações se concentra num intervalo relativamente estreito.

Table 3: Estatísticas descritivas da densidade do solo.

Variável	Mínimo	1.º Quartil	Mediana	Média	3.º Quartil	Máximo
Densidade do solo (g/cm³)	1.272	1.449	1.49	1.49	1.529	1.727



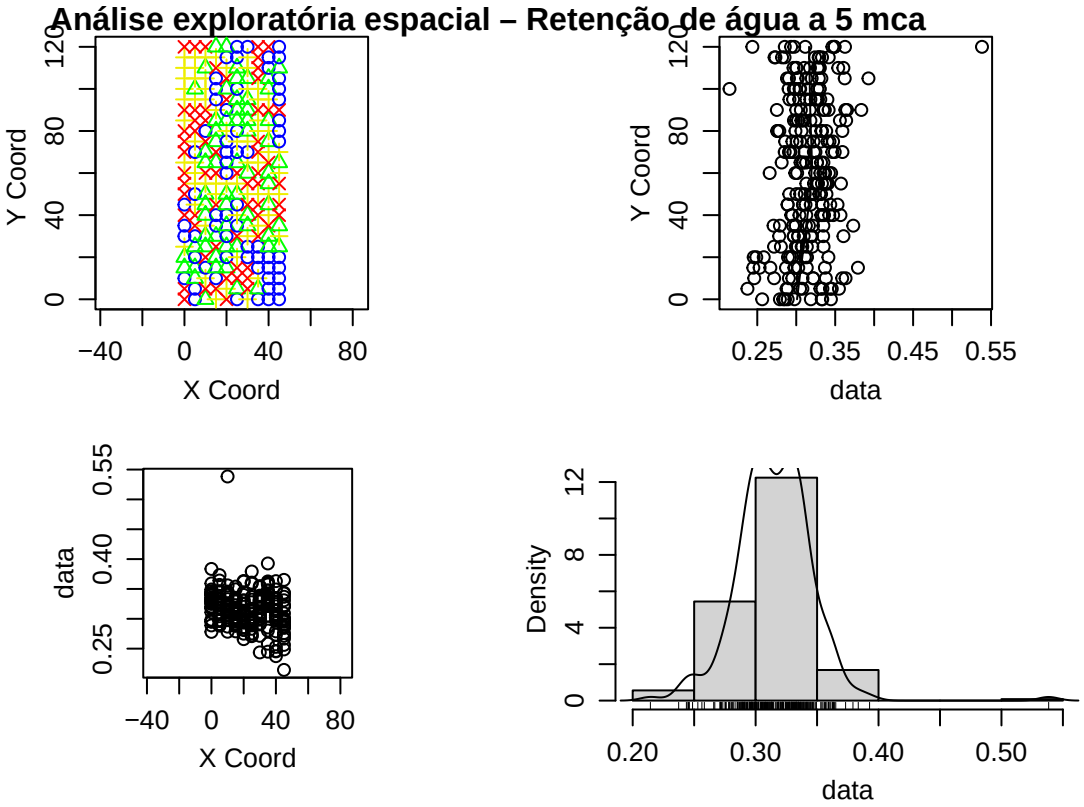
O boxplot mostra uma distribuição centrada, com dispersão moderada e alguns valores extremos tanto abaixo como acima da distribuição principal. O histograma confirma um formato aproximadamente simétrico, com maior concentração de valores em torno de 1.5 g/cm³. A variável apresenta comportamento regular, sem assimetrias pronunciadas, refletindo uma densidade relativamente homogénea no solo ao longo da área amostrada.

2.2 Análise Exploratória Espacial

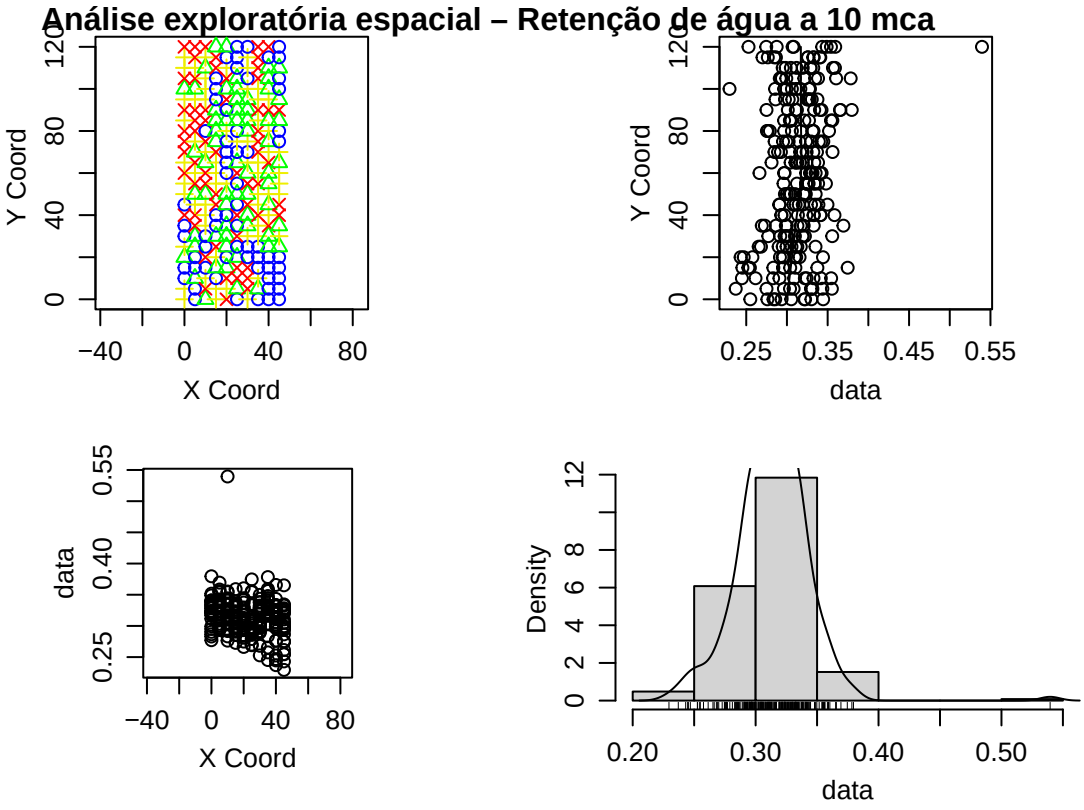
Para a análise espacial, foram considerados os 250 pontos de amostragem do solo como um padrão de pontos georreferenciados associados às medições de retenção de água e às covariáveis físicas do solo. O objetivo desta etapa é caracterizar a distribuição espacial das observações, identificar possíveis tendências ao longo das coordenadas e avaliar a existência de padrões espaciais relevantes. Para tal, analisou-se a dispersão dos pontos no domínio de estudo, examinaram-se estatísticas descritivas das coordenadas e investigaram-se relações entre os valores observados e as posições espaciais.

Através da função ´summary‘ dos objetos geodata verifica-se que as coordenadas variam entre 0 e 45 metros no eixo X e entre 0 e 120 metros no eixo Y, o que confirma que os pontos ocupam todo o domínio da área amostrada. As distâncias entre pontos vão de 5 metros (correspondente ao espaçamento regular da grelha) até cerca de 128 metros, que representa a maior separação possível dentro da região. Isto indica um conjunto de pontos bem distribuído e com cobertura completa do espaço estudado.

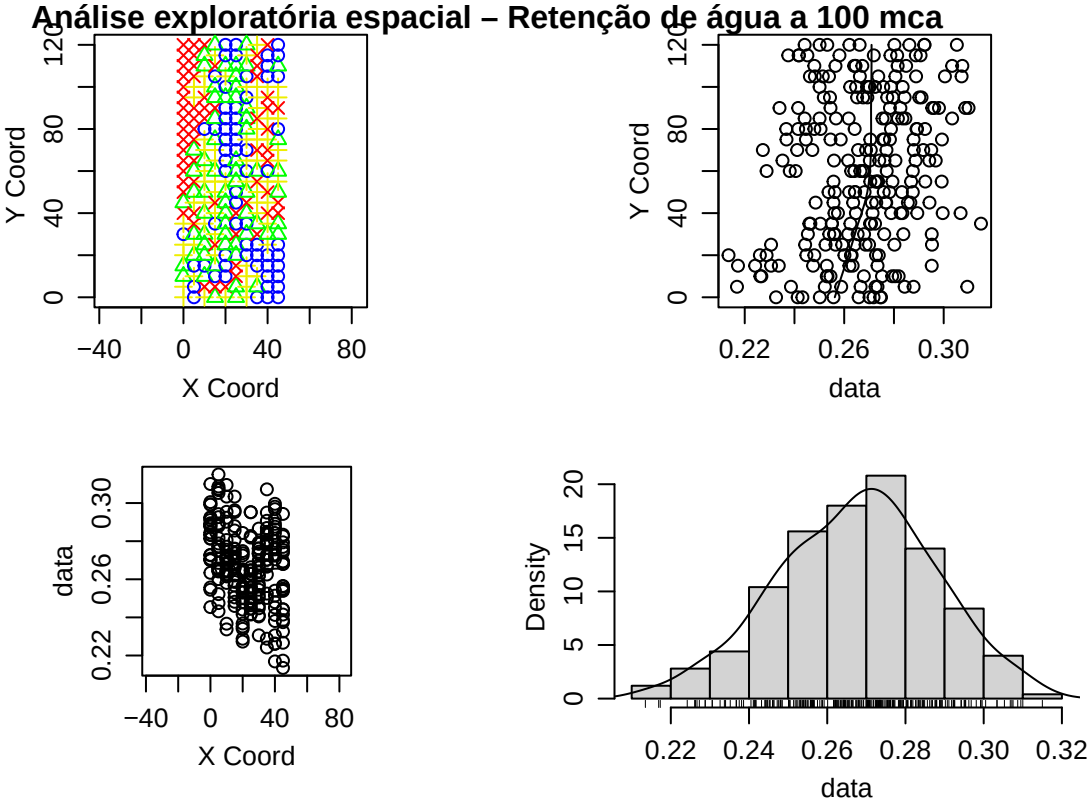
De seguida foram visualizados quatro gráficos obtidos através do plot dos objetos geodata, para cada um dos valores de pressão considerados:



