Universidade do Minho

Mestrado em Matemática e Computação

Estatística Espacial

Trabalho Prático

Hugo Filipe de Sá Rocha (PG52250)

Eduardo Teixeira Dias (PG52249)

5 de janeiro de 2025

Conteúdo

Capítulo 1

Introdução

No âmbito da unidade curricular de Estatística Espacial do Mestrado em Matemática e Computação da Universidade do Minho, foi realizado um projeto prático com o intuito de aplicar dois tipos de modelos lineares espaciais: modelos geoestatísticos e modelos referentes a áreas.

Este trabalho consiste na análise e modelação de dados observados numa região, assumindo-se que se tratam de realizações de um processo estocástico espacial. Inicialmente é feita uma apresentação dos dados e uma análise exploratória espacial e não-espacial dos mesmos, incluindo a descrição das principais estatísticas descritivas e principais representações gráficas. Posteriormente, realizou-se a modelação recorrendo-se a regressão linear para dados espacialmente correlacionados, tendo em conta o tipo de dados (dados geoestatísticos ou referentes a áreas). Para além da inferência sobre os parâmetros do modelo, realizou-se também predição espacial.

Capítulo 2

Base de dados *Meuse River* (dados geoestatísticos)

2.1 Constituição da base de dados

A base de dados *Meuse River* fornece localizações e concentrações de metais pesados no solo superficial, juntamente com uma série de variáveis de solo e paisagem nos locais de observação, recolhidas numa planície de inundação do rio Meuse, perto da aldeia de *Stein* (NL). As concentrações de metais pesados são provenientes de amostras compostas de uma área de aproximadamente 15m x 15m.

A base de dados é composta pelas seguintes colunas:

- **x** - vetor numérico; Coordenada Este (metros) no sistema de coordenadas topográficas *Rijksdriehoek* (RDH) dos Países Baixos;
- **y** - vetor numérico; Coordenada Norte (metros) no sistema RDH;
- ***cadmium*** - concentração de cádmio no solo superficial, em $mg\ kg^{-1}$ de solo ("ppm"); valores de cádmio nulos no conjunto de dados original foram ajustados para 0,2 (metade do menor valor não nulo);
- ***copper*** - concentração de cobre no solo superficial, em $mg\ kg^{-1}$ de solo ("ppm");
- ***lead*** - concentração de chumbo no solo superficial, em $mg\ kg^{-1}$ de solo ("ppm");
- ***zinc*** - concentração de zinco no solo superficial, em $mg\ kg^{-1}$ de solo ("ppm");
- ***elev*** - elevação relativa acima do leito do rio local, em metros;
- ***dist*** - distância ao rio *Meuse*; obtida da célula mais próxima na meuse.grid, derivada por uma operação GIS de propagação (distância espacial) com precisão horizontal de 20 metros; depois normalizada para $[0, 1]$;
- ***om*** - matéria orgânica, $kg\ (100\ kg)^{-1}$ de solo (percentagem);
- ***ffreq*** - classe de frequência de inundação: 1 = uma vez a cada dois anos; 2 = uma vez a cada dez anos; 3 = uma vez a cada 50 anos;
- ***soil*** - tipo de solo de acordo com o mapa de solos 1:50 000 dos Países Baixos. 1 = Rd10A (Solos

de prado calcários pouco desenvolvidos, argila arenosa leve); 2 = Rd90C/VII (Solos de prado não calcários pouco desenvolvidos, argila arenosa pesada a argila leve); 3 = Bkd26/VII (Solo de tijolo vermelho, arenoso fino, argila leve);

- ***lime*** - classe de cal: 0 = ausente; 1 = presente conforme teste de campo com HCl a 5%;
- ***landuse*** - classe de uso do solo: Aa = Agricultura/não especificada, Ab = Agr/beterraba sacarina, Ag = Agr/cereais pequenos, Am = Agr/milho, B = floresta, Bw = árvores em pasto, Fh = árvores frutíferas altas, Fl = árvores frutíferas baixas, Fw = árvores frutíferas em pasto, Ga = jardins residenciais, SPO = campo de desporto, STA = curral, W = pastagem.
- ***dist.m*** - distância ao rio *Meuse* em metros, obtida durante o levantamento de campo;

2.2 Análise estatística dos atributos

A imagem abaixo apresenta um resumo estatístico detalhado das variáveis utilizadas no estudo. Cada variável é descrita em termos das suas estatísticas descritivas básicas, incluindo o valor mínimo, o primeiro quartil, a mediana, a média, o terceiro quartil e o valor máximo para os atributos numéricos e a distribuição de frequência para os atributos categóricos. Essas medidas permitem compreender a distribuição dos dados e identificar possíveis padrões ou *outliers*. Este resumo estatístico foi obtido com o comando *summary* do R.

```
       x                y               cadmium          copper           lead            zinc            elev      
 Min.   :178605   Min.   :329714   Min.   : 0.200   Min.   : 14.00   Min.   : 37.0   Min.   : 113.0   Min.   : 5.180  
 1st Qu.:179371   1st Qu.:330762   1st Qu.: 0.800   1st Qu.: 23.00   1st Qu.: 72.5   1st Qu.: 198.0   1st Qu.: 7.546  
 Median :179991   Median :331633   Median : 2.100   Median : 31.00   Median :123.0   Median : 326.0   Median : 8.180  
 Mean   :180005   Mean   :331635   Mean   : 3.246   Mean   : 40.32   Mean   :153.4   Mean   : 469.7   Mean   : 8.165  
 3rd Qu.:180630   3rd Qu.:332463   3rd Qu.: 3.850   3rd Qu.: 49.50   3rd Qu.:207.0   3rd Qu.: 674.5   3rd Qu.: 8.955  
 Max.   :181390   Max.   :333611   Max.   :18.100   Max.   :128.00   Max.   :654.0   Max.   :1839.0   Max.   :10.520  

      dist              om          ffreq  soil   lime       landuse      dist.m      
 Min.   :0.00000   Min.   : 1.000   1:84   1:97   0:111   W      :50   Min.   :  10.0  
 1st Qu.:0.07569   1st Qu.: 5.300   2:48   2:46   1: 44   Ah     :39   1st Qu.:  80.0  
 Median :0.21184   Median : 6.900   3:23   3:12           Am     :22   Median : 270.0  
 Mean   :0.24002   Mean   : 7.478                         Fw     :10   Mean   : 290.3  
 3rd Qu.:0.36407   3rd Qu.: 9.000                         Ab     : 8   3rd Qu.: 450.0  
 Max.   :0.88039   Max.   :17.000                         (Other):25   Max.   :1000.0  
                   NA's   :2                              NA's   : 1                   
```

Figura 2.1: Análise estatística dos atributos usando o comando *summary* do R.

Algumas conclusões que se podem tirar, por exemplo nas variáveis categóricas, grande maioria das medições não registaram presença de cal no solo, na maioria das zonas verifica-se inundações a cada dois anos e ainda são constituídas por prado calcários pouco desenvolvidos de argila arenosa leve. Além disso, também existe um maior número de usos do solo para pastagem.

Para a realização deste estudo, decidimos selecionar a variável ***zinc*** para ser a nossa variável de interesse, que, como visto anteriormente, representa a concentração de zinco no solo superficial. Os valores desta variável variam entre o mínimo de 113 e o máximo de 1839 apresentando uma média de 469.7 e mediana de 326.00, ou seja, metade das medições encontram-se acima deste valor e metade são inferiores ao mesmo. Conclui-se também através dos valores do primeiro e terceiro quartil que 25% dos valores encontram-se abaixo de 198 e 75% encontram-se abaixo de 674.5, respetivamente.

Com vista a estudar com mais detalhe a nossa variável de interesse, analisou-se o histograma da mesma, representado de seguida.

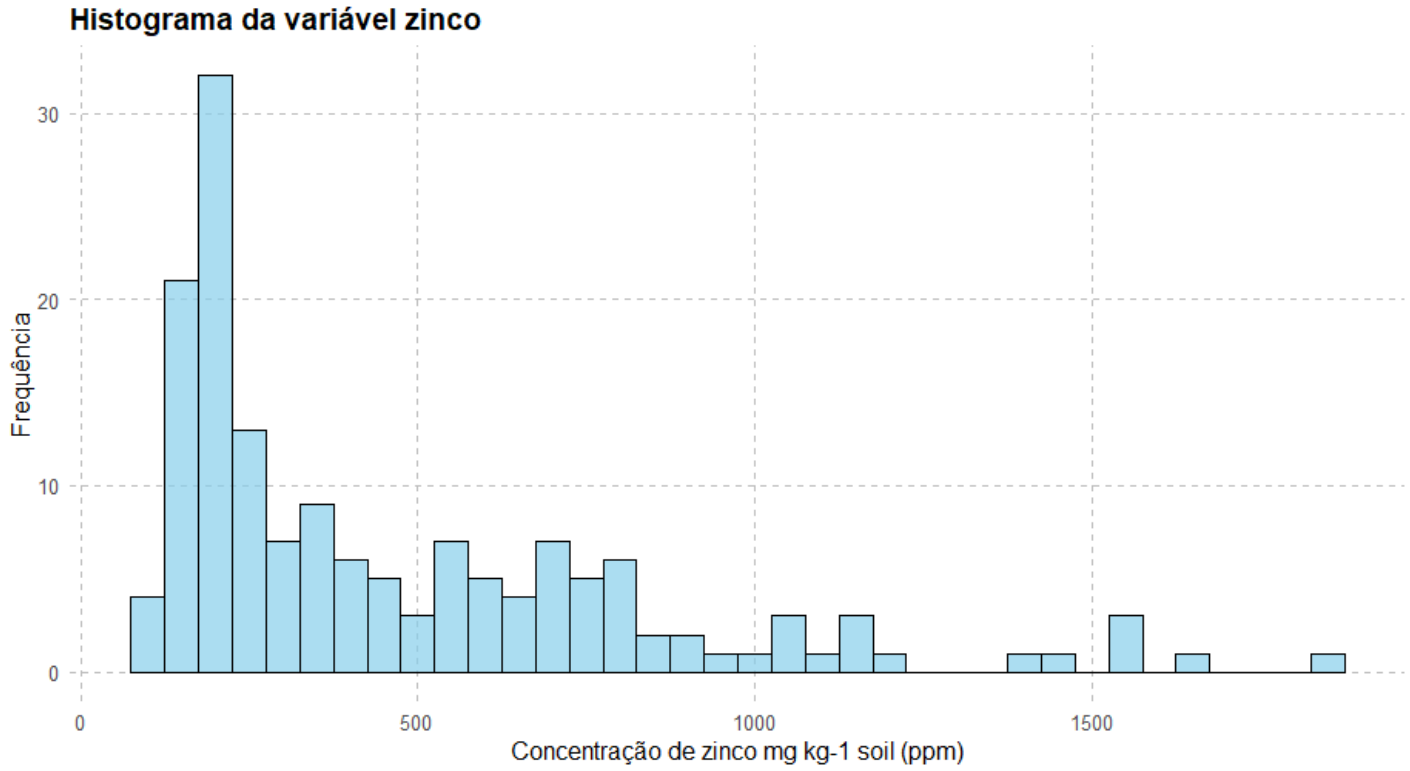


Figura 2.2: Histograma da variável *zinc*.

