

---

# Estadística Espacial

## ## Variograma: Algoritmos de Machine Learning



FACULTAD DE MATEMÁTICAS  
PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE

### Integrantes

- Michelle Vergara
- Milenka Zuvic
- Macarena Geraldo
- Francisco Ortega
- Carlos Quezada

---

## TABLA DE CONTENIDOS

<b>1. Problema 1 (Geoestadístico)</b>	<b>1</b>
1.1. Pregunta 1	1
1.2. Pregunta 2	2
1.3. Pregunta 3	4
1.4. Pregunta 4	6
1.5. Pregunta 5	11
<b>2. Problema : Datos de área</b>	<b>13</b>
2.1. Pregunta 1	14
2.2. Pregunta 2	15
2.4. Pregunta 4	21
2.5. Pregunta 5	24

## 1. Problema 1 (Geoestadístico)

Considere la base de datos `geoR::camg`; contenido de magnesio medido en muestras de la capa 0-20 cm en 178 ubicaciones dentro de una determinada zona de estudio dividida en tres subzonas. Con el objetivo de crear un mapa de la zona de interés para la variable “mg020” (contenido de magnesio), responda las siguientes preguntas:

### 1.1. Pregunta 1

- Dar un formato apropiado a la base de datos para trabajar en R a través de la librería `geoR`.

Con la siguiente línea de código, podemos cargar los datos de `camg` dentro de R:

```
1 # Taller 4
2 library(geoR)
3
4 # cargamos los datos
5 data(camg)
```

Al revisar los datos “`camg`” vienen numerosas variables, entre ellas: espaciales (coordenadas y regiones), numéricas (valores de magnesio y calcio). Además la clase de “`camg`” es un “`data.frame`”, que no es del todo adecuado para realizar el análisis geoestadístico. Para ello se debe transformar la clase del objeto con el siguiente código:

```
# dejamos solo el atributo de interes que es mg020 en formato de geodata
mg020 <- as.geodata(camg, data.col = 6)
class(mg020)
```

De esta forma, con el atributo ‘`data.col=6`’ le indicamos que la columna 6, correspondiente a los valores de “`mg020`”, corresponde al atributo que consideraremos junto a las variables espaciales (y que es lo que se pide para esta pregunta).

Con el comando “`class(mg020)`” validamos la clase de esta variable y que ahora corresponde a “`geodata`”, adecuada para usar la librería de “`geoR`”.

## 1.2. Pregunta 2

- Realice un análisis exploratorio a través de gráficas apropiadas y responda las siguientes preguntas:
  - a) ¿Existe una posible presencia de valores atípicos?
  - b) ¿Se debe remover alguna tendencia antes de analizar la dependencia espacial?
  - c) ¿Se justifica alguna transformación (como Box-Cox) en los datos?

Para responder a estas preguntas podemos ejecutar la siguiente línea para ver diversos gráficos de los datos de la variable “mg020”:

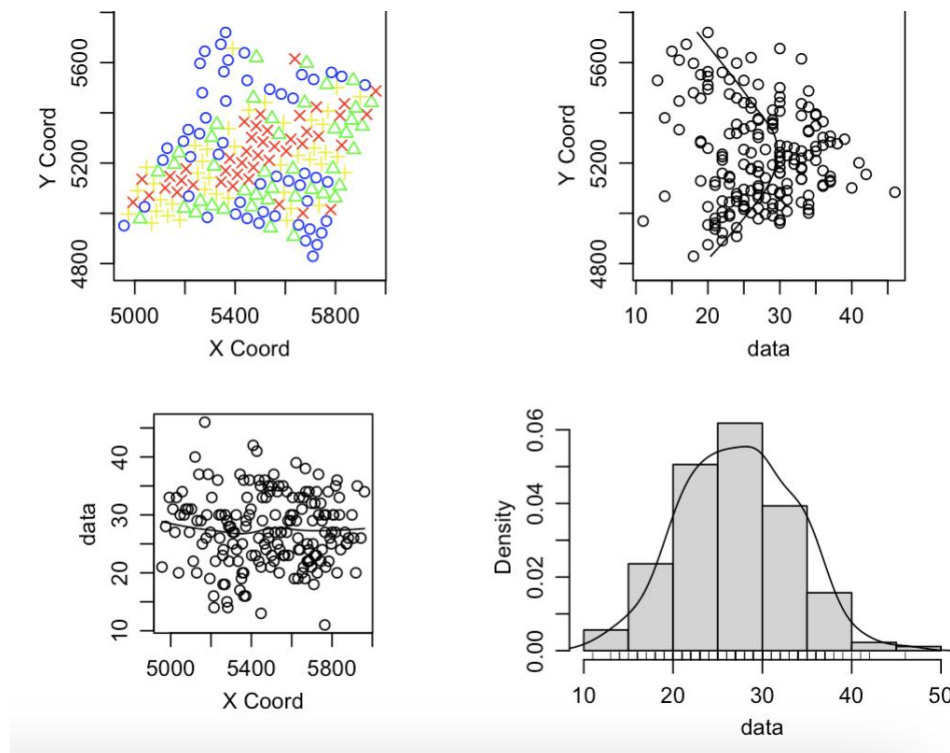


Figura 1: Gráfico del objeto geodata “mg020”

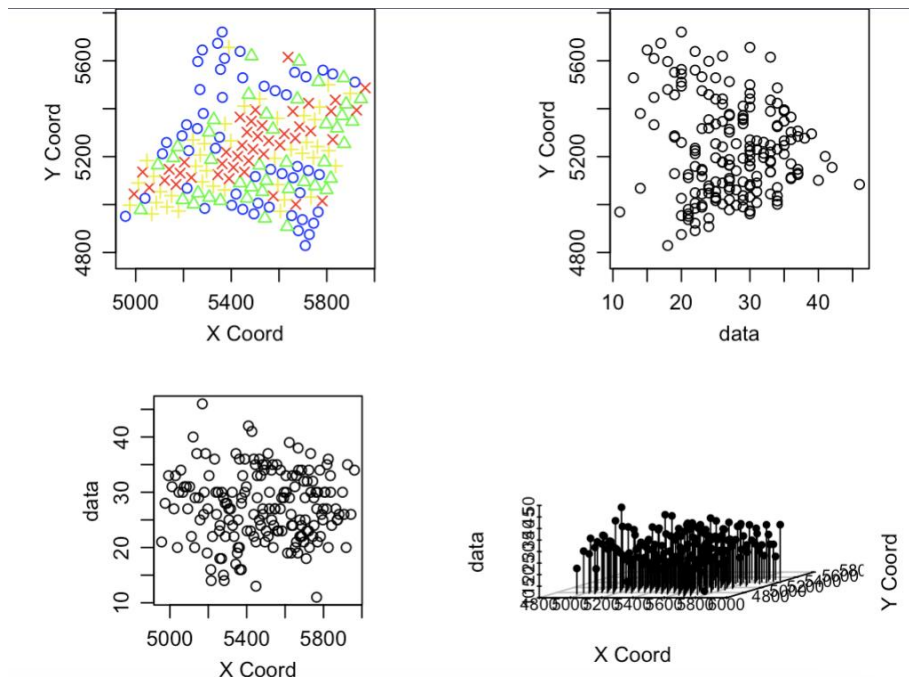


Figura 2: Gráfico del objeto geodata “mg020” con atributo “scatter3D”

Observando la figura 2, que muestra un gráfico 3D para las coordenadas y el valor de magnesio, podemos ver que a simple vista se ve únicamente un único valor atípico (ubicado en el extremo derecho y que resalta en altura respecto a los demás); el resto de los datos se mantienen como una gran sábana uniforme. Dado esto, y que si bien los valores atípicos son muy pocos respecto al total de datos, se tendrán en consideración para la elección del método del variograma.