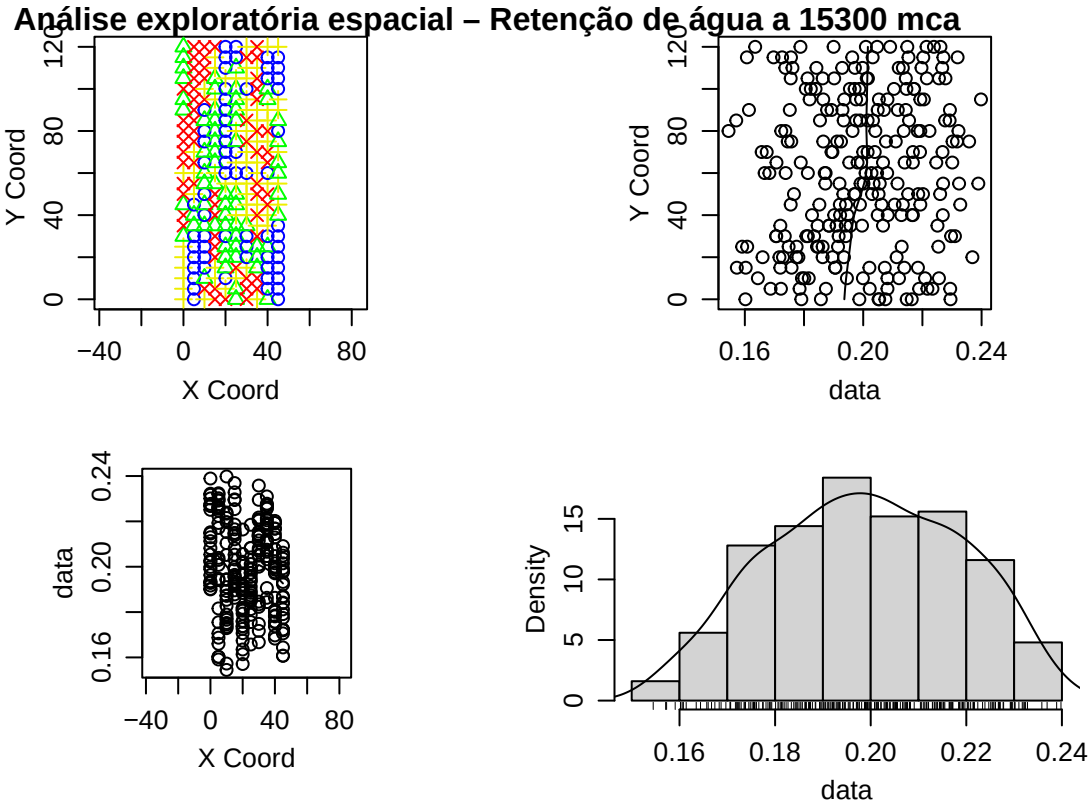
A distribuição dos pontos mostra um grid regular que cobre toda a área de estudo. Parece que à medida que se caminha na direção positiva das abcissas os pontos com valores menores vão surgindo mais. Esta tendência negativa dos dados no aumento da abcissa é confirmada pelo terceiro gráfico do plot do objeto geodata. Em Y não se consegue visualizar com clareza nenhuma tendência.



Para medições efetuadas a 10mca retiram-se conclusões semelhantes às retiradas em 5mca.



Para medições efetuadas a 100mca parece ser ainda mais clara a tendência negativa no aumento da coordenada X, com a presença de valores maiores para valores menores de X e valores inferiores para valores maiores de X. Parece também no segundo gráfico observar-se uma ligeira tendência positiva no aumento de valores de Y, corroborada pelo primeiro gráfico pelo numero de valores superiores (vermelhos) à medida que Y aumenta.



Para medições efetuadas a 15300mca a relação entre os valores e as coordenadas indica apenas variações suaves, sem evidência de tendência clara associada a X ou Y, refletindo um padrão espacial bastante homogéneo.

2.3 Modelação Espacial

Após a análise exploratória espacial, procede-se à modelação geoestatística da retenção de água no solo. O objetivo desta etapa é caracterizar formalmente a estrutura espacial do fenómeno em estudo, distinguindo entre a componente determinística associada à tendência e a componente estocástica responsável pela dependência espacial residual. Esta separação é essencial para garantir uma modelação adequada e para assegurar que os pressupostos subjacentes aos métodos geoestatísticos são satisfeitos.

A modelação espacial foi conduzida de forma sequencial. Numa primeira fase, avaliou-se a existência de tendências espaciais na média do processo, em função das coordenadas espaciais e de covariáveis relevantes. De seguida, a dependência espacial residual foi caracterizada através do variograma empírico, ao qual foram ajustados modelos teóricos adequados, avaliados com recurso a validação cruzada. Por fim, os modelos obtidos foram utilizados para a predição espacial da quantidade de água retida no solo, recorrendo a métodos de krigagem.

2.3.1 Análise de tendências

A verificação de tendências foi realizada através do ajuste de modelos de regressão linear múltipla, considerando como variável resposta a quantidade de água retida no solo e como variáveis explicativas as coordenadas espaciais e as covariáveis físicas do solo. Esta análise permite identificar a existência de uma componente determinística na média do processo espacial, cuja remoção é fundamental antes da modelação da dependência espacial residual.

Para a pressão de 5 mca, os resultados evidenciam a presença de uma tendência espacial significativa ao longo da coordenada X, com um efeito negativo, indicando uma diminuição da retenção de água à medida que X aumenta. Não se observa uma tendência significativa associada à coordenada Y. A densidade do solo apresenta um efeito negativo altamente significativo, sendo a covariável com maior influência na retenção de água. Entre as variáveis granulométricas, a fração argilosa apresenta um efeito positivo significativo, enquanto as frações de areia e silte revelam apenas efeitos marginais.

Table 4: Resultados da regressão linear para a retenção de água a 5 mca.

Variável	Estimativa	P Value	Significância
CoordX	-0.00031	0.00096	***
CoordY	0.00002	0.69975	
Densidade	-0.34370	0.00001	***
Areia	0.00443	0.07323	.
Silte	0.00269	0.07818	.
Argila	0.00275	0.01058	*

Para as medições efetuadas a 10 mca, observa-se um padrão semelhante ao obtido para 5 mca. A coordenada X continua a apresentar um efeito negativo estatisticamente significativo, enquanto a coordenada Y não revela influência significativa. A densidade do solo mantém um efeito negativo forte e altamente significativo. Tal como no caso anterior, a fração argilosa apresenta um efeito positivo significativo, enquanto as restantes componentes granulométricas não evidenciam associação estatisticamente relevante.

Table 5: Resultados da regressão linear para a retenção de água a 10 mca.

Variável	Estimativa	P Value	Significância
CoordX	-0.00036	0.00019	***
CoordY	0.00005	0.21893	
Densidade	-0.32060	0.00001	***
Areia	0.00319	0.20946	
Silte	0.00218	0.16389	
Argila	0.00229	0.03762	*

No caso da retenção de água a 100 mca, a análise revela a presença de tendências espaciais significativas associadas a ambas as coordenadas. A coordenada X mantém um efeito negativo significativo, enquanto a coordenada Y apresenta um efeito positivo, ainda que de menor magnitude. A densidade do solo continua a ser uma variável explicativa relevante, com um efeito negativo altamente significativo. As variáveis granulométricas não apresentam efeitos estatisticamente significativos neste nível de pressão.

Table 6: Resultados da regressão linear para a retenção de água a 100 mca.

Variável	Estimativa	P Value	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	

Para a pressão mais elevada considerada, correspondente à retenção residual de água no solo, não se observa uma tendência significativa associada à coordenada X, enquanto a coordenada Y apresenta um efeito positivo estatisticamente significativo. A densidade do solo continua a exercer um efeito negativo relevante. Ao contrário do observado para pressões mais baixas, as variáveis granulométricas silte e argila apresentam efeitos positivos estatisticamente significativos, sugerindo que, neste regime de elevada pressão, a composição fina do solo desempenha um papel mais importante na retenção de água.

Table 7: Resultados da regressão linear para a retenção de água a 15300 mca.

Variável	Estimativa	P Value	Significância
CoordX	-0.00012	0.17168	
CoordY	0.00009	0.01710	*
Densidade	-0.11190	0.00001	***
Areia	0.00264	0.24907	
Silte	0.00410	0.00386	**
Argila	0.00353	0.00042	***

De forma geral, os resultados indicam a existência de tendências espaciais dependentes do nível de pressão considerado, bem como a influência consistente da densidade do solo na retenção de água. Estes resultados justificam a incorporação explícita de uma componente de tendência na modelação geoestatística, de modo a garantir que a dependência espacial residual é adequadamente caracterizada através do variograma.

2.3.2 Estimação de variogramas empíricos e teóricos (com validação cruzada)

Após a remoção das componentes de tendência identificadas na secção anterior, procedeu-se à estimação dos variogramas empíricos com o objetivo de caracterizar a dependência espacial residual da retenção de água no solo. Em todos os casos, o variograma empírico foi calculado considerando apenas pares de observações até 60% da distância máxima entre pontos amostrados. Esta escolha visa garantir estimativas mais estáveis da semivariância, evitando classes de distância com número reduzido de pares, que tendem a introduzir maior variabilidade e menor fiabilidade nas estimativas.

As tendências removidas em cada caso refletem os resultados da análise de regressão linear previamente realizada. Para as medições a 5 e 10 mca, foram consideradas tendências dependentes da coordenada X, da densidade do solo e da percentagem de argila, enquanto para 100 mca se incluiu adicionalmente a coordenada Y e foi retirada a percentagem de argila. No caso da retenção a 15300 mca, a tendência incorporou a coordenada Y, a densidade do solo e a percentagem de silte e argila, refletindo a maior influência destas componentes a níveis de pressão mais elevados.

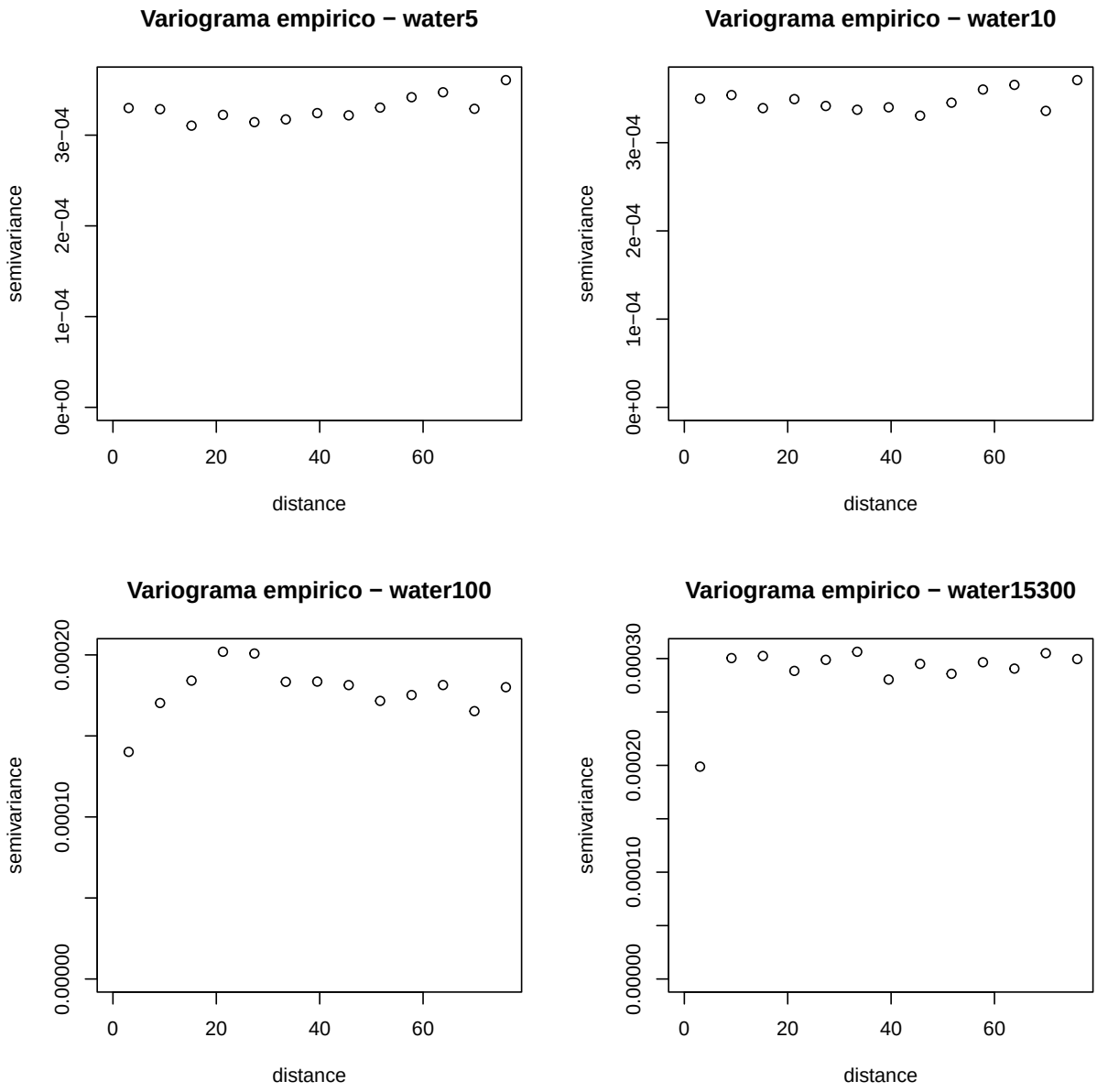
De seguida estão representados os diferentes variogramas obtidos:

```
## variog: computing omnidirectional variogram

## variog: computing omnidirectional variogram

## variog: computing omnidirectional variogram

## variog: computing omnidirectional variogram
```