Como podemos verificar, o histograma é assimétrico à direita, revelando que existe uma maior quantidade de registos cujas concentrações de zinco são baixas, isto é, até ao valor de 250 ppm situam-se o maior número de observações do *dataset* sendo que a partir desse valor de concentração de zinco registadas a frequência de observações vai diminuindo consideravelmente. É de notar também alguns possíveis outliers acima de 1500 ppm que poderão indicar locias onde ocorrem contaminações do rio.

2.3 Correlação entre variáveis

Com o objetivo de estudar a correlação entre as variáveis, obteu-se a matriz de correlação das mesmas como se mostra abaixo.

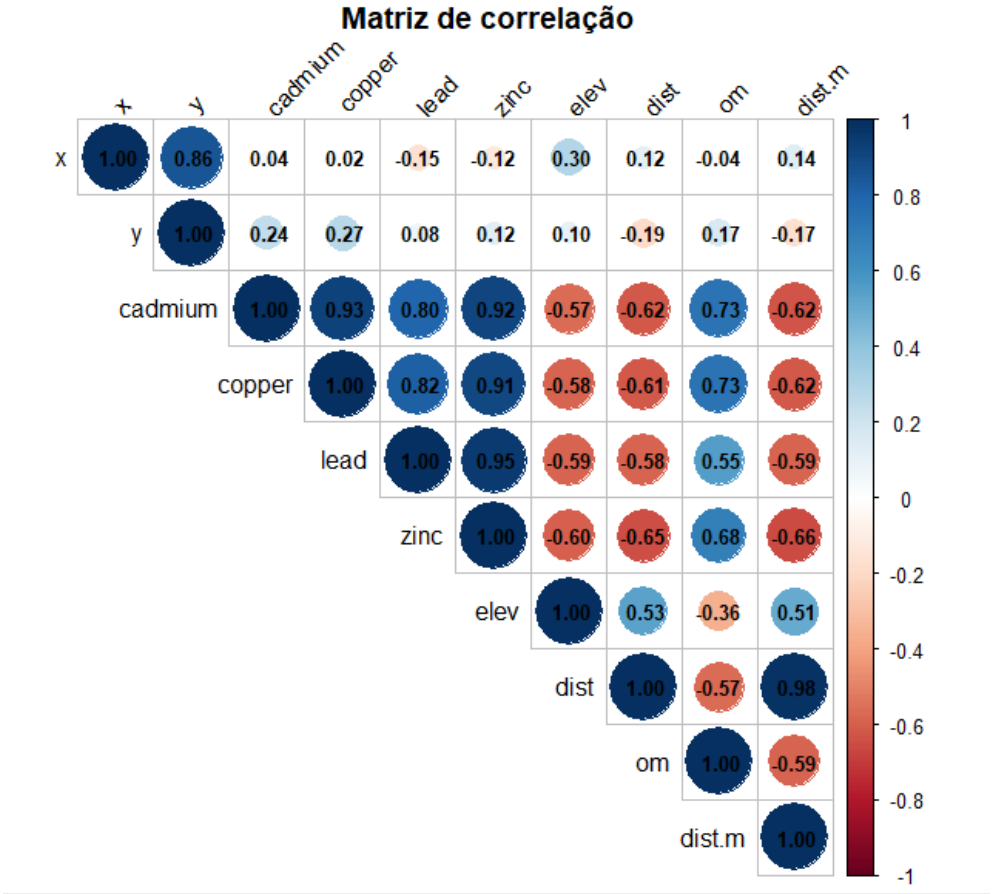


Figura 2.3: Matriz de correlação entre variáveis.

Como se pode verificar na matriz, os metais pesados cádmio, cobre, chumbo e zinco apresentam todos uma correlação elevada entre si. Além disso, a nossa variável de interesse ("*zinc*") está negativamente correlacionada com a elevação e com a distância, isto é, quanto maior a distância do local ao rio e maior a elevação do mesmo, menor a concentração de zinco registada no local e vice versa e positivamente correlacionada com a percentagem de matéria orgânica no solo que poderá acontecer devido, entre inúmeros motivos, à capacidade de absorção de metais por parte deste tipo de matéria.

Além disso, foi também feito um estudo da relação entre os diferentes metais pesados e o atributo "*elev*"referente à elevação. O resultado foi o apresentado abaixo.

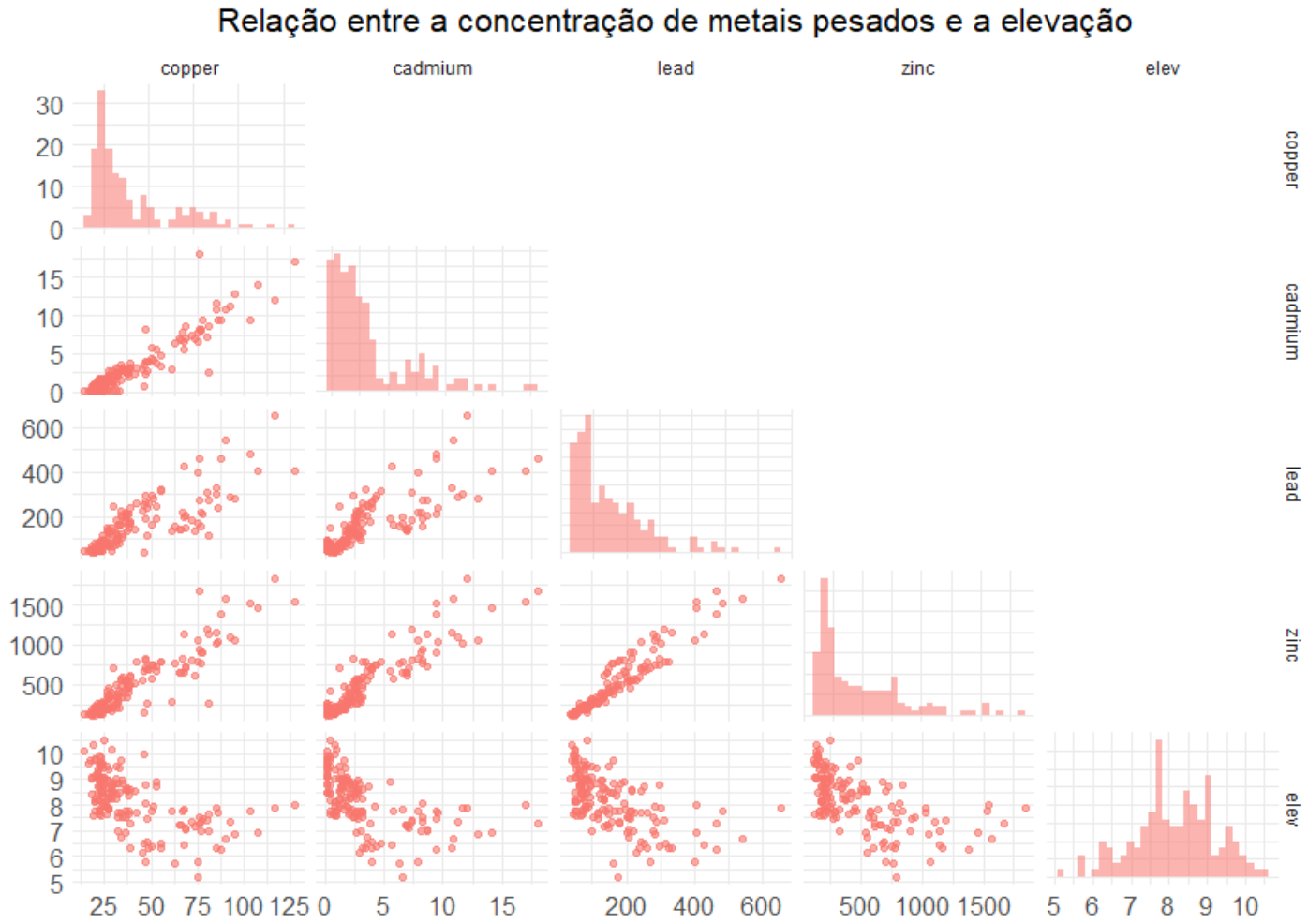


Figura 2.4: Relação entre a concentração dos metais pesados e a elevação.

Como se pode verificar, existe uma relação quase perfeita entre os diferentes metais pesados. No que toca à elevação, essa relação já não é tão evidente pela análise dos gráficos, não havendo um padrão tão definido. É também notório que todos os gistogramas dos metais têm uma distribuição assimétrica à direita, mostrando uma muito maior frequência de observações com concentrações baixas enquanto que a elevação está muito melhor distribuida, tendo uma distribuição muito mais próxima duma distribuição Gaussiana.

2.4 Análise exploratória espacial

Para tentar perceber onde se situam as observações do *dataset* bem como as maiores e menores medições da concentração de zinco foram produzidos os seguintes gráficos.

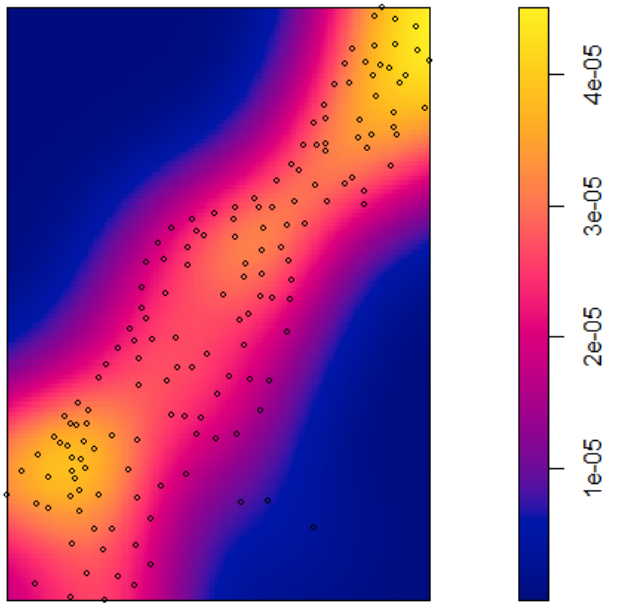


Figura 2.5: Densidade de pontos.