A partir también de la figura 1, podemos ver que en el gráfico de la coordenada Y respecto al valor de magnesio, existe una suerte de tendencia con forma de “campana”, como si los datos se distribuyeran de forma simétrica a lo largo de la coordenada Y. Esta “tendencia” no es necesaria eliminarla, pero probablemente vuelva a estar presente en el variograma, y debería ser considerada para la modelación de esta misma.

Complementando esto con los otros gráficos de la figura 2, no se ve existencia de tendencias evidentes para los datos respecto a cada coordenada y a estas en conjunto. Las nubes de puntos no poseen patrones evidentes, y también se puede validar con la gráfica 3D, donde la saba tampoco posee una caída o subida muy pronunciada.

Finalmente y respecto a si es necesario aplicar una transformación de BOX-COX a los datos, esta no es necesaria dado que, al ver el histograma de la figura 1, podemos ver que los valores de magnesio siguen una distribución similar a la de una distribución normal, siendo

simétrica respecto a la medida. Por ende, no es necesario corregir un sesgo y aplicar esta transformación.

### 1.3. Pregunta 3

- Haga un análisis detallado de dependencia espacial usando el variograma como herramienta y responda las siguientes preguntas:
  - a) ¿Se justifica el uso de un modelo geoestadístico?
  - b) Proponga tres modelos de variogramas identificando valores iniciales para el nugget, el sill y el rango. Además, indique claramente el tipo de media asumida.

Antes de aplicar directamente el variograma a los datos del magnesio, y dada la información obtenida en la pregunta anterior, **al no existir una evidencia de tendencia para los datos de magnesio, supondremos que la media es constante**. A partir de esto, se obtiene el siguiente variograma:

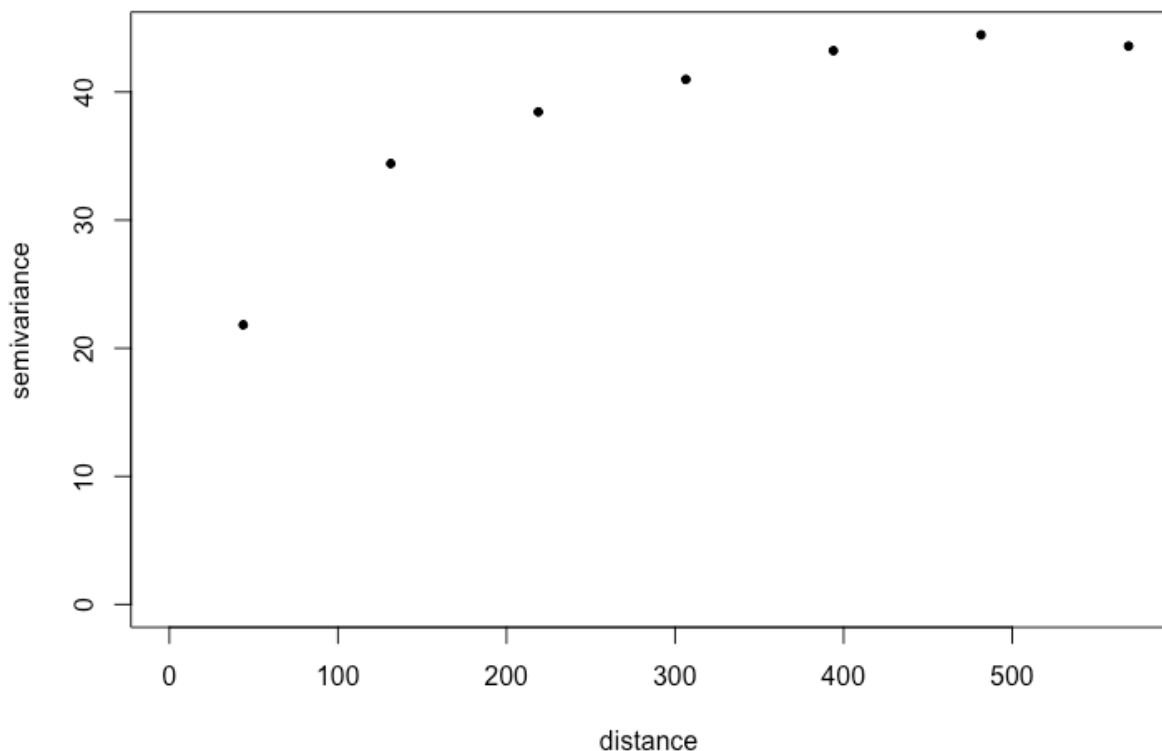
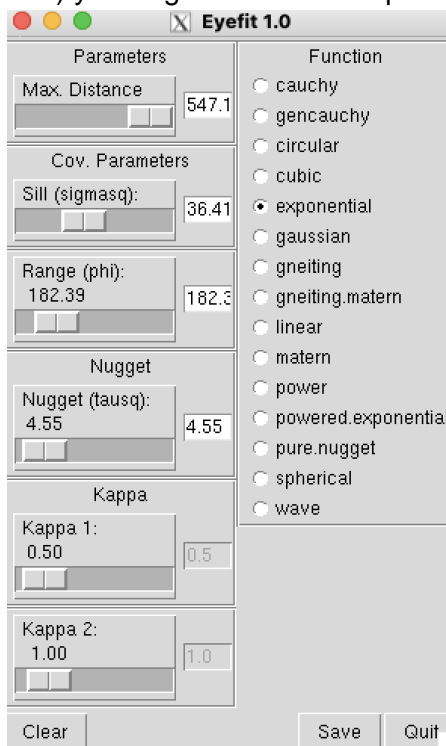