A inspeção dos variogramas empíricos revela diferenças claras entre os níveis de pressão analisados. Para as medições a 5, 10 e 15300 mca, observa-se uma variação relativamente reduzida da semivariância em função da distância. Em particular, os valores da semivariância mantêm-se aproximadamente constantes ao longo das classes de distância consideradas, não evidenciando um crescimento claro com o aumento da separação espacial entre observações. Este comportamento sugere uma fraca dependência espacial residual, indicando que, após a remoção da tendência, as observações apresentam um grau reduzido de correlação espacial.

Em contraste, o variograma empírico correspondente às medições a 100 mca apresenta uma estrutura espacial mais pronunciada. Neste caso, observa-se uma maior variação da semivariância com a distância, sugerindo a presença de correlação espacial residual relevante. Este padrão indica que, para este nível de pressão, a variabilidade da retenção de água não é totalmente explicada pela tendência e que a componente espacial desempenha um papel mais importante.

Face a estes resultados, a modelação de variogramas teóricos será aprofundada principalmente para o caso de 100 mca, onde a estrutura de dependência espacial se mostra mais evidente e suscetível de ser adequadamente capturada por um modelo geoestatístico. Para os restantes níveis de pressão, a fraca variação observada nos

variogramas empíricos sugere uma contribuição espacial limitada, o que condiciona a utilidade prática de uma modelação variográfica mais complexa.

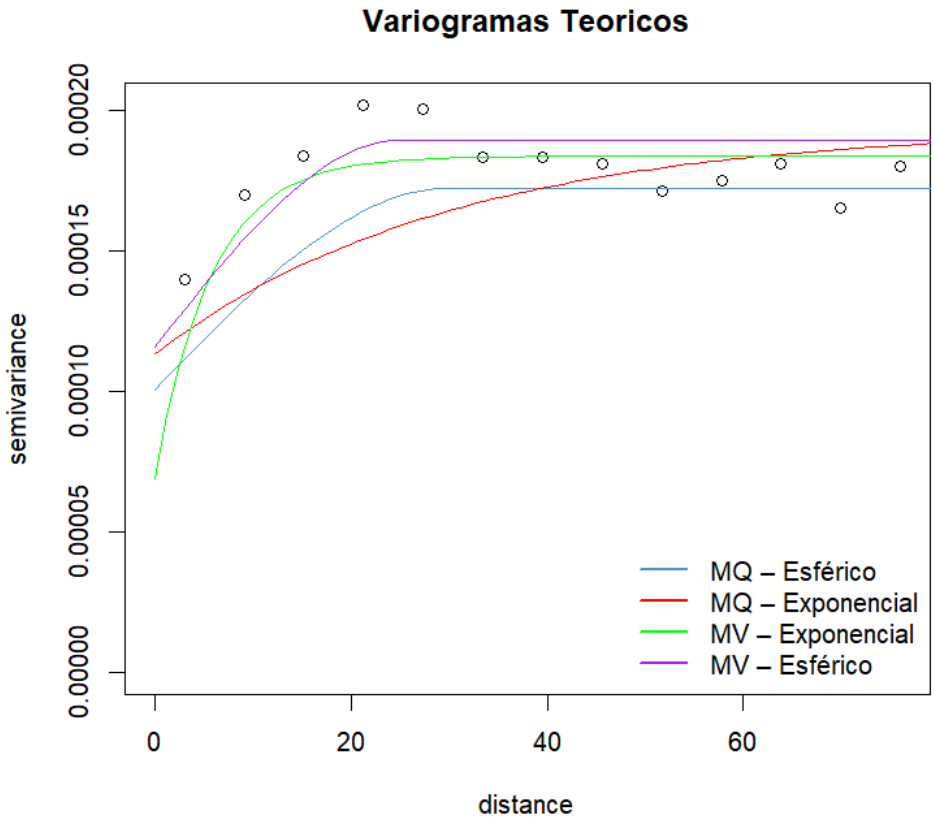
Procedeu-se à estimação de modelos de variograma teóricos com o objetivo de descrever formalmente a estrutura de dependência espacial residual. Dado que este nível de pressão apresentou uma variação mais evidente da variância com a distância, a modelação teórica foi aprofundada apenas para este caso.

Foram considerados dois modelos clássicos de variograma, o esférico e o exponencial, frequentemente utilizados em aplicações geoestatísticas. A estimação dos parâmetros foi realizada através de dois métodos distintos: mínimos quadrados, utilizando a função variofit, e máxima verosimilhança, recorrendo à função likfit.

Os valores iniciais dos parâmetros do variograma, nomeadamente a variância espacial (sigma^2), o range (phi) e o nugget (tau^2) foram inferidos a partir da inspeção do variograma empírico, tendo em conta o nível de estabilização da semivariância e a distância à qual esta estabilização ocorre. Estes valores iniciais serviram como ponto de partida para os procedimentos iterativos de estimação.

No caso do método dos mínimos quadrados, os modelos teóricos foram ajustados diretamente ao variograma empírico, após remoção da tendência previamente identificada. Para o método de máxima verosimilhança, a estimação foi efetuada incorporando a tendência definida em função das coordenadas espaciais e da densidade do solo.

A figura seguinte apresenta a sobreposição dos variogramas teóricos ajustados aos pontos do variograma empírico, evidenciando diferenças na forma como cada modelo e método capturam a estrutura espacial observada. De um modo geral, ambos os modelos considerados conseguem reproduzir adequadamente o comportamento global da semivariância, embora se observem diferenças na rapidez com que atingem o patamar e na forma de transição para a estabilização.



A escolha do modelo de variograma foi realizada com base em validação cruzada leave-one-out (loocv), avaliando simultaneamente a média dos erros de predição e a média do quadrado dos erros padronizados. Um modelo bem ajustado deverá apresentar erros médios próximos de zero e valores do standard error ao quadrado próximos de um.

De acordo com os resultados apresentados na Tabela 8, todos os modelos considerados produzem erros médios muito próximos de zero, sugerindo predições globalmente não enviesadas. No entanto, os modelos ajustados por máxima verosimilhança apresentam um desempenho superior na calibração da variância preditiva. Em particular, o modelo exponencial estimado por máxima verosimilhança destaca-se por apresentar simultaneamente uma média do erro praticamente nula e um valor do erro padronizado ao quadrado muito próximo de um. Com base nestes critérios, este modelo foi selecionado como o mais adequado para descrever a dependência espacial residual da retenção de água a 100 mca, sendo posteriormente utilizado na etapa de predição espacial por krigagem.

Table 8: Resultados da validação cruzada dos modelos de variograma ajustados para a retenção de água a 100 mca.

Modelo	Média.do.erro	Média.do.quadrado.do.std.error
MQ – Esférico	-3.87e-05	1.170
MQ – Exponencial	-2.84e-05	1.130
MV – Esférico	-3.16e-05	1.005
MV – Exponencial	-2.73e-05	1.006

Table 9: Coeficientes estimados do modelo geoestatístico (máxima verosimilhança, covariância exponencial) para a retenção de água a 100 mca.

Parâmetro	Estimativa
Intercept	0.5275
CoordX	-0.0004
CoordY	0.0000
Densidade	-0.1695
nugget	0.0001
sigma^2	0.0001
range	5.8067

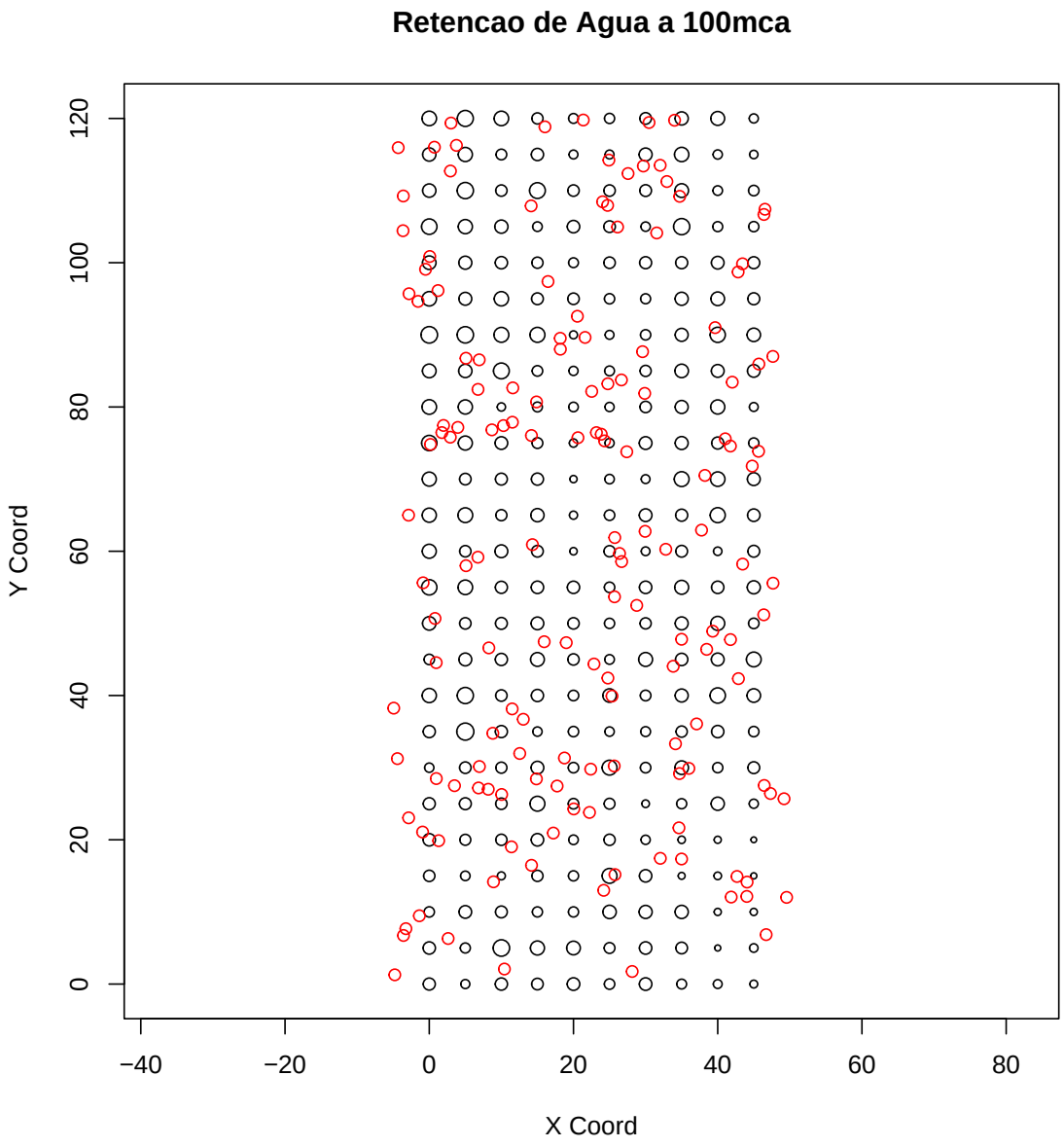
O modelo indica um efeito negativo da densidade do solo na retenção de água a 100 mca, enquanto os efeitos das coordenadas espaciais são residuais após a remoção da tendência. O alcance estimado sugere dependência espacial moderada, sendo a variância do nugget reduzida, o que indica baixo erro de medição ou variabilidade não explicada.