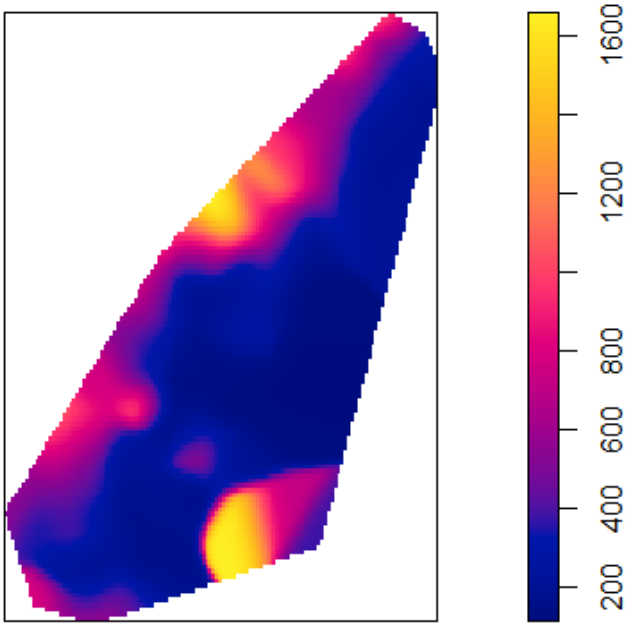


Figura 2.6: Concentrações de zinco.

Como se pode ver no primeiro gráfico, os pontos parecem estar bem distribuídos pelo rio e a zona onde se efetuaram observações segue o aspeto que se previa visto que estamos perante medições nas margens e planícies de inundação de um rio.

Já no segundo gráfico, observam-se os valores de medição de zinco que é a nossa variável de interesse e que são mais elevados na margem superior do rio e na zona a sudeste do mesmo.

Para além disso, pela análise visual dos gráficos da Figura 2.7, é plausível concluir que não existe estacionariedade na média visto que não existe um padrão notório nos gráficos que relacionam a abcissa e a ordenada com a variável de interesse, e portanto, existe uma tendência não constante.

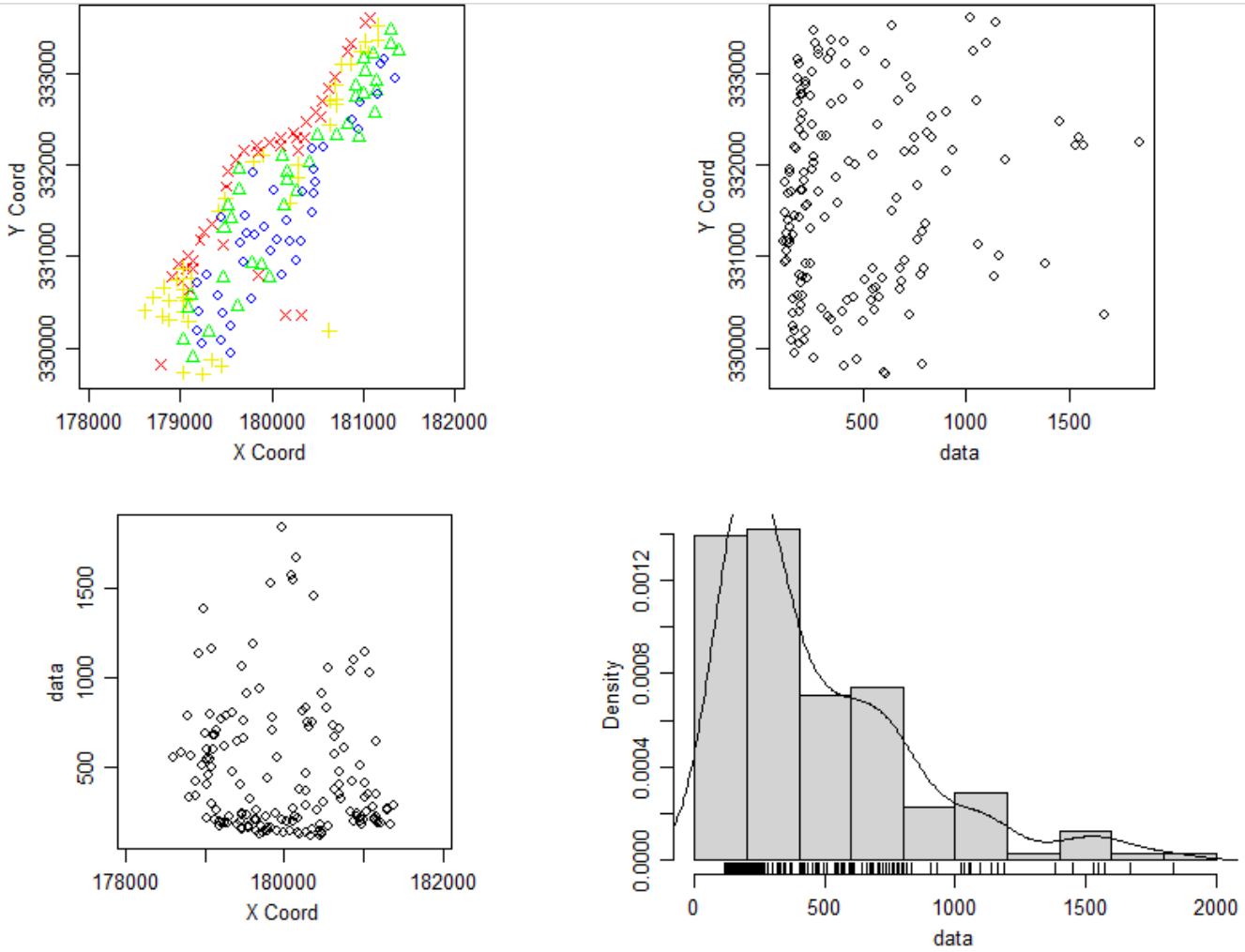


Figura 2.7: Estimação da tendência espacial.

2.5 Estimação da tendência espacial

Com o objetivo de estimar a tendência espacial, criou-se um modelo de regressão linear. Torna-se importante destacar que antes de iniciar esse processo foram removidos os restantes metais pesados do modelo visto que estes são extremamente correlacinados e queremos evitar multicolinearidade. Inicialmente, todos os restantes atributos foram utilizados.

É também importante frisar que, para mais à frente realizar podermos realizar previsões, dividimos o nosso dataset em conjunto de treino e teste, tendo mantido 151 observações para treino e as restantes 4 para teste.

```
Call:
lm(formula = train$zinc ~ x + y + elev + dist + om + ffreq +
    soil + lime + landuse + dist.m, data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-489.06 -120.47  -12.33   75.49  715.82

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.229e+04  1.176e+04   1.045  0.29813
x           -4.947e-02  7.214e-02  -0.686  0.49416
y           -7.608e-03  5.857e-02  -0.130  0.89686
elev        -4.587e+01  3.377e+01  -1.358  0.17686
dist         1.498e+02  5.994e+02   0.250  0.80309
om           3.509e+01  8.521e+00   4.119 6.97e-05 ***
ffreq2      -2.046e+02  7.666e+01  -2.669  0.00864 **
ffreq3      -2.025e+02  1.024e+02  -1.977  0.05029 .
soil2        3.481e+01  6.574e+01   0.529  0.59747
soil3        9.120e+01  1.026e+02   0.889  0.37576
lime1        1.362e+02  6.085e+01   2.237  0.02708 *
landuseAb   -1.995e+02  1.826e+02  -1.092  0.27678
landuseAg   -2.380e+02  1.946e+02  -1.223  0.22371
landuseAh   -1.181e+02  1.650e+02  -0.716  0.47532
landuseAm   -1.072e+02  1.670e+02  -0.642  0.52195
landuseB    -1.211e+02  2.114e+02  -0.573  0.56789
landuseBw    3.904e+01  2.012e+02   0.194  0.84644
landuseDEN  -7.051e+01  2.872e+02  -0.245  0.80649
landuseFh   -2.777e+02  2.745e+02  -1.012  0.31363
landuseFw   -1.064e+02  1.783e+02  -0.597  0.55186
landuseGa   -3.106e+02  2.283e+02  -1.360  0.17628
landuseSPO  -2.232e+02  2.806e+02  -0.795  0.42793
landuseSTA  -2.114e+02  2.313e+02  -0.914  0.36255
landuseTv   -4.127e+02  2.737e+02  -1.508  0.13419
landuseW    -8.132e+01  1.684e+02  -0.483  0.63011
dist.m      -5.396e-01  5.054e-01  -1.068  0.28777
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 217.5 on 122 degrees of freedom
  (3 observations deleted due to missingness)
Multiple R-squared:  0.7123,    Adjusted R-squared:  0.6533
F-statistic: 12.08 on 25 and 122 DF,  p-value: < 2.2e-16
```

Figura 2.8: Modelo de regressão linear.

Estes foram os valores obtidos usando o comando *summary* do R sobre o nosso modelo.

Diversas das covariáveis deste modelo inicial eram não significativas, como tal, fomos removendo uma a uma as variáveis que tinham maiores p-values até ter apenas variáveis explicativas significativas.

Após remoção dos atributos "y", "*landuse*", "*dist*"e "*soil*"ficamos com o modelo de regressão linear final cujo resultado do comando "*summary*"do R está apresentado abaixo.

```
Call:
lm(formula = train$zinc ~ x + elev + om + ffreq + lime + dist.m,
    data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-468.13 -134.12  -10.39   81.55  734.19

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.179e+04  6.359e+03   1.853  0.06592 .
x           -6.138e-02  3.599e-02  -1.705  0.09030 .
elev        -5.036e+01  2.739e+01  -1.839  0.06805 .
om           3.889e+01  6.981e+00   5.571 1.24e-07 ***
ffreq2      -1.795e+02  6.316e+01  -2.843  0.00514 **
ffreq3      -1.611e+02  7.480e+01  -2.154  0.03292 *
lime1        1.258e+02  5.446e+01   2.309  0.02237 *
dist.m      -3.301e-01  1.047e-01  -3.152  0.00198 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 212.2 on 141 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:  0.6888,    Adjusted R-squared:  0.6733
F-statistic: 44.58 on 7 and 141 DF,  p-value: < 2.2e-16
```

Figura 2.9: Modelo de regressão linear final.