Figura 3: Variograma con media constante para los valores de magnesio. Se considera como rango máximo la mitad de la distancia máxima.

La semivarianza se estabiliza aproximadamente para aquellos puntos que están a una distancia superior a 500 unidades, por lo que podríamos considerar el silo alrededor de una semivarianza de 40.

Se puede afirmar la existencia de dependencia espacial en el rango mostrado en la figura 3; en esta zona la varianza cambia constantemente, dando cuenta de una correlación en los datos.

A partir de este variograma se hace un ajuste visual de 3 modelos propuestos: exponencial, gaussiano y circular considerando una media constante (por la falta de una tendencia evidente explicada en la pregunta anterior) y los siguientes valores para el nugget, silo y rango:



The screenshot shows the Eyefit 1.0 software window. It has a title bar with standard macOS window controls and the text 'Eyefit 1.0'. The interface is divided into two main columns. The left column contains several parameter input fields, each with a slider and a numerical value: 'Max. Distance' (547.1), 'Sill (sigmasq):' (36.41), 'Range (phi):' (182.39), 'Nugget (tausq):' (4.55), 'Kappa 1:' (0.50), and 'Kappa 2:' (1.00). The right column is titled 'Function' and contains a list of radio buttons for different mathematical models: 'cauchy', 'gencauchy', 'circular', 'cubic', 'exponential' (which is selected), 'gaussian', 'gneiting', 'gneiting.matern', 'linear', 'matern', 'power', 'powered.exponential', 'pure.nugget', 'spherical', and 'wave'. At the bottom of the window are three buttons: 'Clear', 'Save', and 'Quit'.

Figura 4: Valores iniciales para el silo, rango y nugget para ajustar el variograma empírico considerando una media constante

Luego se ajustan las respectivas funciones, obteniéndose el siguiente resultado:

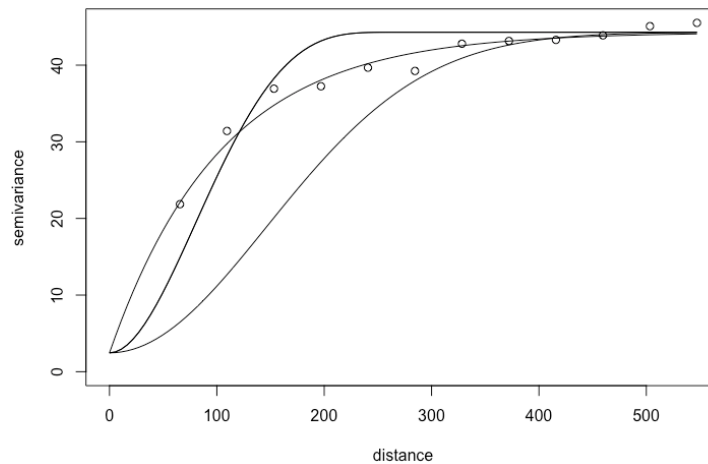


Figura 4: Ajuste visual de modelos exponencial, gaussiano y cúbico para el variograma.  
Puntos de partida para determinar los mejores parámetros de ajuste.

#### 1.4. Pregunta 4

- Ajuste los modelos de variograma teóricos propuestos en el ítem anterior, mediante máxima verosimilitud restringida. Para cada modelo de covarianza estime el modelo con media constante y media lineal. Utilice el criterio de Akaike (AIC) para seleccionar el modelo más apropiado

Continuando con los resultados expuestos en la pregunta anterior y asumiendo media constante, se emplea código para obtener los modelos, estimando los parámetros a mediante máxima verosimilitud, obteniéndose los siguientes AIC:

Ajuste	AIC
<i>Exponencial</i>	1093.4
<i>Gaussiano</i>	1125.141
<i>Cúbico</i>	1169.12



El gráfico del variograma para los 3 ajustes descritos previamente es el siguiente:

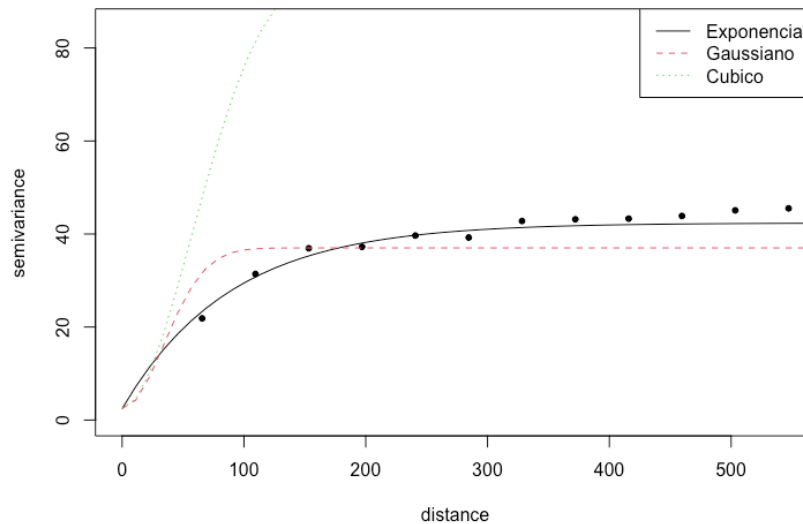


Figura 5: Variograma con 3 ajustes distintos propuestos

De acuerdo a lo anterior, el modelo con menor AIC es el de mejor ajuste en este caso corresponde al ajuste exponencial, que tiene un AIC de 1091,644; corresponde a la línea segmentada de color negro en la figura 5.

Si ahora, consideramos que la media es lineal, debemos recalcular el variograma para considerar este cambio. Esto se efectúa a partir del siguiente código:

```
# si la media no es constante, sino lineal  
vg_lineal <- variog(mg020, trend = "1st", max.dist = hmax/2)  
plot(vg_lineal, pch=20, max.dist = hmax/2)
```

Este variograma, que considera que la media es lineal se ve de la siguiente forma:

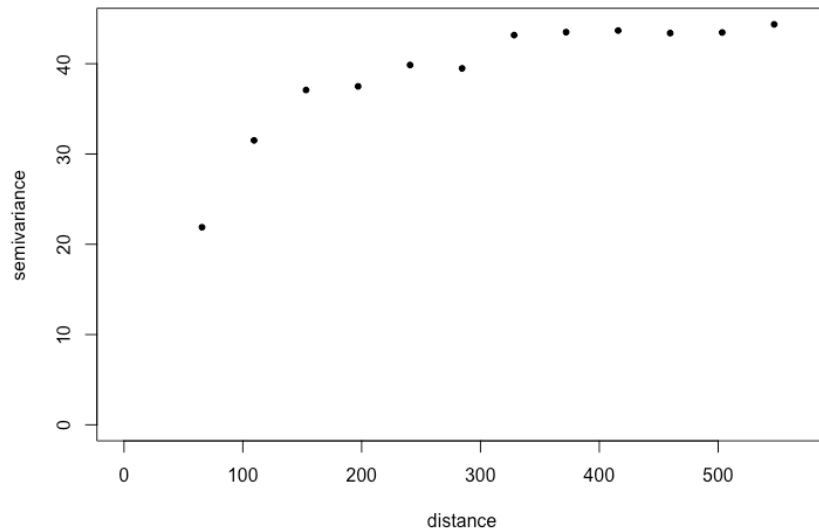


Figura 6: Variograma para los datos de magnesio considerando una media lineal

Ahora corresponde efectuar el ajuste del modelo para el variograma empírico, al igual que en el paso anterior. Los valores iniciales para determinarlo en términos del silo, rango y nugget con los siguientes:

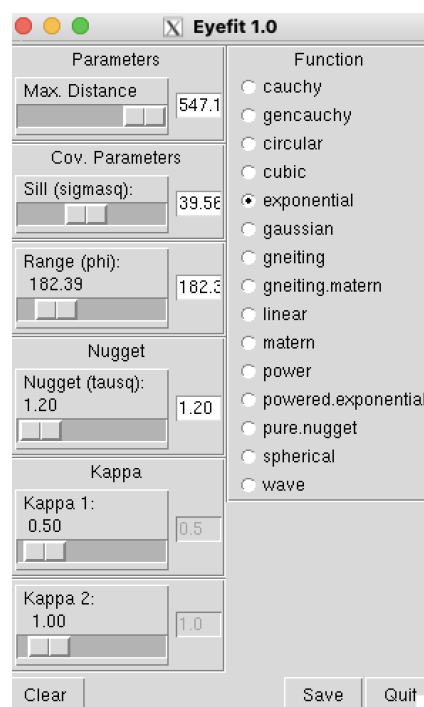


Figura 7: Valores iniciales para el silo, rango y nugget para ajustar el variograma empírico considerando una media lineal

El ajuste visual para este variograma considerando los modelos: exponencial, gaussiano y cúbico es el siguiente:

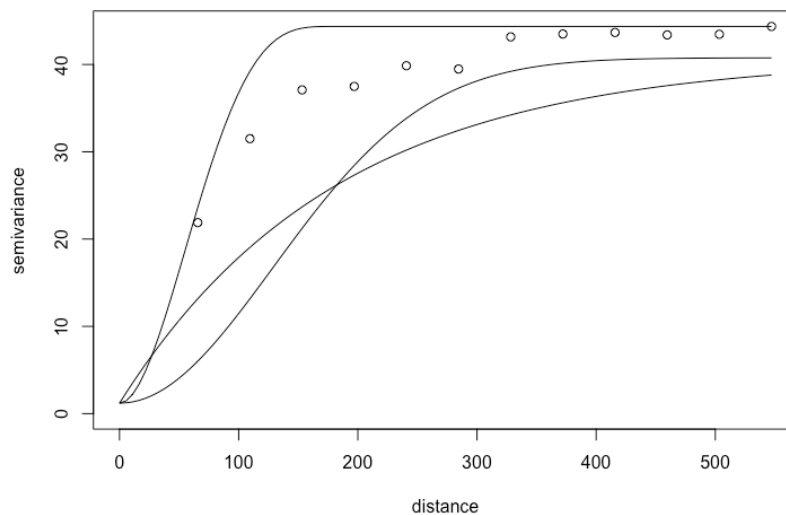


Figura 8: Ajuste visual de modelos exponencial, gaussiano y cúbico para el variograma, considerando una media lineal.