2.3.3 Predição com krigagem

Após a modelação da estrutura de dependência espacial e a seleção do modelo de variograma mais adequado, procede-se à etapa de predição espacial da retenção de água no solo. O objetivo desta fase é estimar os valores da variável de interesse em localizações não observadas, explorando simultaneamente a informação fornecida pela tendência estimada e pela dependência espacial residual capturada pelo variograma teórico selecionado. Para este efeito, recorre-se a métodos de krigagem, que permitem obter predições ótimas no sentido de mínima variância e uma quantificação explícita da incerteza associada às estimativas.

Uma vez que os dados originais foram recolhidos numa grelha espacial regular, optou-se por não realizar a predição exatamente sobre essa grelha, de modo a ilustrar a capacidade do modelo em produzir estimativas

em novos locais. Assim, a predição foi efetuada em 150 pontos gerados aleatoriamente no interior e nas proximidades da área de amostragem original. Para garantir a reprodutibilidade dos resultados, a geração destes pontos foi realizada com recurso à fixação de uma seed.



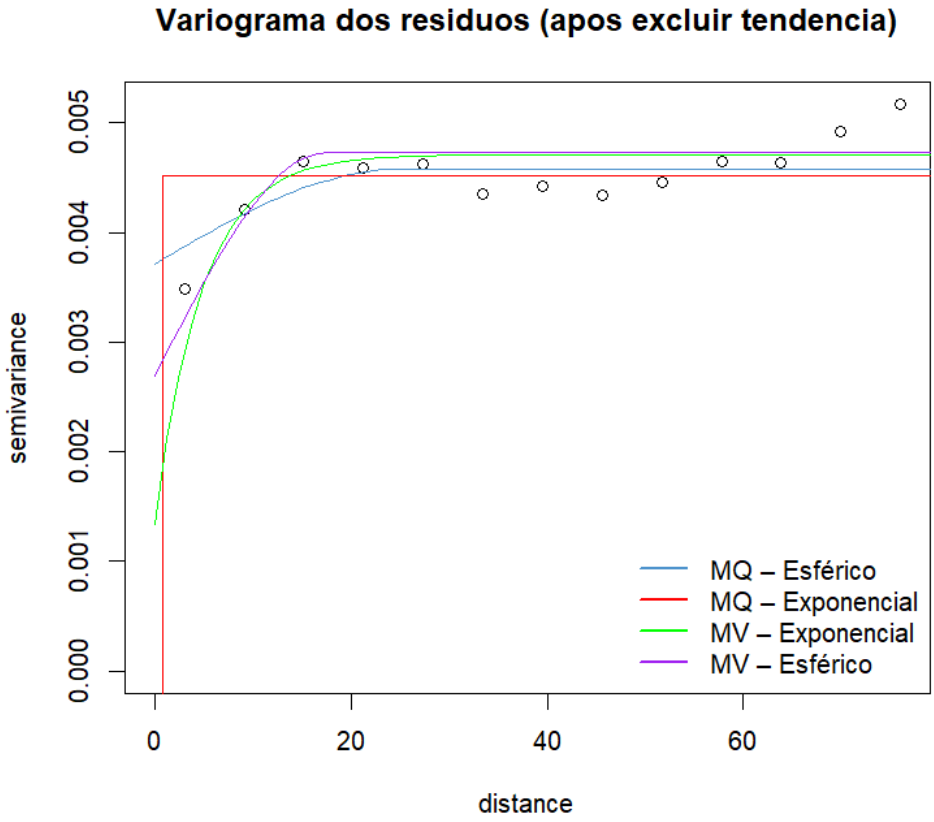
Numa fase seguinte, e uma vez que a predição da retenção de água será realizada através de krigagem com tendência externa, tornou-se necessário obter os valores da covariável densidade do solo nas novas localizações de predição. Como estas localizações não correspondem a pontos originalmente observados, foi previamente necessário estimar a densidade do solo nesses locais.

Para esse efeito, recorreu-se a um procedimento análogo ao utilizado na modelação da variável principal, baseado em métodos geoestatísticos. Em particular, procedeu-se inicialmente à análise exploratória da densidade do solo, avaliando a existência de tendências espaciais. De seguida, foi estimado o variograma empírico dos resíduos, ao qual foram ajustados modelos de variograma teóricos. A seleção do modelo final foi realizada com base em validação cruzada do tipo leave-one-out, assegurando uma escolha informada do modelo que melhor descreve a dependência espacial da covariável.

Table 10: Resultados da regressão linear para a densidade do solo.

Variável	Estimativa	P Value	Significância
X	0.00052	0.12195	
Y	-0.00055	0.00018	***
Areia	0.00208	0.81815	
Silte	-0.00661	0.23471	
Argila	-0.00091	0.81575	

A análise de tendência para a densidade do solo indica a presença de uma componente espacial significativa associada à coordenada Y, com um efeito negativo estatisticamente significativo. As restantes variáveis consideradas, incluindo a coordenada X e as variáveis granulométricas, não apresentam evidência estatística de influência na média da densidade. Estes resultados justificam a inclusão de uma tendência dependente da coordenada Y na modelação geoestatística da densidade do solo, utilizada posteriormente na etapa de krigagem.



O variograma empírico dos resíduos, após a remoção da tendência, evidencia a presença de dependência espacial, com um aumento da semivariância para pequenas distâncias e posterior estabilização num patamar bem definido. Este comportamento indica que observações próximas apresentam valores de densidade do solo mais semelhantes do que observações afastadas.

A comparação entre os modelos teóricos ajustados mostra que o modelo exponencial estimado por máxima verosimilhança apresenta o melhor desempenho, em particular na representação da variabilidade a curtas distâncias. Este resultado é consistente com os critérios de validação cruzada previamente utilizados, nomeadamente a média dos erros próxima de zero e valores do standard error ao quadrado próximos da

unidade. Assim, o modelo exponencial ajustado por máxima verosimilhança foi selecionado para a predição da densidade do solo nas novas localizações.

Table 11: Coeficientes estimados do modelo geoestatístico (máxima verosimilhança, covariância exponencial) para a densidade do solo.

Parâmetro	Estimativa
Intercept	1.5264
CoordY	-0.0006
nugget	0.0013
sigma^2	0.0034
range	4.8056

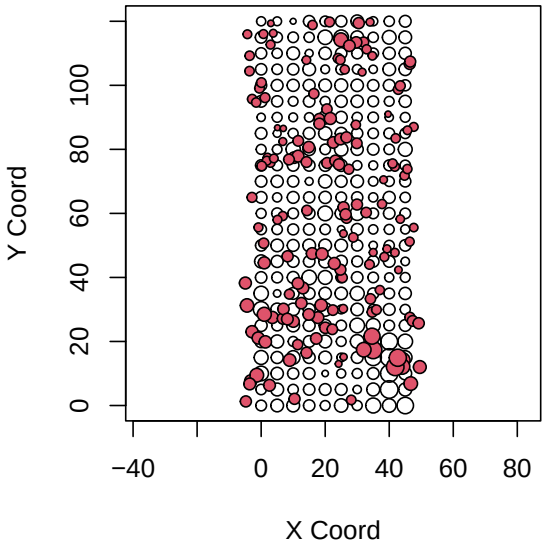
O modelo geoestatístico ajustado para a densidade do solo evidencia uma dependência espacial moderada, com alcance estimado em cerca de 4.8 unidades. A variância espacial é superior ao efeito nugget, indicando que a maior parte da variabilidade da densidade é explicada pela estrutura espacial, enquanto o erro de medição ou micro-variabilidade é relativamente reduzido.

A predição da densidade do solo nas novas localizações foi realizada através de krigagem ordinária, recorrendo ao modelo de variograma previamente selecionado (modelo exponencial ajustado por máxima verosimilhança). A função krige.control foi utilizada para definir a estrutura do modelo, especificando a tendência considerada nos dados observados (trend.d =~ dens_aux$coords[,2]) e nas localizações de predição (trend.l=~ novos.locs[,2]), bem como o modelo de dependência espacial estimado. De seguida, a função krige.conv permitiu obter as estimativas de krigagem para a densidade do solo nos novos pontos.