Como se pode observar, todos os betas, se considerarmos um intervalo de confiança de 90%, são estatisticamente significativos. Convém destacar que sendo a variável x significativa mostra-se que de facto há tendência/ o processo não é estacionário na média, tal como tínhamos presumido. Neste modelo final ficamos então com as variáveis "x", "*elev*", "*om*", "*ffreq*", "*lime*"e "*dist.m*", sendo que a tendência pode então ser representada pela seguinte fórmula:

$$\text{Y(x) = concentração de zinco na localização x}$$

$\hat{\mu}(\text{x}) = 1.179\text{x}10^4 - 0.06138 * abcissa(x) - 50.36 * elevacao(x) + 38.89 * materiaorganica(x) - 179.5 * (sefreq(x) = 2) - 161.1 * (sefreq(x) = 3) + 125.8 * (selime(x) = 1) - 0.3301 * dist.m(x)$

Além disso, foi também usada a função "*vif*"do R sobre o nosso modelo para calcular o Fator de Inflação da Variância (Variance Inflation Factor), que é uma métrica para avaliar a multicolinearidade entre os preditores num modelo de regressão. Os resultados obtidos foram os seguintes.

```
            GVIF Df GVIF^(1/(2*Df))
x       2.376765  1        1.541676
elev    2.776409  1        1.666256
om      1.863646  1        1.365154
ffreq   3.270927  2        1.344831
lime    1.987343  1        1.409732
dist.m  1.876999  1        1.370036
```

Figura 2.10: Avaliar multicolinearidade entre os preditores usando a função *vif*.

Visto que todos os valores são inferiores a 10, não existe multicolinearidade entre os preditores (algo que já procuramos evitar anteriormente ao retirar os restantes metais pesados).

2.6 Estimação dos variogramas empírico e teórico

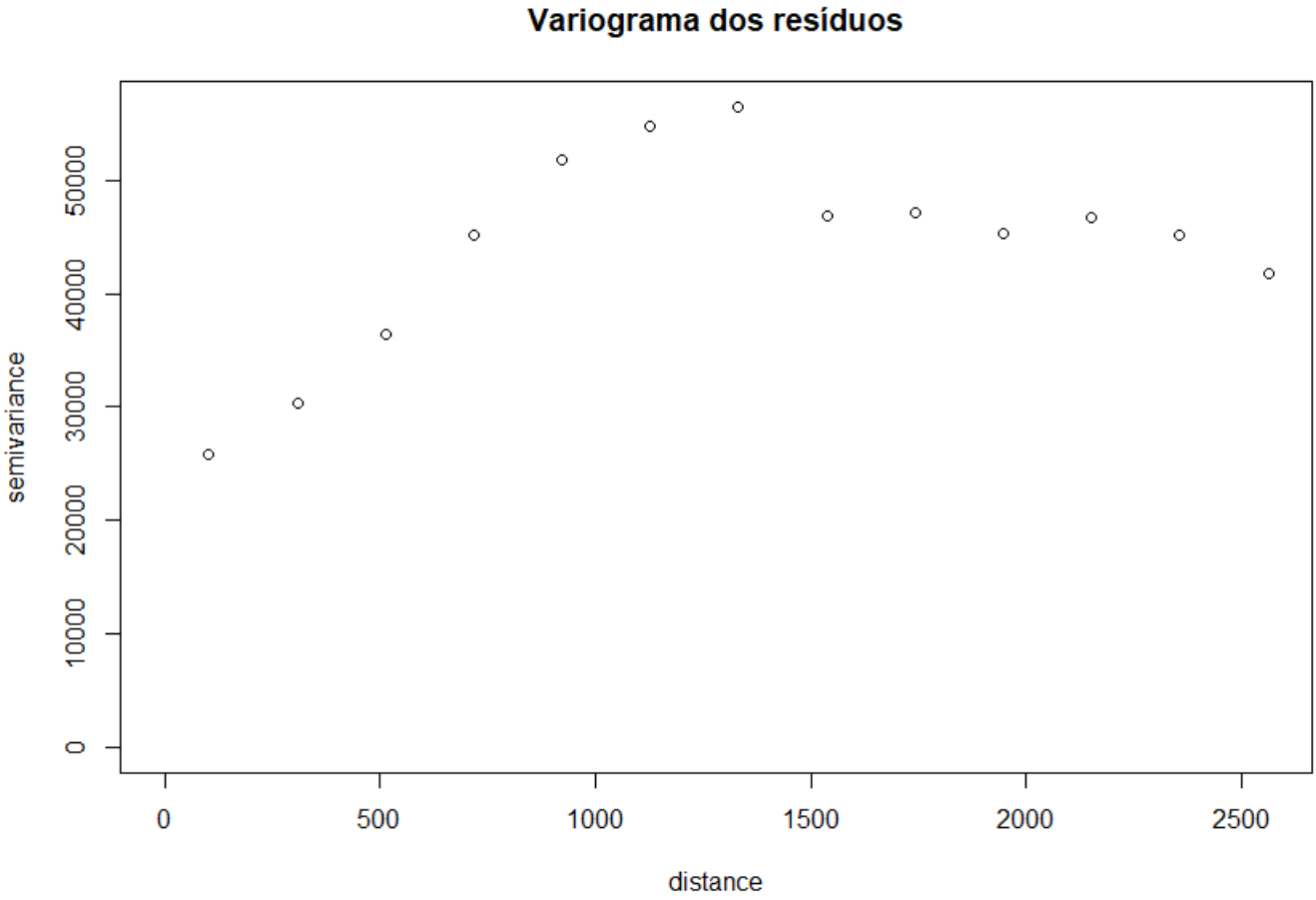


Figura 2.11: Variograma empírico estimado.

```
v$n
[1] 301 771 941 995 999 910 784 776 726 616 550 472 447
```

Figura 2.12: Número de pontos usados para cálculo de cada *bin* variograma empírico.

Na imagem acima é possível observar o variograma empírico estimado e ainda o resultado do comando $n sobre o mesmo que nos permite saber o número de pontos utilizados para o cálculo de cada *bin* do variograma onde se constata que em todos ultrapassa o mínimo de 30 pontos para cálculo.

De seguida, ajustaram-se diferentes modelos teóricos, nomeadamente, o modelo Exponencial, Esférico, Gaussiano e Matérn com kappa=2 usando o método dos mínimos quadrados e da máxima verossimilhança obtendo o seguinte resultado:

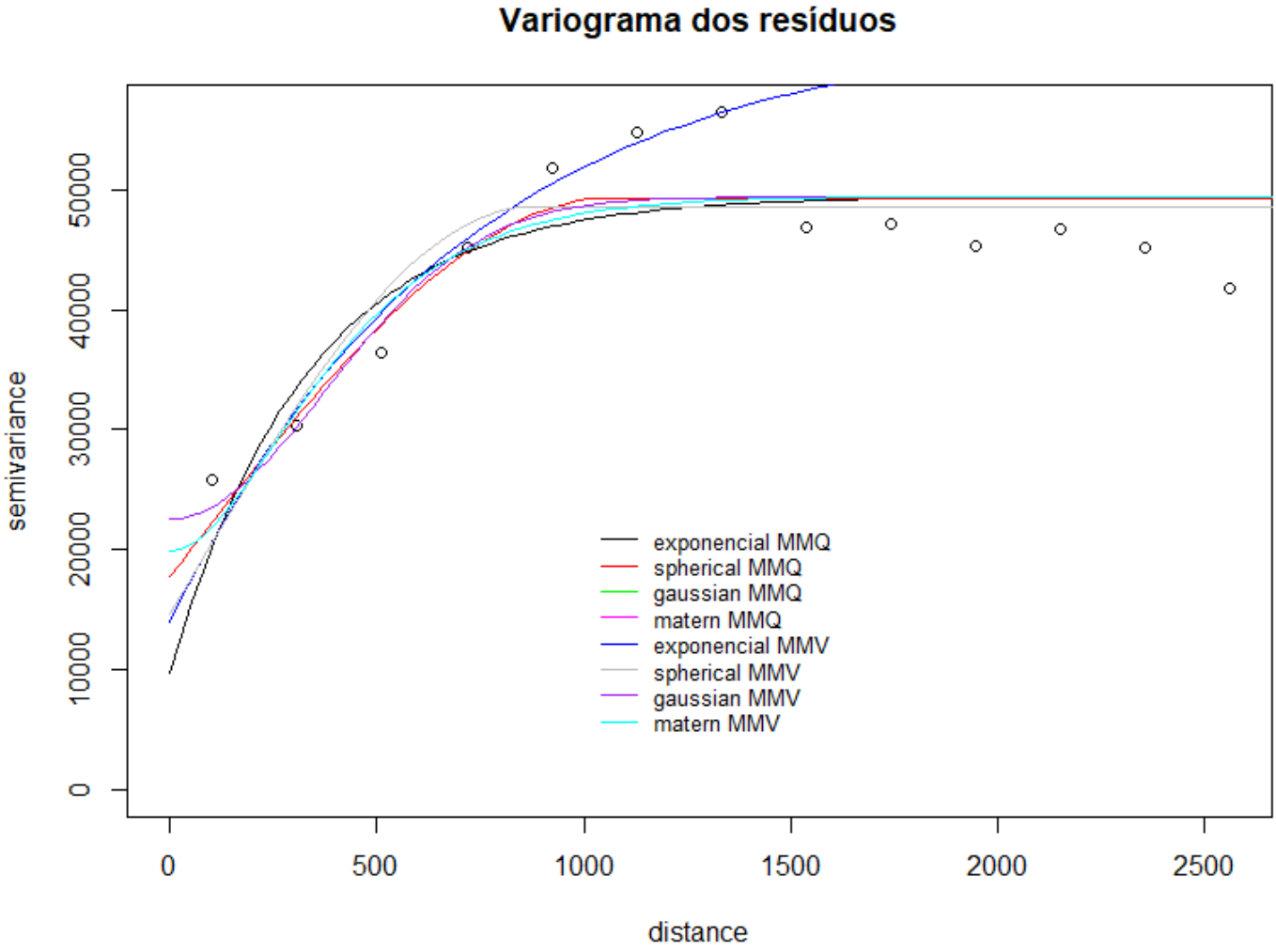


Figura 2.13: Diferentes modelos teóricos.

2.7 Validação-cruzada

Para determinar o modelo mais adequado, efetuou-se validação cruzada para cada um tendo se obtido os valores do erro médio, média dos erros padronizados, desvio padrão dos erros padronizados e média do erro quadrado, que estão todos condensados na tabela abaixo.

Modelos	ErroMédio	MédiaErrosPadronizados	DesvioPadrãoErrosPadronizados	MédiaErroQuadrado
MQ EXP	2.0696	0.0060	1.0252	1.0441
MQ SPH	**1.4065**	**0.0041**	**1.0090**	**1.0113**
MQ GAU	1.4228	0.0042	1.0117	1.0166
MQ MAT	1.6598	0.0048	1.0161	1.0256
MV EXP	1.6755	0.0050	1.0333	1.0605
MV SPH	1.7623	0.0052	1.0320	1.0579
MV GAU	1.5340	0.0045	1.0319	1.0577
MV MAT	1.8197	0.0053	1.0320	1.0578

Tabela 2.1: Valores da validação cruzada dos diferentes modelos.

Nota: Na nomenclatura utilizada acima para os modelos, inicialmente MQ significa que foi utilizado o método dos mínimos quadrados e MV significa que foi utilizado o método da máxima verossimilhança. De

seguida, EXP, SPH, GAU, MAT é referente ao modelo Exponencial, Esférico, Gaussiano e Matérn utilizando kappa=2, respetivamente.

O modelo que obteve valores para o erro médio e média dos erros padronizados mais próximos de zero e os valores para o desvio padrão dos erros padronizado e média dos erros quadrados mais próximos de 1, que é o ideal, foi o modelo de correlação esférica obtido usando o método dos mínimos quadrados, como podemos ver a negrito na tabela, tendo sido este o modelo selecionado por nós para a realização da interpolação espacial.

2.8 Interpolação Espacial

Por último, restava prever o nível da concentração de zinco em localizações não observadas, neste caso, nos nossos dados de teste, compostos por 4 localizações, através de kriging com tendência externa, assumindo o modelo para a tendência e o modelo para a correlação definidos nos capítulos 2.5 e 2.7, respetivamente. Os resultados obtidos encontram-se na tabela abaixo, juntamente com os valores reais e o desvio padrão das previsões. Para além disso, na Figura 2.14 é possível vizualizar geograficamente os 4 locais sobre os quais realizamos as previsões.

Concentração real de zinco nos novos locais	Previsão	Desvio padrão das previsões
248.56	282	165.79
896.76	801	170.53
383.19	342	169.39
726.06	593	160.25

Tabela 2.2: Previsões da concentração de zinco em locais não observados

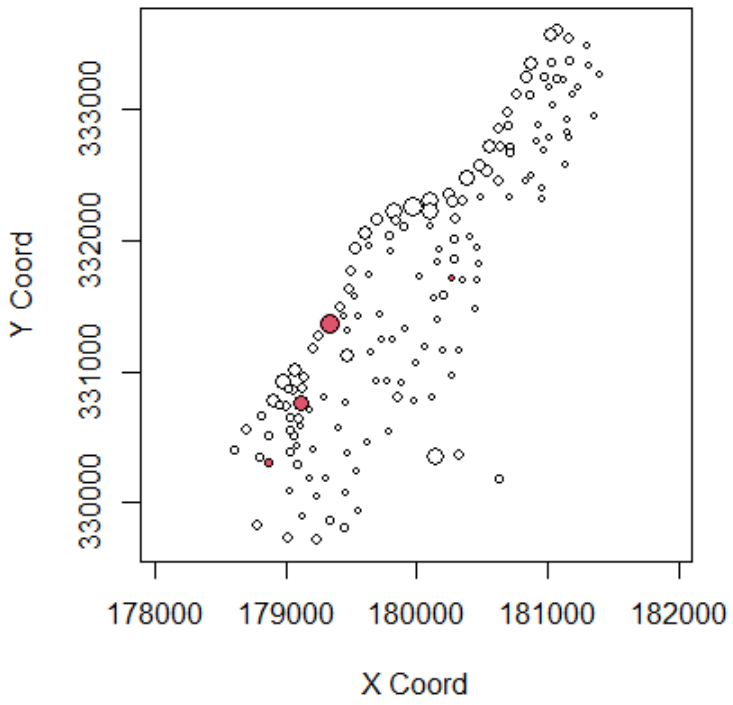


Figura 2.14: Localizações onde foi realizada a previsão.

Capítulo 3

Base de dados *World* (dados agregados por área)

3.1 Constituição da base de dados

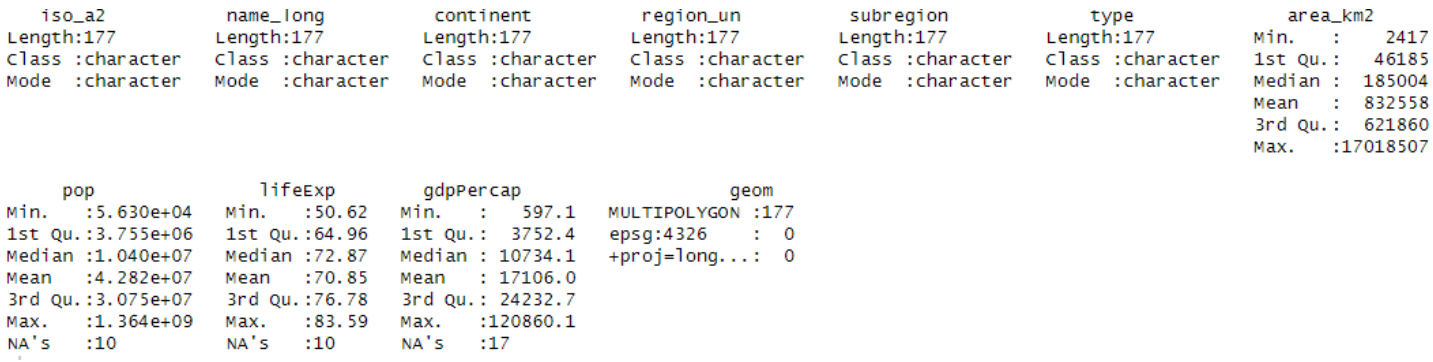
A base de dados *World* é um objeto sf no R que contém dados de um mapa mundial da Natural Earth, complementado com algumas variáveis provenientes do Banco Mundial. Este tipo de objeto sf (simple features) é comumente utilizado em análises geoespaciais em R, permitindo a manipulação e visualização de dados espaciais de forma eficiente. Além de conter informações geográficas como coordenadas e fronteiras dos países, o objeto inclui variáveis adicionais relacionadas a dados socioeconómicos e demográficos extraídos do Banco Mundial. O *dataset* contém 177 observações e 11 variáveis.

A base de dados é composta pelos seguintes atributos:

- ***iso_a2***: vetor de caracteres com os códigos ISO de 2 caracteres dos países.
- ***name_long***: vetor de caracteres com os nomes dos países.
- ***continent***: vetor de caracteres com os nomes dos continentes.
- ***region_un***: vetor de caracteres com os nomes das regiões.
- ***subregion***: vetor de caracteres com os nomes das sub-regiões.
- ***type***: vetor de caracteres com os tipos de entidades geográficas.
- ***area_km2***: vetor inteiro com os valores das áreas em quilómetros quadrados.
- ***pop***: vetor inteiro com a população em 2014.
- ***lifeExp***: vetor inteiro com a expetativa de vida ao nascer em 2014.
- ***gdpPercap***: vetor inteiro com o PIB per capita em 2014.
- ***geom***: sfc_MULTIPOLYGON.

3.2 Análise estatística dos atributos

A imagem abaixo apresenta um resumo estatístico detalhado das variáveis utilizadas no estudo. Cada variável é descrita em termos das suas estatísticas descritivas básicas, incluindo o valor mínimo, o primeiro quartil, a mediana, a média, o terceiro quartil e o valor máximo para os atributos numéricos e a distribuição de frequência para os atributos categóricos. Essas medidas permitem compreender a distribuição dos dados e identificar possíveis padrões ou *outliers*. Este resumo estatístico foi obtido com o comando *summary* do R.



```
   iso_a2           name_long          continent          region_un          subregion            type             area_km2
Length:177         Length:177         Length:177         Length:177         Length:177         Length:177         Min.   :    2417
Class :character   Class :character   Class :character   Class :character   Class :character   Class :character   1st Qu.:   46185
Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Median :  185004
                                                                                                                 Mean   :  832558
                                                                                                                 3rd Qu.:  621860
                                                                                                                 Max.   :17018507

      pop                lifeExp         gdpPercap                  geom
Min.   :5.630e+04   Min.   :50.62   Min.   :   597.1   MULTIPOLYGON :177
1st Qu.:3.755e+06   1st Qu.:64.96   1st Qu.:  3752.4   epsg:4326    :  0
Median :1.040e+07   Median :72.87   Median : 10734.1   +proj=long...:  0
Mean   :4.282e+07   Mean   :70.85   Mean   : 17106.0
3rd Qu.:3.075e+07   3rd Qu.:76.78   3rd Qu.: 24232.7
Max.   :1.364e+09   Max.   :83.59   Max.   :120860.1
NA's   :10          NA's   :10      NA's   :17
```

Figura 3.1: Análise estatística dos atributos usando o comando *summary* do R.

Algumas conclusões que se podem tirar, por exemplo nas variáveis numéricas, existe uma grande discrepância nos valores de *gdpPercap* (PIB per capita) podendo variar entre o mínimo de 597.1 e o máximo de 120860.1. Além disso, a população também difere bastante de região para região registando um mínimo de 5.63×10^4 e um máximo de 1.364×10^9.

Por escolha própria, para a realizão deste estudo, decidimos selecionar a variável ***lifeExp*** como nossa variável de interesse, que, como visto anteriormente, representa a esperança média de vida, em anos. Os valores desta variável variam entre o mínimo de 50.62 anos e o máximo de 83.59 anos apresentando uma média de 70.85 anos e mediana de 72.87 anos. Conclui-se também através dos valores do primeiro e terceiro quartil que 75% das observações encontram-se acima de 64.96 anos e 75% dos valores encontram-se abaixo de 76.78 anos, respetivamente.

Com vista a estudar com mais detalhe a nossa variável de interesse, analisou-se o histograma da mesma, representado de seguida.