As observações originais e as estimativas obtidas foram representadas graficamente, permitindo uma visualização conjunta dos dados observados e dos valores preditos. Os pontos correspondentes às observações iniciais são apresentados a preto, enquanto as estimativas de krigagem para as novas localizações surgem assinaladas a vermelho.

```
## kappa not used for the exponential correlation function
## ---------------------------------------------------------------
## likfit: likelihood maximisation using the function optim.
## likfit: Use control() to pass additional
##          arguments for the maximisation function.
##         For further details see documentation for optim.
## likfit: It is highly advisable to run this function several
##         times with different initial values for the parameters.
## likfit: WARNING: This step can be time demanding!
## ---------------------------------------------------------------
## likfit: end of numerical maximisation.

## krige.conv: model with mean defined by covariates provided by the user
## krige.conv: Kriging performed using global neighbourhood
```



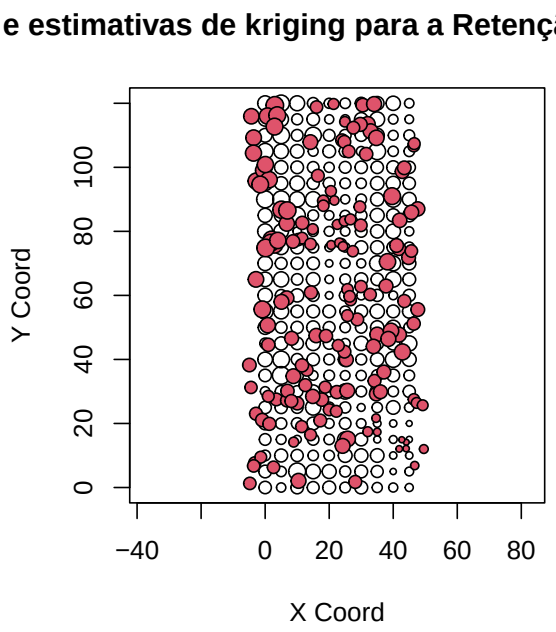
A figura ilustra a distribuição espacial das observações de densidade do solo e das respetivas estimativas de krigagem em novas localizações. Observa-se que os valores preditos tendem a acompanhar os padrões espaciais definidos pelas observações mais próximas, refletindo a natureza local do método de krigagem. Este comportamento evidencia a importância dos pontos vizinhos no processo de predição, uma vez que as estimativas em cada nova localização resultam de combinações ponderadas das observações circundantes, com pesos determinados pela distância e pela estrutura de dependência espacial capturada pelo variograma.

Após a predição da densidade do solo nas novas localizações, esta covariável foi incorporada como tendência externa no modelo de krigagem da retenção de água a 100 mca. Para esse efeito, recorreu-se à krigagem ordinária com tendência, considerando como termos determinísticos as coordenadas espaciais e a densidade do solo predita, assegurando que a informação adicional fornecida por esta covariável fosse devidamente integrada no processo de predição.

A especificação do modelo foi realizada através da função krige.control, definindo a tendência nos dados observados e nas localizações de predição, bem como o modelo de variograma previamente selecionado para a retenção de água. De seguida, a função krige.conv permitiu obter as estimativas de krigagem para a variável de interesse nas novas localizações.

```
## kappa not used for the exponential correlation function
## ----------------------------------------------------------------
## likfit: likelihood maximisation using the function optim.
## likfit: Use control() to pass additional
##          arguments for the maximisation function.
##         For further details see documentation for optim.
## likfit: It is highly advisable to run this function several
##         times with different initial values for the parameters.
## likfit: WARNING: This step can be time demanding!
## ----------------------------------------------------------------
## likfit: end of numerical maximisation.

## krige.conv: model with mean defined by covariates provided by the user
## krige.conv: Kriging performed using global neighbourhood
```



A figura acima apresenta, de forma conjunta, as observações originais e as estimativas de krigagem obtidas com inclusão da densidade do solo como tendência externa. Observa-se que as predições acompanham os padrões espaciais definidos pelas observações vizinhas, refletindo a natureza local do método de krigagem. A incorporação da densidade contribui para uma maior coerência espacial das estimativas, sobretudo em regiões onde a variabilidade da retenção de água está fortemente associada às propriedades físicas do solo. Estes resultados ilustram a relevância da utilização de covariáveis espacialmente estruturadas na melhoria das predições geoestatísticas.

De notar que os pontos estimados fora da grelha inicial terão uma variância maior do que os pontos estimados dentro da grelha, uma vez que estes têm observações mais próximas. Isto foi confirmado através da variância de kriging presente no objeto proveniente da função krige.conv.

2.4 Conclusão

A análise geoestatística realizada permitiu caracterizar de forma detalhada a variabilidade espacial da retenção de água no solo, bem como a influência de covariáveis físicas relevantes, com particular destaque para a densidade do solo. A análise exploratória evidenciou a presença de tendências espaciais dependentes do nível de pressão considerado, justificando a inclusão explícita de componentes determinísticas na modelação da média do processo.

A estimação dos variogramas empíricos mostrou que, para a maioria dos níveis de pressão analisados, a dependência espacial residual é reduzida após a remoção da tendência. Em contraste, para a retenção de água a 100 mca foi identificada uma estrutura espacial mais pronunciada, permitindo uma modelação variográfica consistente. A comparação entre modelos teóricos, através de validação cruzada, conduziu à seleção do modelo exponencial estimado por máxima verosimilhança, que apresentou simultaneamente predições não enviesadas e uma adequada calibração da incerteza.

A etapa de predição espacial evidenciou a utilidade dos métodos de krigagem na estimação de valores em localizações não observadas, assegurando continuidade espacial e integração eficiente da informação disponível. A incorporação da densidade do solo como tendência externa revelou-se particularmente relevante, permitindo melhorar a coerência espacial das predições da retenção de água a 100 mca. No seu conjunto, os resultados

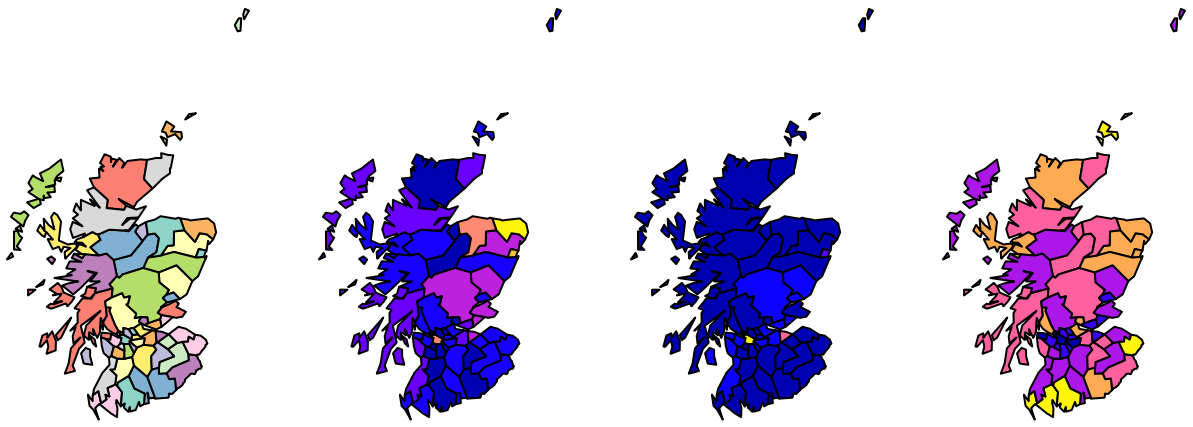
obtidos demonstram a adequação da abordagem geoestatística adotada para a análise de fenómenos contínuos no espaço e evidenciam a importância da consideração simultânea da estrutura espacial e das covariáveis físicas do solo

3 Dados Agregados por Área - scotland_sf

Nesta secção procede-se à análise de dados agregados por área, recorrendo à base de dados scotland_sf, que reúne informação sobre a incidência de cancro do lábio em 56 condados da Escócia, no período de 1975 a 1980. Para cada unidade espacial encontram-se disponíveis o número de casos observados, o número esperado de casos e uma covariável socioeconómica associada à proporção da população empregada nos setores da agricultura e pesca (AFF).

A natureza agregada dos dados e a estrutura espacial das unidades administrativas tornam necessária a adoção de metodologias específicas de Estatística Espacial, capazes de lidar com a dependência espacial entre áreas vizinhas e com a variabilidade inerente a dados de contagem. A inclusão da geometria dos condados, representada em formato sf, permite explorar a distribuição espacial da incidência da doença, avaliar a presença de padrões espaciais e investigar possíveis associações espaciais entre as áreas.

A análise desenvolve-se de forma sequencial, iniciando-se com uma análise exploratória não espacial e espacial dos dados, seguida de testes formais de associação espacial. Posteriormente, são considerados modelos estatísticos espaciais adequados para dados agregados por área, com o objetivo de explicar a variabilidade observada e produzir predições espaciais, culminando numa avaliação global dos resultados obtidos.



3.1 Análise Exploratória Não Espacial

Numa primeira fase, foi realizada uma análise exploratória não espacial com o objetivo de caracterizar globalmente as variáveis da base de dados, sem considerar explicitamente a estrutura espacial das unidades geográficas. Esta etapa permite compreender a escala, a variabilidade e a distribuição das variáveis em estudo, constituindo um ponto de partida fundamental para a análise espacial subsequente. Para esse efeito, analisaram-se medidas descritivas de localização e dispersão, bem como representações gráficas simples, como histogramas e boxplots.

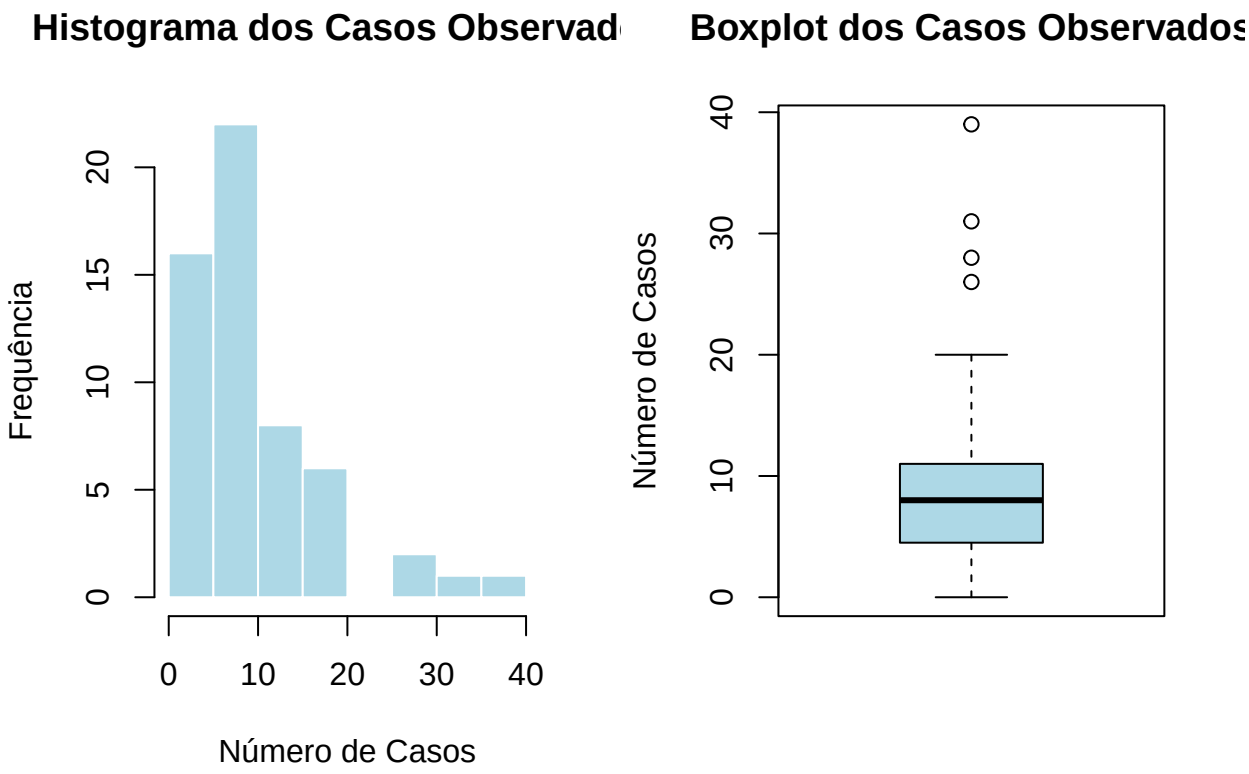
Table 12: Estatísticas descritivas das variáveis da base de dados scotland_sf.