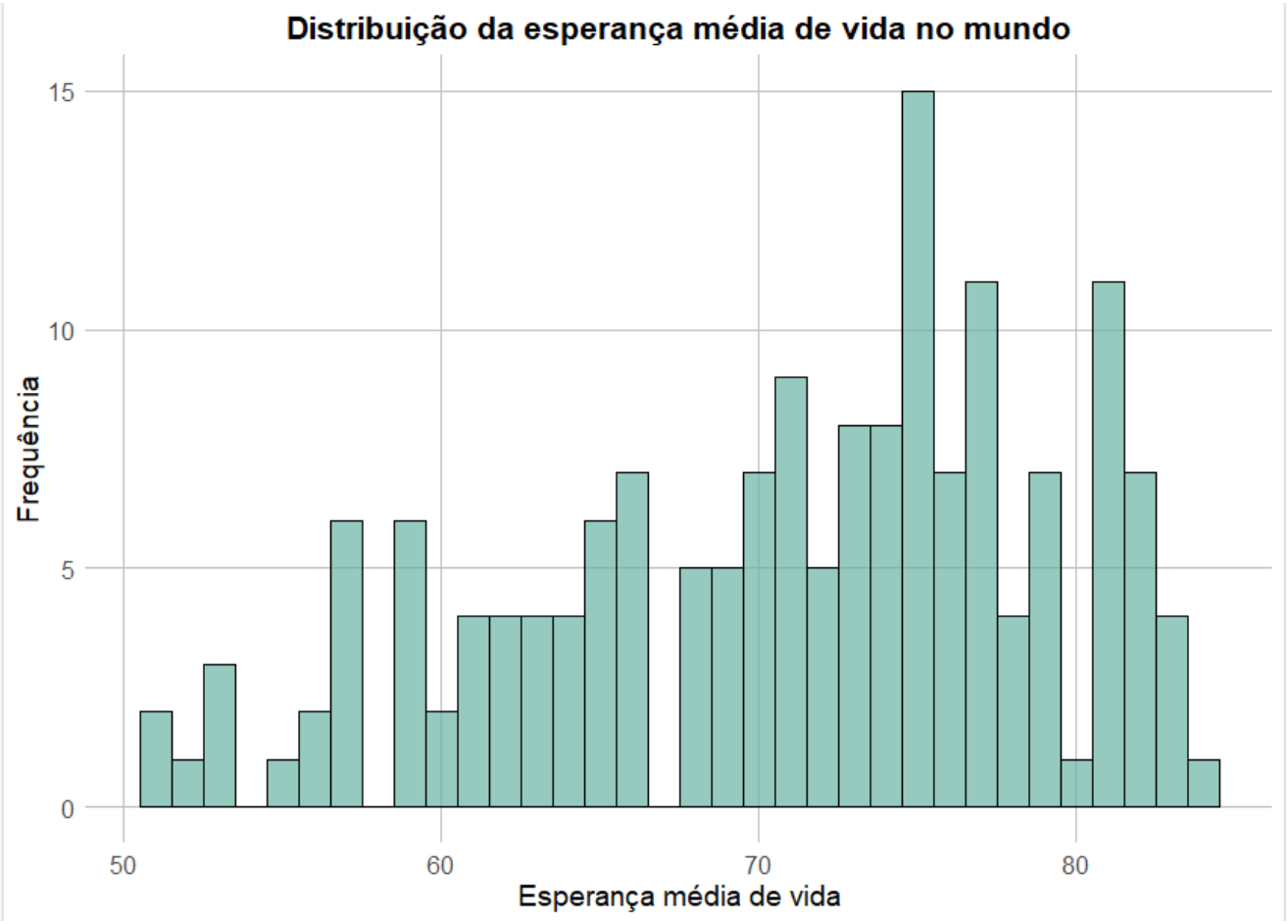


Figura 3.2: Histograma da variável *lifeExp*.

Como podemos verificar, o valor da expetativa de vida com maior frequência são os 75 anos e parece existir uma tendência para haver uma maior frequência de países com uma expetativa de vida alta (> 70 anos).

3.3 Correlação entre variáveis

Com o objetivo de estudar a correlação entre as variáveis numéricas, obteu-se a matriz de correlação das mesmas como se mostra abaixo.

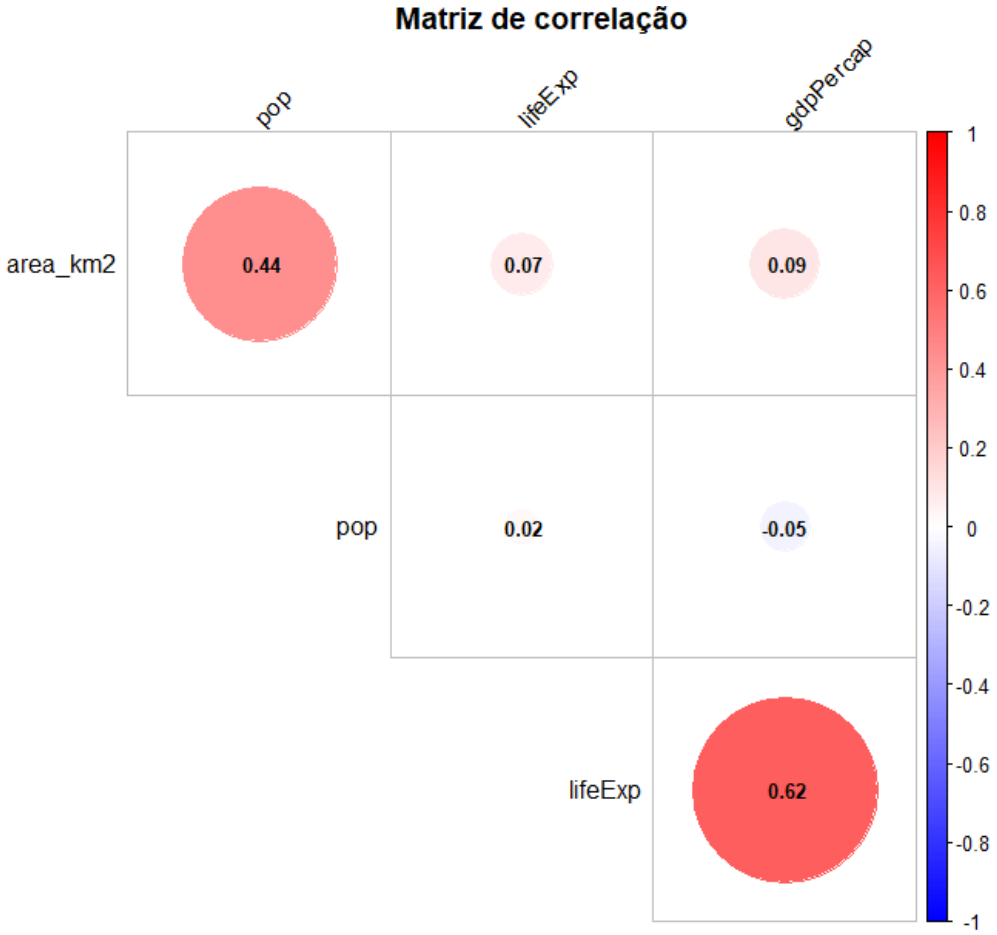


Figura 3.3: Matriz de correlação entre variáveis.

Como se pode verificar na matriz, a população está positivamente correlacionada com a área em km^2, o que parece fazer sentido sendo que se uma região tem uma maior área faz sentido que a população seja maior. Além disso, a nossa variável de interesse não está significativamente correlacionada com a população nem com a área. No entanto, a expetativa de vida está positivamente correlacionada com o PIB per capita, o que indicia que países mais ricos/desenvolvidos apresentam uma maior expetativa de vida ao nascer.

Posteriormente, categorizou-se os valores do atributo "gdpPercap"(PIB per capita) em quatro categorias baseadas nos quartis (25%, 50%, 75%) dos dados de PIB per capita. Neste caso ficou-se com quatro categorias denominadas:

- **GDP Baixo** - quartil mais baixo (0% a 25%).
- **GDP Médio-Baixo** - segundo quartil (25% a 50%).
- **GDP Médio-Alto** - terceiro quartil (50% a 75%).
- **GDP Alto** - quartil mais alto (75% a 100%).

Com base nesta última transformação, realizou-se um *boxplot* que relaciona estas quatro categorias com a nossa variável de interesse *lifeExp*, obtendo-se o seguinte resultado:

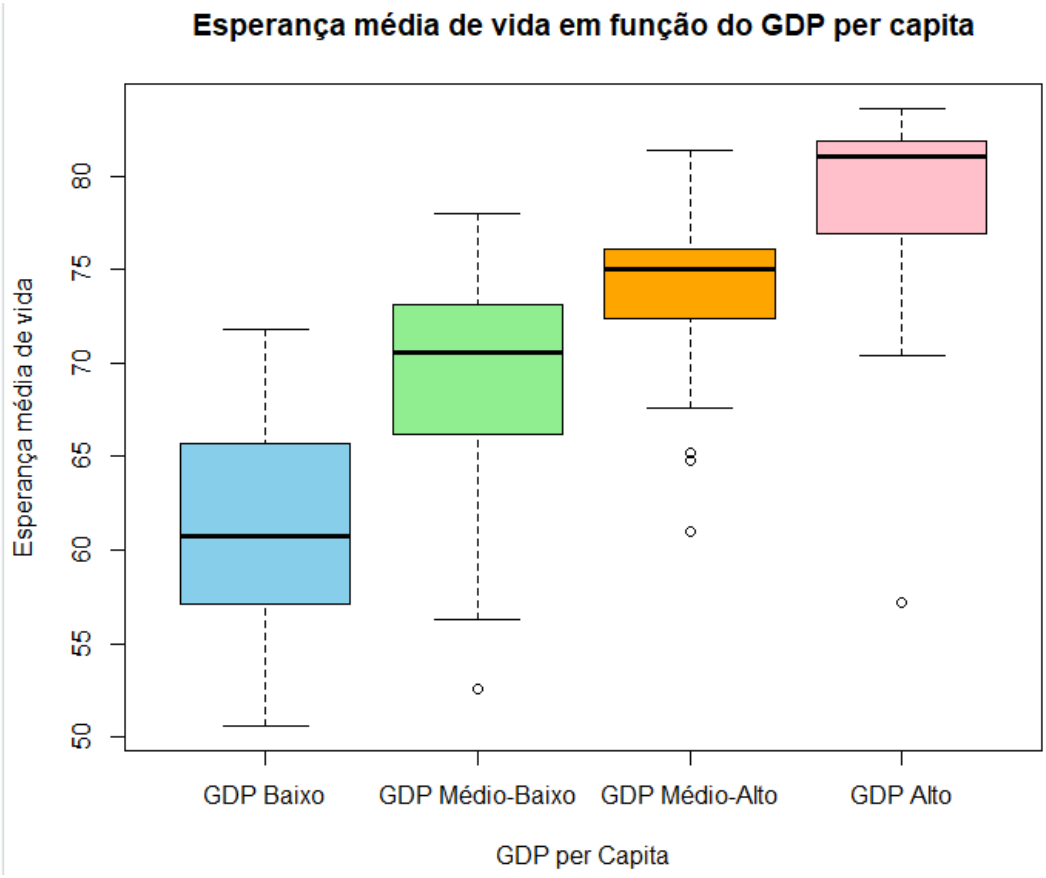


Figura 3.4: *Boxplot* que relaciona PIB per capita com expetativa de vida.

Analisando mais especificamente este *boxplot*, a ideia evidenciada da existência de correlação entre o PIB per capita e a expetativa de vida torna-se ainda mais visível e é possível observar um aumento nessa expetativa de vida entre cada categoria GDP.