Variavel	Minimo	Q1	Mediana	Media	Q3	Maximo	Desvio.padrão
Casos observados	0.0	4.75	8.00	9.570	11.000	39.00	7.910
Casos esperados	1.1	4.05	6.30	9.580	10.130	88.70	13.180
AFF	0.0	0.01	0.07	0.087	0.115	0.24	0.068

A variável cases apresenta valores entre 0 e 39, com uma mediana de 8 casos e um desvio-padrão de aproximadamente 7.9. Isto indica variabilidade moderada e a presença de alguns condados com números de casos bastante superiores ao típico, o que explica a diferença entre a média (9.57) e a mediana.

A variável expected varia entre 1.1 e 88.7, com mediana 6.3 e desvio-padrão de 13.2. Esta elevada dispersão reflete diferenças consideráveis na dimensão da população em risco entre os condados, sendo que alguns poucos apresentam valores muito superiores aos restantes.

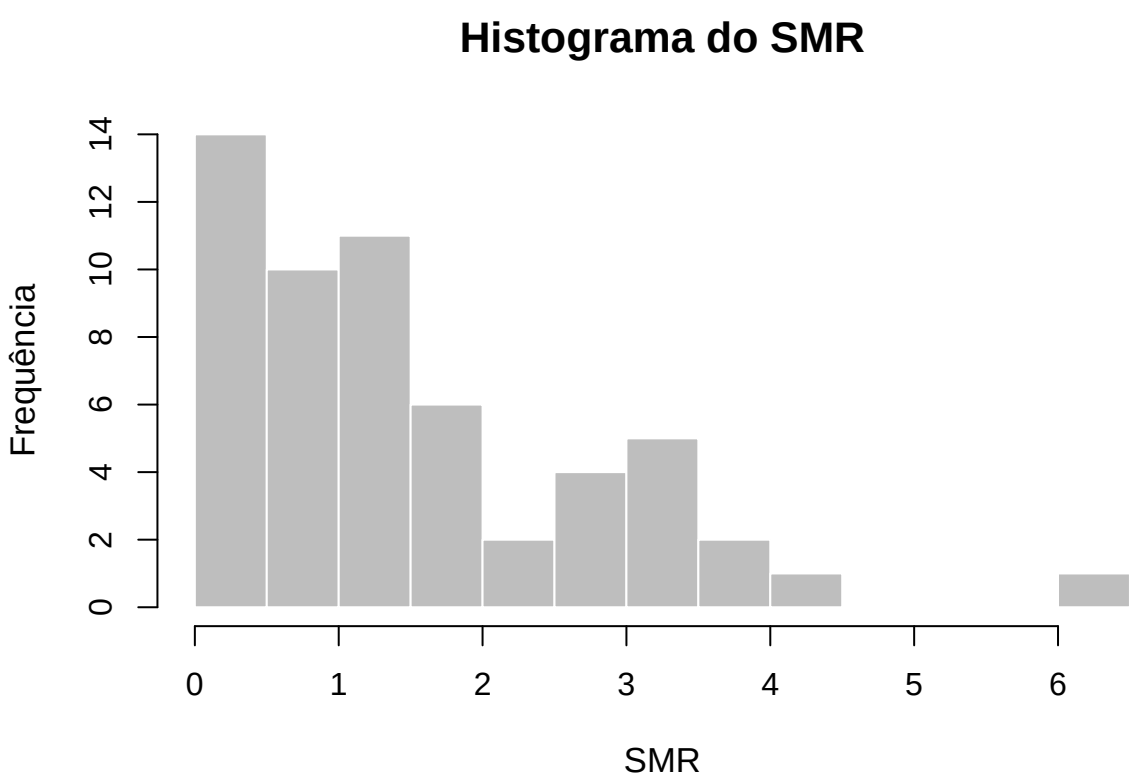
A covariável AFF, que representa a proporção da população empregada nos setores agrícola, piscatório e pecuário, assume valores entre 0 e 0.24, com mediana de 0.07 e desvio-padrão de 0.068. A distribuição é assimétrica para a direita, evidenciando que a maioria dos condados tem uma fração pequena da população nestes setores, enquanto alguns apresentam valores mais elevados.



O histograma mostra que a maioria dos condados apresenta um número relativamente reduzido de casos de cancro do lábio, concentrando-se sobretudo entre 5 e 15 casos. A distribuição apresenta assimetria à direita, evidenciada pela presença de alguns condados com valores substancialmente mais elevados, chegando a um máximo de 39 casos.

O boxplot confirma esta variabilidade, a mediana situa-se próximo dos 8 casos e o intervalo interquartil é relativamente estreito, indicando que a maioria das observações se encontra numa faixa moderada de valores. Contudo, surgem vários pontos acima dos limites superiores da caixa, representando condados com incidência consideravelmente superior ao padrão dominante. Estes valores não devem ser vistos necessariamente como erros, mas sim como áreas potencialmente relevantes do ponto de vista epidemiológico.

O rácio padronizado de mortalidade (SMR), definido como o quociente entre o número de casos observados e o número esperado em cada condado, permite avaliar a variação relativa do risco de cancro do lábio após ajuste pela população em risco. Valores de SMR superiores a 1 indicam maior incidência do que a esperada, enquanto valores inferiores a 1 sugerem risco abaixo do esperado.

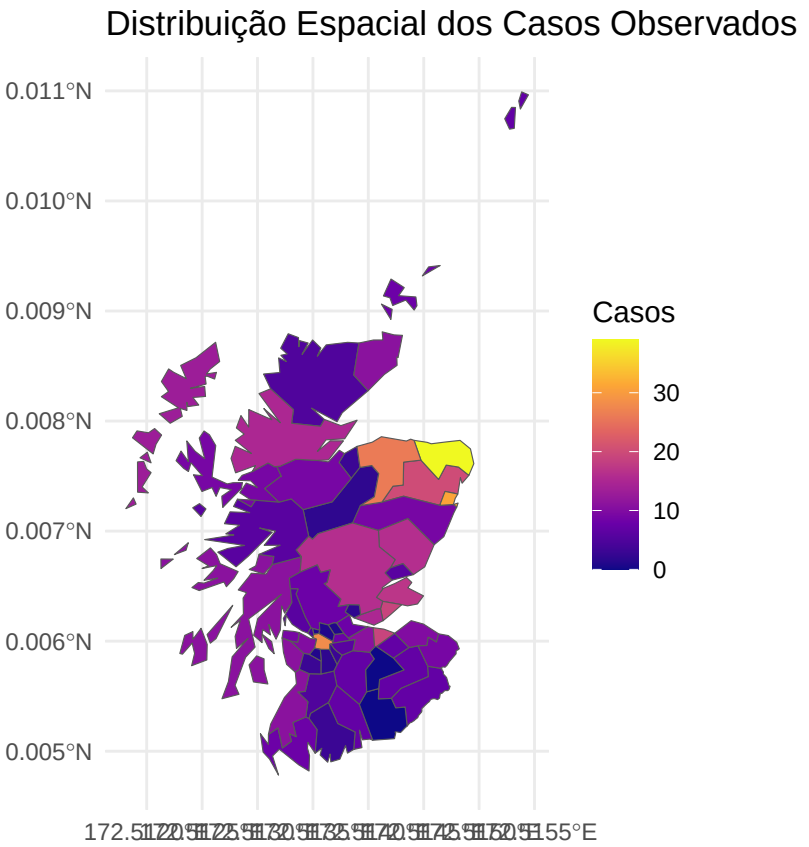


A análise descritiva mostra que o SMR varia entre 0 e 6.43, com uma mediana de 1.11, o que significa que, em metade dos condados, o número de casos observados é aproximadamente igual ou ligeiramente superior ao esperado. O terceiro quartil (2.24) revela que um conjunto considerável de condados apresenta valores substancialmente acima da média, refletindo uma distribuição assimétrica à direita. A média de 1.52, superior à mediana, confirma a influência de alguns condados com SMR muito elevado.

O histograma reforça esta assimetria, mostrando elevada concentração de condados com SMR entre 0 e 2, acompanhada de uma cauda longa que inclui vários valores superiores a 3 e um extremo acima de 6. Este padrão sugere heterogeneidade espacial no risco relativo, com a maioria das áreas apresentando risco moderado, mas algumas regiões destacando-se por incidência expressivamente superior ao esperado.

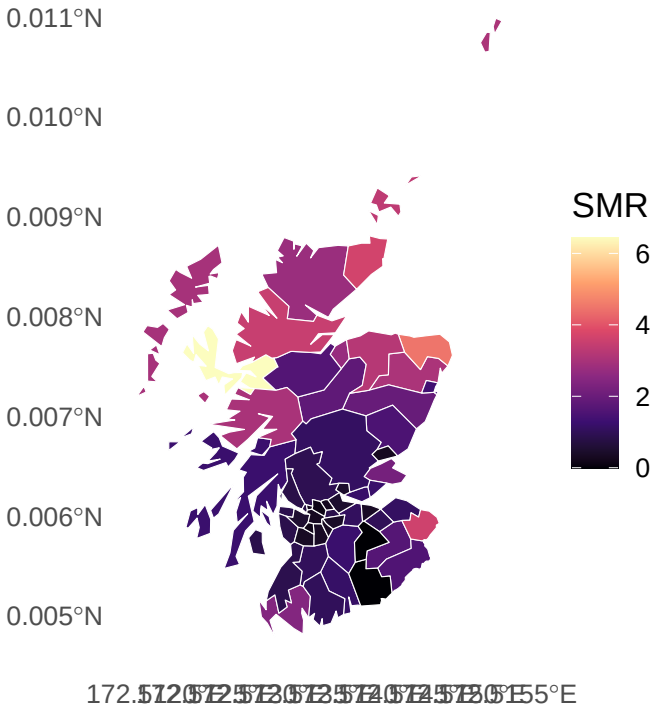
3.2 Análise Exploratória Espacial

A análise exploratória espacial tem como objetivo compreender como os casos de cancro do lábio se distribuem geograficamente pelos 56 condados da Escócia. A representação espacial permite identificar padrões de concentração, possíveis áreas de maior incidência e desigualdades espaciais que não são evidentes através de estatísticas globais. Este primeiro passo é essencial para avaliar a presença de heterogeneidade espacial e motivar análises formais de correlação e modelação espacial.



O mapa mostra uma clara variação espacial no número de casos observados. A maioria dos condados apresenta valores baixos a moderados indicando incidência reduzida. No entanto, destaca-se um conjunto de condados no nordeste do país correspondentes a valores mais elevados, incluindo o máximo de 39 casos. Este padrão sugere que a incidência da doença não é uniformemente distribuída no território e que poderá existir um agrupamento regional de maior risco naquela zona, justificando a aplicação de medidas formais de autocorrelação espacial, como o coeficiente de Moran.

De seguida foi realizado o plot do SMR por condado na Escócia. A representação espacial do SMR permite identificar áreas com potencial risco aumentado e padrões de variação regional que não são evidentes em análises não espaciais.