3.4 Análise de variáveis agregadas por continente

Posteriormente, analisou-se novamente as variáveis, mas desta feita, com separação por continente tendo se obtido o seguinte gráfico:

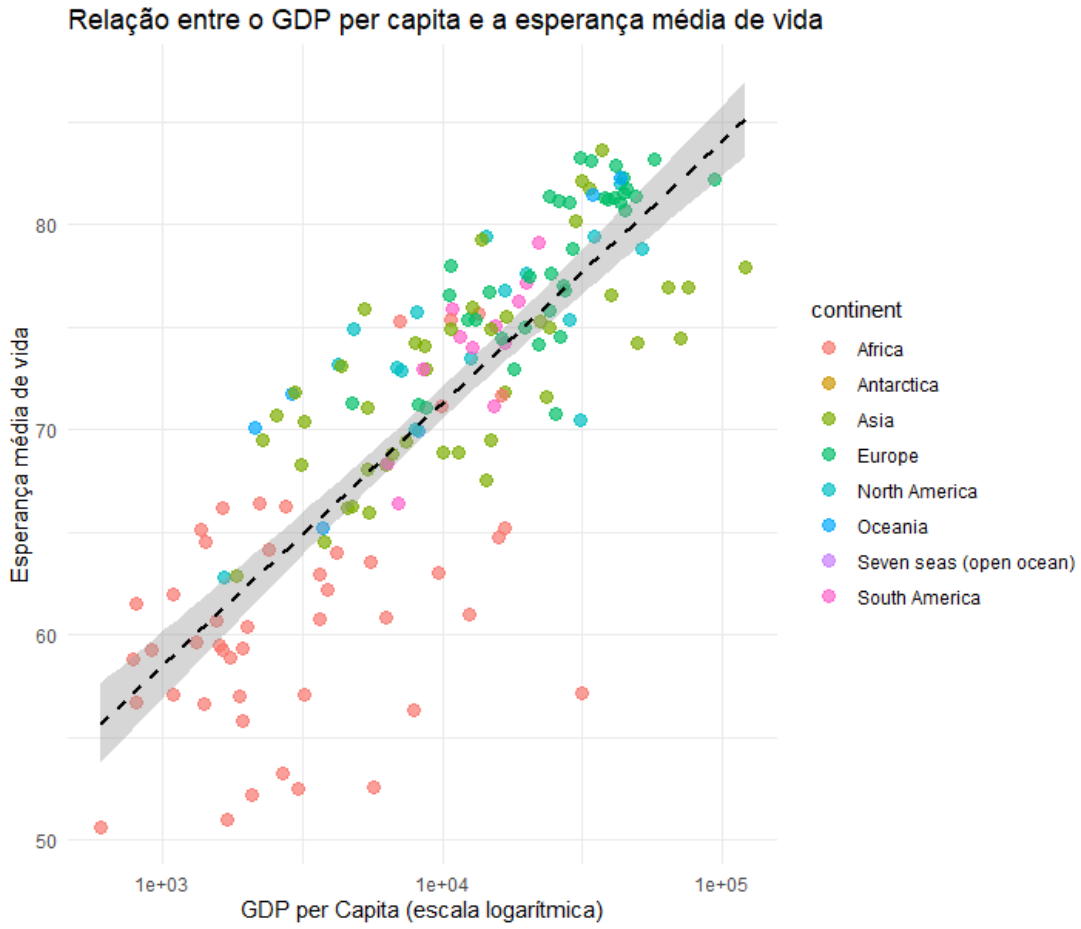


Figura 3.5: Relação entre o GDP per capita e a esperança média de vida.

Como é visível, África é o continente com GDP per capita mais baixo e consequentemente apresenta uma menor expetativa de vida. Por outro lado, continentes como Europa, Ásia e América do Norte que apresentam um GDP per capita superior apresentam também uma expetativa de vida ao nascer superior.

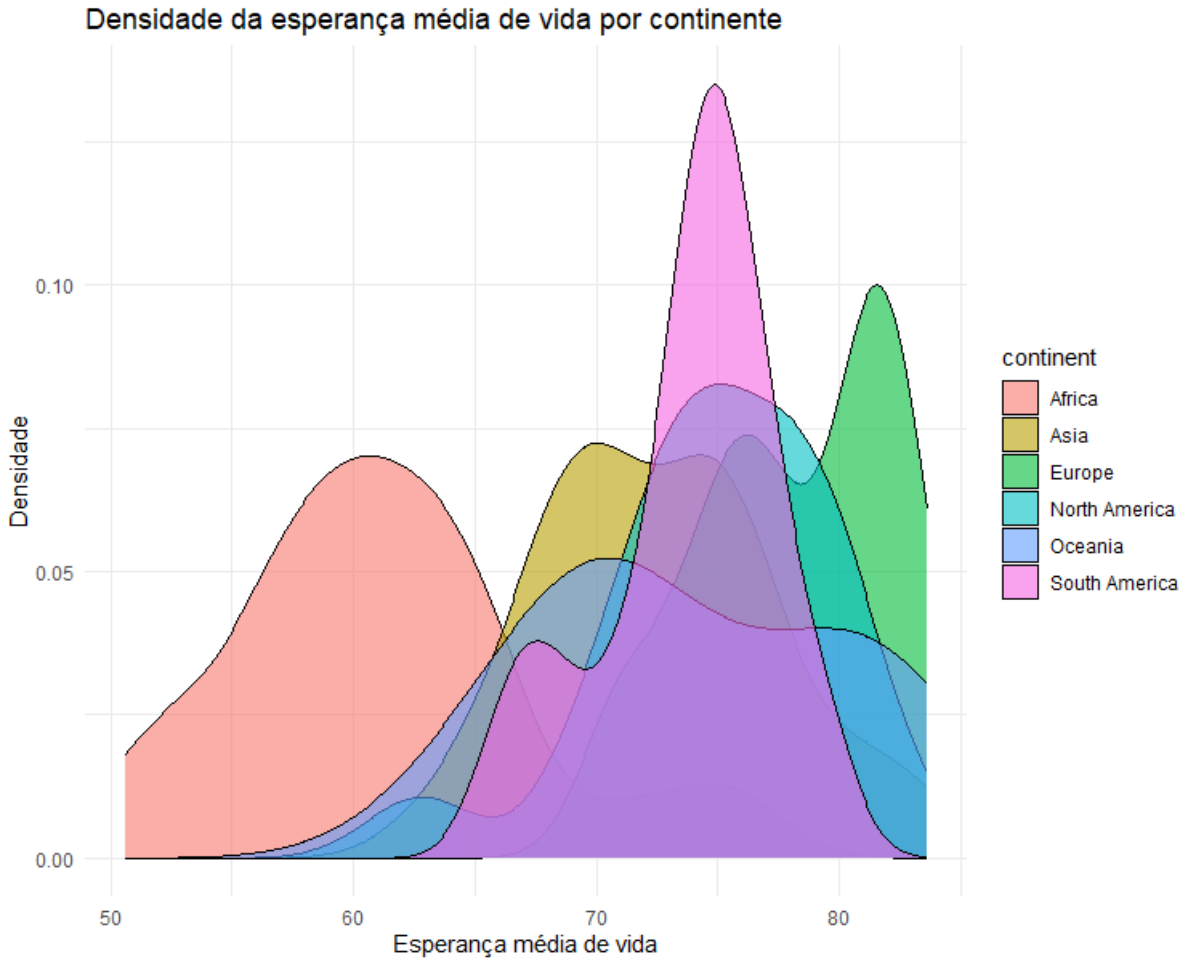


Figura 3.6: Densidade da esperança média de vida por continente.

Analisando também este gráfico da densidade da esperança média de vida por continente, os mesmos cenários acontecem onde África continua a ser o continente com menores esperanças médias de vida. Além disso, algum destaque para a Europa que é o continente que tem as esperanças médias de vida mais altas e também para a América do Sul que tem uma grande densidade de esperanças de vida entre 70 e 80 anos.

Já este gráfico exibe a densidade da experança de vida por continente. A Europa apresenta a maior expectativa de vida, com um pico acentuado próximo dos 80 anos, indicando baixa variabilidade. A África apresenta a menor expectativa de vida, com uma distribuição ampla centrada abaixo dos 60 anos. A Ásia, a América do Norte e a América do Sul apresentam variabilidade moderada, com expectativas de vida variando de 60 a 80 anos. A Oceania é o continente que apresenta maior variação na expectativa de vida. Ou seja, a esperança média de vida varia significativamente entre diferentes países ou regiões refletindo desigualdades ou disparidades na saúde e na qualidade de vida.

3.5 Análise do continente Ásia

Para o resto do nosso projeto, decidimos focar apenas num continente, a Ásia. O resultado do *summary* dos dados filtrados pelo continente Ásia, foram os seguintes:

```
   iso_a2           name_long          continent          region_un          subregion             type
Length:47          Length:47          Length:47          Length:47          Length:47          Length:47
Class :character   Class :character   Class :character   Class :character   Class :character   Class :character
Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character


      pop               lifeExp         gdpPercap                 geom          gdpPercapCategory
Min.   :4.117e+05   Min.   :62.90   Min.   :  1839   MULTIPOLYGON :47   GDP Baixo      : 6
1st Qu.:5.466e+06   1st Qu.:68.86   1st Qu.:  5318   epsg:4326    : 0   GDP Médio-Baixo:16
Median :1.920e+07   Median :71.80   Median : 10650   +proj=long...: 0   GDP Médio-Alto :11
Mean   :9.581e+07   Mean   :72.59   Mean   : 20026                      GDP Alto       :10
3rd Qu.:5.192e+07   3rd Qu.:75.47   3rd Qu.: 23891                      NA's           : 4
Max.   :1.364e+09   Max.   :83.59   Max.   :120860
NA's   :2           NA's   :2       NA's   :4
```

Figura 3.7: Resultado do comando *summary* sobre os dados do continente Ásia.

Nesta análise, é de destacar que a maior parte dos países tem um PIB per capita de Médio-Baixo para cima e uma esperança média de vida que varia entre o mínimo de 62.90 anos e o máximo de 83.59 anos. Além disso, a média situa-se em 72.59 anos, a mediana em 71.80 anos e os valores do primeiro e terceiro quartil em 68.86 anos e 75.47 anos respetivamente, o que nos indica que 25% dos valores encontram-se abaixo de 68.86 anos e 75% dos valores encontram-se abaixo de 75.47 anos.

Com base no seguinte gráfico que nos mostra, por país, a esperança média de vida do mesmo, procurou-se analisar se existiam regiões do continente em questão onde fosse percetível uma maior ou menor esperança média de vida. Como se pode observar no gráfico abaixo, Japão e Israel, por exemplo, destacam-se como tendo uma alta expetativa de vida ao nascer. Do lado oposto, temos, por exemplo, Afeganistão e Iémen, que apresentam esperanças médias de vida baixas.