O mapa mostra uma variação expressiva do SMR entre os condados escoceses. A maioria das áreas apresenta valores moderados, próximos ou ligeiramente acima de 1, representados por tons roxos intermédios. Contudo, destacam-se alguns condados com valores muito elevados (SMR > 4), assinalados em tons amarelos e laranja, indicando incidência muito superior ao esperado, nomeadamente na região nordeste e em algumas áreas isoladas no norte. Destaque para o condado de Lochalsh, que apesar de não possuir muitos casos, tem o SMR maior.

Por outro lado, vários condados do sul exibem valores inferiores ao esperado (tons mais escuros), sugerindo menor risco relativo naquela zona. Este contraste evidencia uma clara heterogeneidade espacial no SMR, que reforça a necessidade de métodos estatísticos capazes de captar dependência espacial.

3.3 Testes de Associação Espacial

Nos modelos espaciais para dados agregados por área, a definição da estrutura de vizinhança é um passo fundamental. A matriz de vizinhança descreve quais áreas são consideradas adjacentes e, portanto, potencialmente semelhantes devido a fatores territoriais, socioeconómicos ou ambientais comuns. Esta estrutura é usada em estatísticas de autocorrelação espacial, como o coeficiente de Moran, e em modelos CAR e SAR, onde a dependência entre regiões é expressa diretamente através da adjacência.

Ao aplicar a função poly2nb, obtém-se a lista de vizinhos (nb) segundo o critério mais simples "queen contiguity", em que duas áreas são vizinhas se partilham um ponto ou um segmento da fronteira. A análise da estrutura de vizinhança mostra que os condados da Escócia apresentam, em média, cerca de quatro vizinhos, com a maioria situando-se entre três e cinco áreas adjacentes. Existem alguns condados isolados, sem quaisquer vizinhos, caso das ilhas. Por outro lado, o condado mais conectado possui onze vizinhos, evidenciando uma posição central no mapa. Esta variação no número de vizinhos é importante, pois afeta a forma como a dependência espacial pode manifestar-se entre as áreas.

Para prosseguir com a utilização desta base de dados para a modelação espacial é necessário confirmar se há correlação espacial entre unidades vizinhas. O coeficiente de Moran I quantifica esse mesmo grau de correlação espacial com base na matriz de pesos definida pela vizinhança. Através do teste de Moran, sob a hipótese nula de ausência de autocorrelação espacial, é possível avaliar se os valores observados tendem a agrupar-se geograficamente.

Table 13: Resultados do teste de Moran global para o número de casos de cancro do lábio.

Estatística	Valor
Moran's I	0.24560
Valor esperado	-0.01920
Variância	0.00830
Z-value	2.90300
p-value	0.00185

O teste apresenta um valor de Moran I = 0.246, claramente superior ao valor esperado sob ausência de autocorrelação (–0.019). O desvio-padrão calculado conduz a um valor-z de 2.903, com um p-value de aproximadamente 0.0018, indicando forte evidência contra a hipótese nula de independência espacial.

Assim, conclui-se que existe autocorrelação espacial positiva, ou seja, condados vizinhos tendem a apresentar números de casos semelhantes. Este resultado justifica a utilização de métodos de modelação espacial, como modelos CAR ou SAR, para capturar adequadamente esta dependência nos dados.

3.4 Modelação Espacial

Após a análise exploratória e a identificação de autocorrelação espacial significativa, procede-se à modelação estatística dos dados agregados por área. Dado que a variável resposta corresponde ao número de casos observados de cancro do lábio em cada condado, foram considerados modelos adequados para dados de contagem, assumindo uma distribuição de Poisson. Esta escolha permite modelar diretamente a natureza discreta e não negativa da variável resposta, bem como a sua relação com covariáveis explicativas.

Em todos os modelos considerados, foi incluído o termo de exposição através do offset correspondente ao logaritmo do número esperado de casos. Este termo permite ajustar as diferenças no risco basal entre áreas, garantindo que os efeitos estimados das covariáveis e da estrutura espacial refletem variações relativas ao risco esperado. Adicionalmente, a covariável socioeconómica AFF foi incluída como efeito fixo em todos os modelos, de modo a avaliar a sua associação com a incidência da doença.

3.4.1 Construção dos Modelos

A modelação foi desenvolvida de forma progressiva, iniciando-se com um modelo de referência não espacial, seguido de modelos que incorporam explicitamente a dependência espacial entre áreas. Em particular, foram considerados modelos Poisson com efeitos aleatórios estruturados espacialmente, bem como um modelo de Poisson completo, que permite decompor a heterogeneidade não explicada em componentes estruturadas espacialmente e não estruturadas. Esta abordagem possibilita avaliar o impacto da dependência espacial na explicação da variabilidade observada e comparar o desempenho dos diferentes modelos considerados.

3.4.2 Modelo Poisson não espacial (modelo de referência)

O modelo de referência assume que o número de casos observados em cada área segue uma distribuição de Poisson, condicionada ao risco esperado e à covariável socioeconómica AFF. Neste modelo não é considerada qualquer dependência espacial entre áreas. De seguida está apresentada a equação do modelo:

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

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

onde Y_i representa o número de casos observados na área i, e_i corresponde ao número esperado de casos (offset), β_0 é o Intercept e β_1 mede o efeito da covariável AFF sobre o risco relativo.

Table 14: Estimativas do modelo Poisson não espacial com offset para os casos esperados.

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

No modelo Poisson não espacial, a covariável AFF apresenta um efeito positivo e estatisticamente significativo sobre o número de casos observados de cancro do lábio. Este resultado indica que áreas com maior proporção de população empregada na agricultura, pesca e atividades afins tendem a apresentar um risco relativo superior, após o ajustamento pelo número esperado de casos. O bom ajustamento global do modelo é refletido pela redução substancial da deviance em relação ao modelo nulo, embora este modelo não considere explicitamente a dependência espacial entre áreas. Este modelo é utilizado como referência para avaliar o impacto da introdução de efeitos espaciais nos modelos subsequentes.

3.4.3 Modelo Poisson espacial com efeito estruturado (CAR)

Para incorporar a dependência espacial entre áreas vizinhas, considera-se um modelo Poisson que inclui um efeito aleatório estruturado espacialmente. Este efeito permite captar padrões espaciais não explicados pelas covariáveis observadas. De seguida está apresentada a equação do modelo:

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

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

$$\nu_i \sim \text{CAR}(\sigma_\nu^2)$$

onde ν_i representa um efeito aleatório espacialmente estruturado, modelado através de um processo autor-regressivo condicional (CAR), definido a partir da matriz de vizinhança entre áreas.

Table 15: Estimativas do modelo Poisson espacial com estrutura CAR e offset para os casos esperados.

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

Table 16: Parâmetros da componente espacial CAR no modelo Poisson.

Parâmetro	Estimativa
Variância dependente (de)	2.1776
Parâmetro de dependência (range)	0.5622
Variância extra	0.0721

Ao incorporar um efeito aleatório estruturado espacialmente através de um modelo CAR, observa-se uma alteração substancial na estimativa e na significância da covariável AFF. Em contraste com o modelo Poisson não espacial, o efeito de AFF deixa de ser estatisticamente significativo, sugerindo que a associação previamente observada pode ser explicada, pelo menos em parte, pela dependência espacial entre áreas vizinhas.

Os parâmetros da estrutura espacial indicam a presença de autocorrelação espacial relevante, capturada pela componente CAR, confirmando os resultados do teste de Moran global. Este resultado evidencia a importância de considerar explicitamente a dependência espacial na modelação de dados agregados por área, uma vez que a omissão desta estrutura pode conduzir a inferências enviesadas sobre os efeitos das covariáveis.

3.4.4 Modelo Poisson espacial Completo

O modelo completo estende o modelo anterior ao decompor a variabilidade residual em dois componentes distintos: um efeito aleatório não estruturado e um efeito aleatório estruturado espacialmente. Esta formulação permite distinguir heterogeneidade específica de cada área de padrões espaciais sistemáticos. De seguida está apresentada a equação do modelo:

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

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

$$\phi_i \sim \mathcal{N}(0, \sigma_\phi^2), \qquad \nu_i \sim \text{CAR}(\sigma_\nu^2)$$

onde ϕ_i corresponde a um efeito aleatório não estruturado, assumido independente entre áreas, e ν_i representa o efeito aleatório estruturado espacialmente.

Table 17: Estimativas dos efeitos fixos do modelo Poisson completo.

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

Table 18: Parâmetros da componente espacial estruturada (CAR) no modelo de Poisson completo.

Parâmetro	Estimativa
Variância dependente (de)	0.3861
Parâmetro de dependência (range)	0.9791
Variância extra	0.0006

Table 19: Variância do efeito aleatório não estruturado no modelo de Poisson Completo.

Componente	Variância
Efeito aleatório não estruturado (iid)	0.3492

No modelo completo, que inclui simultaneamente um efeito aleatório estruturado espacialmente e um efeito aleatório não estruturado, a covariável AFF continua a não apresentar um efeito estatisticamente significativo sobre o número de casos observados. Este resultado reforça a evidência de que a associação positiva observada no modelo Poisson não espacial é largamente explicada pela dependência espacial entre áreas e por heterogeneidade não estruturada específica de cada condado.

Os parâmetros associados à componente espacial estruturada apresentam valores inferiores aos obtidos no modelo CAR simples, sugerindo que parte da variabilidade espacial é agora capturada pelo efeito aleatório não estruturado. A variância estimada para este último confirma a existência de heterogeneidade adicional entre áreas que não é explicada nem pelas covariáveis nem pela estrutura espacial.

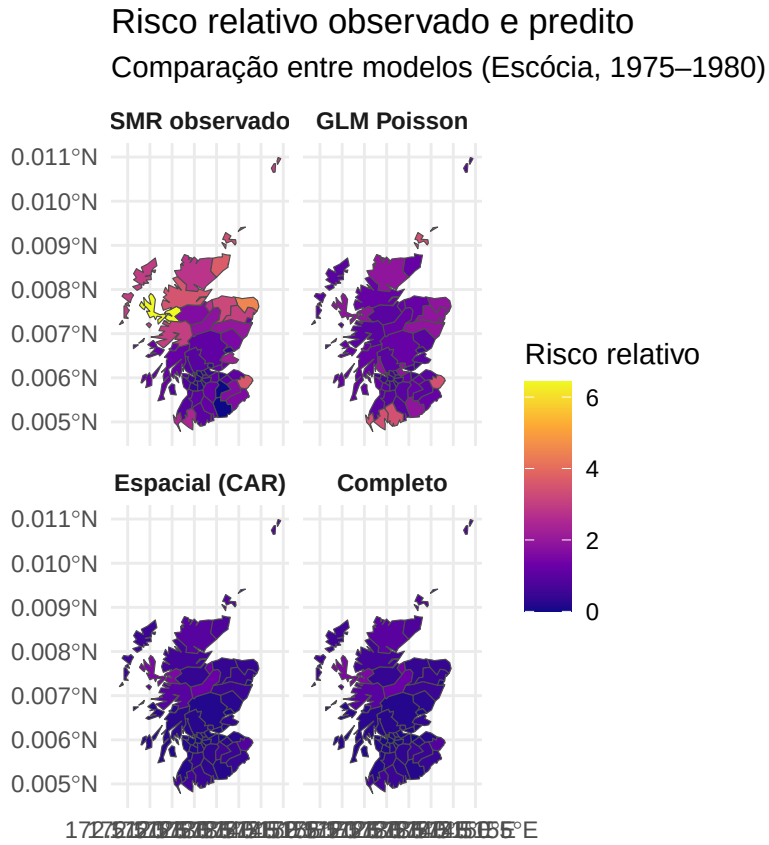
3.4.5 Predição Espacial

Após o ajuste dos diferentes modelos, procedeu-se à obtenção das previsões do número de casos e do risco relativo para cada um dos condados da Escócia. Dado que o objetivo nesta fase é comparar o comportamento dos modelos nos mesmos locais observados, as previsões foram realizadas ao nível das áreas já existentes na base de dados.

No caso do modelo Poisson não espacial, as previsões foram obtidas através da função predict, considerando a escala da resposta (type = "response"), o que permite obter diretamente as estimativas do número esperado de casos em cada condado. A partir destas estimativas, foi posteriormente calculado o risco relativo como a razão entre os casos previstos e os casos esperados.

Para os modelos espaciais, nomeadamente o modelo Poisson com estrutura CAR e o modelo de Poisson completo, recorreu-se à função augment, que permite extrair, de forma integrada, os valores ajustados (fitted values) associados a cada área. Estes valores correspondem às estimativas do número médio de casos segundo cada modelo, já incorporando os efeitos espaciais estruturados e, no caso do modelo completo, também os efeitos não estruturados. De forma análoga ao modelo não espacial, o risco relativo foi obtido através da razão entre os valores ajustados e o número esperado de casos.

Desta forma, as previsões obtidas para cada modelo são diretamente comparáveis, uma vez que se referem às mesmas áreas geográficas e utilizam a mesma definição de risco relativo. Estas previsões servem de base à avaliação quantitativa do desempenho dos modelos, apresentada na secção seguinte através de diferentes métricas de ajuste e erro preditivo.



A comparação dos mapas de risco relativo evidencia diferenças claras entre o modelo Poisson não espacial e os modelos que incorporam dependência espacial. O risco relativo observado apresenta variações acentuadas entre condados, com alguns valores extremos que refletem elevada instabilidade associada a áreas com poucos casos esperados.

O modelo Poisson não espacial tende a reproduzir essas variações de forma direta, resultando em padrões de risco mais irregulares e com maior amplitude. Em contraste, outros modelos produzem estimativas de risco relativo mais contidas e suaves no espaço. Este comportamento resulta da partilha de informação entre áreas vizinhas, característica fundamental dos modelos espaciais, que conduz a um efeito de smoothing.

Entre os dois modelos espaciais as diferenças entre o modelo CAR e o modelo completo são relativamente subtis, indicando que ambos conseguem capturar de forma semelhante a dependência espacial subjacente aos dados.

Em síntese, embora o modelo Poisson não espacial reproduza melhor alguns extremos observados, os modelos espaciais fornecem uma representação mais realista e interpretável do risco relativo, reduzindo a influência de flutuações aleatórias e destacando padrões espaciais consistentes.

3.4.6 Comparação dos Modelos

Para comparar o desempenho dos modelos ajustados de forma quantitativa, foram consideradas métricas de ajuste global e de qualidade preditiva, cada uma captando diferentes aspetos do comportamento dos modelos.

O AIC (Akaike Information Criterion) avalia o compromisso entre qualidade de ajuste e complexidade do modelo, penalizando modelos com maior número de parâmetros. Valores mais baixos indicam modelos preferíveis do ponto de vista parcimonioso.

A deviance mede a discrepância entre o modelo ajustado e o modelo saturado, sendo uma medida de falta de ajuste. Valores elevados de deviance indicam um fraco ajuste global aos dados.

As métricas RMSE (Root Mean Squared Error) e MAE (Mean Absolute Error) avaliam a capacidade preditiva dos modelos, quantificando o erro médio entre os valores observados e os valores ajustados. Enquanto o RMSE penaliza fortemente erros grandes, o MAE fornece uma medida mais robusta do erro médio absoluto.

Adicionalmente, foi avaliado o erro médio absoluto do risco relativo (razão entre casos observados e esperados), permitindo comparar diretamente a capacidade dos modelos em reproduzir padrões de risco espacial.

Table 20: Comparação das métricas de ajuste e desempenho preditivo entre os 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

Os resultados mostram que o modelo Poisson não espacial apresenta valores mais baixos de AIC, RMSE e MAE, sugerindo, à primeira vista, um melhor desempenho em termos de ajuste global e erro preditivo médio. No entanto, este resultado deve ser interpretado com cautela.

Apesar do menor AIC, o modelo não espacial apresenta uma deviance bastante elevada, indicando uma fraca capacidade de explicar adequadamente a estrutura dos dados. Este comportamento é consistente com a existência de dependência espacial não modelada, levando o modelo a absorver padrões espaciais através dos efeitos fixos, nomeadamente da covariável AFF.

Os modelos espaciais apresentam valores de AIC e métricas preditivas ligeiramente piores, o que é esperado dada a maior complexidade estrutural. No entanto, a diferença entre os modelos espaciais e o modelo não espacial não é substancial, sobretudo quando se considera que os modelos espaciais conseguem explicitar e separar a variabilidade associada à dependência espacial.

Importa ainda salientar que os erros associados ao risco relativo são apenas moderadamente superiores nos modelos espaciais, indicando que estes modelos mantêm uma capacidade preditiva aceitável, ao mesmo tempo que fornecem uma interpretação estatisticamente mais adequada do fenómeno em estudo.

Em síntese, embora o modelo Poisson não espacial apresente melhores métricas numéricas em alguns critérios, os modelos espaciais revelam-se metodologicamente mais consistentes, permitindo capturar explicitamente a autocorrelação espacial presente nos dados e evitando conclusões potencialmente enviesadas sobre o efeito das covariáveis.

3.4.7 Análise de Resíduos

A análise de resíduos constitui uma etapa fundamental na avaliação da adequação dos modelos ajustados, permitindo verificar não apenas a magnitude da variabilidade não explicada, mas também a presença de dependência espacial residual. Para esse efeito, analisaram-se a variância dos resíduos e a autocorrelação espacial através do teste de Moran.

Table 21: Variância dos resíduos para os diferentes modelos considerados.

Modelo	Variância.dos.resíduos
Poisson não espacial	0.775
Poisson espacial (CAR)	0.430
Poisson espacial Completo	0.389

Observa-se uma redução clara da variância dos resíduos à medida que se introduzem componentes espaciais no modelo. O modelo Poisson não espacial apresenta a maior variabilidade residual, enquanto os modelos

espaciais, em particular o modelo completo, apresentam valores substancialmente inferiores. Este resultado indica que uma parte significativa da variabilidade não explicada no modelo simples é absorvida pelos efeitos espaciais estruturados e não estruturados.

Para avaliar a independência espacial dos resíduos, foi aplicado o teste de Moran global, considerando como hipótese nula a independência espacial.

Table 22: Resultados do teste de Moran aplicado aos resíduos dos modelos.

Modelo	Moran.I	p.value
Poisson não espacial	0.373	< 0.001
Poisson espacial (CAR)	-0.034	0.563
Poisson espacial (Completo)	-0.001	0.423

Os resultados evidenciam diferenças marcantes entre o modelo não espacial e os modelos espaciais. No modelo Poisson não espacial, o teste de Moran rejeita claramente a hipótese de independência espacial dos resíduos (p-value < 0.001), indicando a presença de autocorrelação espacial não modelada. Este resultado confirma que o modelo simples não é capaz de capturar adequadamente a estrutura espacial subjacente aos dados.

Em contraste, para os modelos espaciais, o teste de Moran não rejeita a hipótese nula de independência espacial dos resíduos, sugerindo que a dependência espacial foi eficazmente incorporada na modelação. Adicionalmente, a redução da variância residual nestes modelos reforça a evidência de um melhor ajuste estrutural, apesar de as métricas preditivas globais não apresentarem melhorias substanciais face ao modelo mais simples.

Em síntese, a análise de resíduos confirma que, embora os modelos espaciais possam apresentar valores ligeiramente superiores em algumas métricas globais, estes oferecem uma representação estatisticamente mais adequada do fenómeno em estudo, ao eliminar a autocorrelação espacial residual e reduzir a variabilidade não explicada.

3.5 Conclusão

A análise dos dados agregados por área evidenciou a presença de dependência espacial significativa na incidência de cancro do lábio nos condados da Escócia, confirmando a necessidade de recorrer a modelos espaciais. O modelo Poisson não espacial, embora apresente melhores valores em algumas métricas globais, revelou resíduos espacialmente correlacionados e uma deviance elevada, indicando uma especificação inadequada.

Os modelos espaciais mostraram-se mais apropriados, ao incorporar explicitamente a estrutura de vizinhança entre áreas, reduzindo a variância dos resíduos e eliminando a autocorrelação espacial residual. Estes modelos produziram estimativas de risco relativo mais estáveis e espacialmente coerentes, permitindo uma interpretação mais fiável dos padrões espaciais observados.

Em síntese, apesar de uma ligeira perda de desempenho em algumas métricas preditivas, os modelos espaciais oferecem uma representação estatisticamente mais adequada do fenómeno em estudo, sendo fundamentais na análise de dados epidemiológicos agregados por área.

4 Referências

Diggle, P. J., & Ribeiro, P. J. (2007). *Model-based Geostatistics*. Springer Series in Statistics. Springer.

Cressie, N. A. C. (1993). *Statistics for Spatial Data*. Wiley, New York.

Soares, A. (2000). *Geoestatística para Ciências da Terra e do Ambiente*. Ensino da Ciência e Tecnologia 9. IST Press.

Lawson, A. B. (2006). *Statistical Methods in Spatial Epidemiology* (2ª ed.). Wiley.

Bivand, R. S., Pebesma, E., & Gómez-Rubio, V. (2013). *Applied Spatial Data Analysis with R* (2ª ed.). Springer.

Elliott, P., Wakefield, J. C., Best, N. G., & Briggs, D. J. (2000). *Spatial Epidemiology: Methods and Applications*. Oxford University Press.

Besag, J., York, J., & Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, 43(1), 1–20.

Ribeiro, P. J., & Diggle, P. J. (2001). geoR: A package for geostatistical analysis. *R News*, 1(2), 14–18.

Dumelle, M., & Ver Hoef, J. M. (2021). spmodel: Spatial statistical modeling and prediction in R. *Journal of Statistical Software*, 99(1), 1–33.

Ver Hoef, J. M., & Dumelle, M. (2023). *spmodel Workshop: Spatial Statistical Modeling with R*. Disponível em: https://usepa.github.io/spmodel.spatialstat2023/

Bailey, T. C. (2008). *An Introduction to Spatial and Spatio-Temporal Modelling of Small Area Disease Rates*. University of Exeter, United Kingdom.

Diggle, P. J. (2000). *Spatial Statistics for Environmental Epidemiology*. University of Lancaster, United Kingdom.