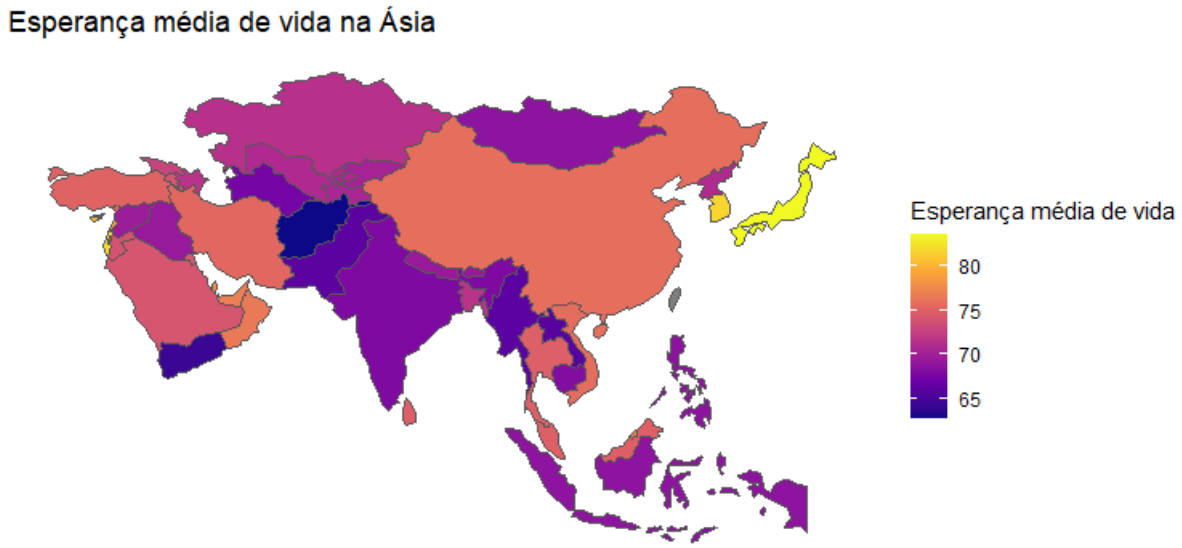


Figura 3.8: Expetativa de vida ao nascer dos vários países asiáticos.

3.6 Estatísticas I de Moran e c de Geary

De forma a compreender se há alguma auto-correlação espacial entre a esperança média de vida nos países do continente asiático foram calculadas as estatísticas I de Moran c de Geary. Para tal, começamos por calcular a matriz de vizinhanças W usando duas abordagens distintas. A primeira consiste na utilização da distância entre os centróides de cada país e a segunda consiste no uso do polígno, ou seja, as formas geométricas reais dos países, os pesos são baseados na relação espacial entre esses polígonos, que pode ser definida usando as suas fronteiras.

Abordagem	Estatística I de Moran	p-value	Estatística c de Geary	p-value
Polígono	0.15689451	0.1439	0.54883929	**0.0003173**
Centróides	0.060624926	0.3339	0.897643540	0.1228

Tabela 3.1: Estatísticas I de Moran e c de Geary para as duas abordagens implementadas.

Pela tabela podemos observar que para ambas as abordagens os p-values do teste de I de Moran são superiores a 0.05, não nos permitindo rejeitar a hipótese nula, e como tal indicando que não há correlação espacial entre as esperanças médias de vida doa países asiáticos. Os valores da estatística c de Geary foram ambos inferiores a 1 indicando correlação espacial positiva, ou seja, que os valores semelhantes para a esperança média de vida encontram-se maioritáriamente agrupados/ são vizinhos. Dito isto, o p-value deste teste na abordagem dos polígonos indica-nos que há evidência estatística para rejeitar a hipótese nula, e como tal, que existe correlação entre a esperança média de vida dos países asiáticos.

Nota: Os países que não continham informação sobre a esperança média de vida dos seus habitantes foram removidos antes do cálculo destas estatísticas.

3.7 Ajuste de diversos modelos aos dados

3.7.1 Modelo de Regressão Linear

Após toda a análise realizada anteriormente ao *dataset*, criou-se um modelo de regressão linear com vista a ajustar uma equação linear para prever a expetativa de vida com base nos preditores.

```
Call:
lm(formula = lifeExp ~ subregion + type + area_km2 + pop + gdpPercap,
    data = asia_no_na)

Residuals:
   Min     1Q Median     3Q    Max
-7.193 -2.071  0.000  1.986  6.252

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                 7.827e+01  4.138e+00  18.913  < 2e-16 ***
subregionEastern Asia       6.118e+00  2.763e+00   2.214  0.03385 *
subregionSouth-Eastern Asia -4.506e-01  2.136e+00  -0.211  0.83423
subregionSouthern Asia      -6.706e-01  2.321e+00  -0.289  0.77448
subregionWestern Asia        1.453e+00  2.137e+00   0.680  0.50126
typeDisputed                -6.959e+00  5.208e+00  -1.336  0.19057
typeSovereign country       -7.754e+00  3.507e+00  -2.211  0.03409 *
area_km2                    -1.378e-06  8.051e-07  -1.712  0.09631 .
pop                          2.380e-09  4.060e-09   0.586  0.56176
gdpPercap                    8.349e-05  2.719e-05   3.071  0.00426 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.796 on 33 degrees of freedom
Multiple R-squared:  0.5224,    Adjusted R-squared:  0.3922
F-statistic: 4.011 on 9 and 33 DF,  p-value: 0.001538
```

Figura 3.9: Modelo de regressão linear.

Estes foram os valores obtidos usando o comando *summary* do R sobre o nosso modelo. Destarcar que o AIC deste modelo foi de 247.3776.

Após remoção dos atributos "*pop*", "*type*"e "$area\_km^2$"ficamos com o modelo de regressão linear final cujo resultado do comando "*summary*"do R está apresentado abaixo.

```
Call:
lm(formula = lifeExp ~ subregion + gdpPercap, data = asia_no_na)

Residuals:
    Min      1Q  Median      3Q     Max
-7.9954 -1.9632  0.0865  1.8562  7.3454

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)               6.946e+01  1.765e+00  39.345  < 2e-16 ***
subregionEastern Asia     6.121e+00  2.645e+00   2.314  0.02630 *
subregionSouth-Eastern Asia  1.702e-01  2.143e+00   0.079  0.93714
subregionSouthern Asia   -2.183e-01  2.226e+00  -0.098  0.92241
subregionWestern Asia     2.753e+00  2.087e+00   1.319  0.19532
gdpPercap                 8.197e-05  2.727e-05   3.006  0.00473 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.902 on 37 degrees of freedom
Multiple R-squared:  0.4344,    Adjusted R-squared:  0.3579
F-statistic: 5.683 on 5 and 37 DF,  p-value: 0.0005506
```

Figura 3.10: Modelo de regressão linear final.

O nosso modelo linear final é então dado pela seguinte fórmula:

$lifeExp_i$(x) = 69.46 + 6.121*EasternAsia(x) + 0.1702*SouthEasternAsia(x) - 0.2183*SouthernAsia(x) + 2.753*WesternAsia(x) + 8.197 $\times$ 10$^-$5*gdpPercap$_i$ + ε_i

Destacar que o AIC deste modelo foi de 246.6515.

3.7.2 Modelos SAR, SMA e CAR

De seguida utilizamos três tipos de modelos espaciais: modelo SAR, modelo SMA e modelo CAR. Estes modelos são fundamentais para capturar e analisar a dependência espacial existente nos dados, permitindo um melhor entendimento da relação entre as variáveis. O principal objetivo desta análise é comparar o desempenho dos modelos SAR, SMA e CAR, avaliando como cada um captura a estrutura espacial dos dados. Além disso, verificamos a presença de autocorrelação espacial nos resíduos dos modelos utilizando o teste de Moran.

Os resultados destes três modelos juntamente com o modelo linear estão condensados na tabela abaixo.

Modelos	Erro Padrão	AIC	Resíduos Independentes (?)
Modelo Linear	3.902	**246.6515**	sim (p-value = 0.1893)
Modelo SAR	**3.5986**	248.32	sim (p-value = 0.6614) mas λ não significativo
Modelo SMA	3.6022	248.5	sim (p-value = 0.7786) mas λ não significativo
Modelo CAR	3.6193	248.65	sim (p-value = 0.9324) mas λ não significativo

Tabela 3.2: Comparação dos diferentes modelos.

Destacar na tabela acima que o modelo linear teve o menor valor de AIC e o modelo SAR obteve o menor valor de erro padrão. Em todos os quatro modelos, os p-values obtidos no teste I de Moran foram sempre

superiores a 0.05 e inclusive a 0.1. Logo, há evidência estatística para concluir que, em todos os modelos, os resíduos são independentes ou espacialmente não-correlacionados, como pretendido.

Capítulo 4

Conclusão

Neste trabalho tivemos como objetivo analisar e realizar predição espacial em dois *datasets*: Meuse River e World. Através da aplicação de técnicas de estatística espacial, foram identificadas relações significativas entre as variáveis e padrões espaciais nos dados.

Para o conjunto de dados Meuse River, a análise exploratória revelou uma forte correlação entre as concentrações de diferentes metais pesados e uma relação negativa entre a concentração de zinco e a elevação. A modelação geoestatística indicou a presença de dependência espacial nos dados, e o modelo de correlação esférica foi selecionado como o mais adequado para descrever a variabilidade espacial do zinco.

No conjunto de dados World, a análise exploratória revelou uma forte correlação entre o PIB per capita e a expetativa de vida. A modelação espacial, através dos modelos SAR, SMA e CAR, permitiu aprofundar a análise ao considerar a estrutura espacial dos dados. Essa abordagem revelou a importância de levar em conta a influência espacial na modelagem da expectativa de vida, complementando os resultados obtidos pelo modelo linear.

Em suma, este trabalho permitiu-nos colocar em prática diferentes técnicas de Estatística Espacial tanto para dados geoestatísticos como para dados agredados por área. Em ambas as componentes, foi-nos permitido fazer uma análise diferenciada dos dados e realizar predição espacial através de diferentes modelos lineares e espaciais, consolidando assim de uma forma mais prática todos os conhecimentos adquiridos em sala de aula.

Bibliografia

[1] S. Banerjee, B.P. Carlin, and A.E. Gelfand. *Hierarchical Modeling and Analysis for Spatial Data*. Chapman&Hall /CRC, 2nd edition, 2014.

[2] Roger S. Bivand, Edzer J. Pebesma, and V. Gómez-Rubio. *Applied Spatial Data Analysis with R*. UseR! Series. Springer, 2nd edition, 2013.

[3] M.L Carvalho and I. Natário. *Análise de Dados Espaciais*. Sociedade Portuguesa de Estatística, 2008.

[4] N.A.C. Cressie. *Statistics for Spatial Data*. Wiley, New York, 1993.

[5] P. Diggle and P. Ribeiro. *Model-based Geostatistics*. Springer Series in Statistics. Springer, 2007.

[6] M. Sherman. *Spatial Statistics and Spatio-Temporal Data: Covariance Functions and Directional Properties*. Wiley Series in Probability and Statistics. Wiley, 2011.