El resultado de los 3 ajustes en términos del AIC es el siguiente:

Ajuste	AIC
<i>Exponencial</i>	1094.085
<i>Gaussiano</i>	1164.069
<i>Cúbico</i>	1130.842

La gráfica con estos 3 ajustes sobre el variograma con media lineal es el siguiente:

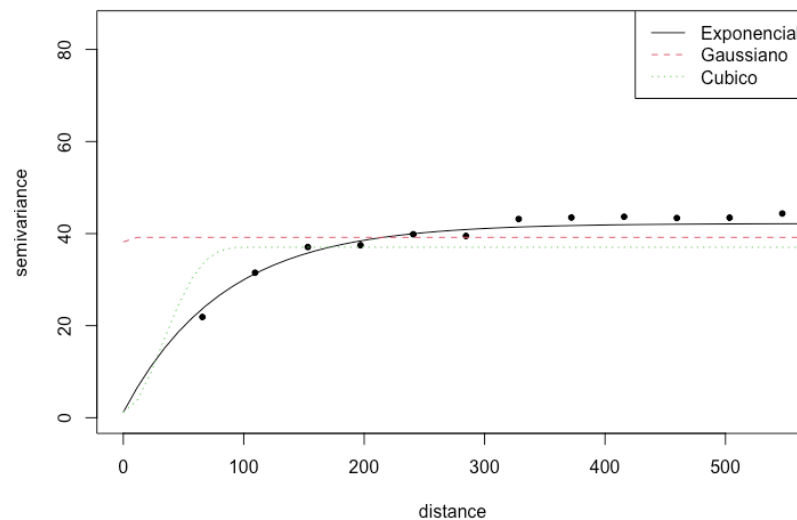


Figura 9: Variograma de media lineal con 3 ajustes distintos propuestos

En este escenario, nuevamente el ajuste exponencial para el variograma empírico es el que tiene el menor AIC.

Un cuadro resumen para los ajustes considerando la media constante y lineal, se muestra a continuación:

	Media constante		Media lineal	
	Espacial	No espacial	Espacial	No espacial
<b>Exponencial</b>	1093.4	1162	1094.085	1162
<b>Gaussiano</b>	1125.141	1162	1164.069	1162
<b>Cubico</b>	1169.12	1162	1130.842	1162

Figura 10: Cuadro resumen con los AIC para variogramas con media constante y lineal, en comparación del AIC no espacial

Se puede ver que para el escenario del modelo exponencial y gaussiano (media constante o lineal) el AIC de los modelos ajustados para el variograma empírico es siempre menor que al AIC no espacial, dando cuenta de que la estructura espacial es algo importante para el conjunto de datos. No así, con el caso del modelo cúbico, que en el caso de considerar la media constante, el modelo no espacial tiene un mejor “rendimiento” que el modelo que considera un ajuste cúbico para el variograma empírico.

Adicionalmente, para ambos escenarios el modelo exponencial es el de menor AIC, y bajo este criterio la mejor opción. Sin embargo, si se considera un ajuste de media constante, el AIC es menor que en el caso de que la media sea constante, pero por muy poca diferencia.

El mejor modelo en este caso, es el modelo exponencial, con media constante, para ajustar el variograma empírico de los datos de magnesio.

### 1.5. Pregunta 5

- Realice un kriging apropiado (con la función de media y función de covarianza escogida en los puntos anteriores) para este conjunto de datos e interprételo. Grafique el kriging y la varianza del kriging estimado.

Para poder visualizar el modelo, es necesario efectuar primero una grilla para poder mostrar las predicciones, esto se hace con el siguiente código:

```
# creamos la grilla para la prediccion
n = 50
grilla_mg <- expand.grid(east=seq(4957,5961,l=n),
                        north=seq(4829,5710,l=n))

plot(grilla_mg,pch=20,col="gray70")
points(mg020$coords, col=2, pch=20)
```

Lo que entrega el siguiente resultado:

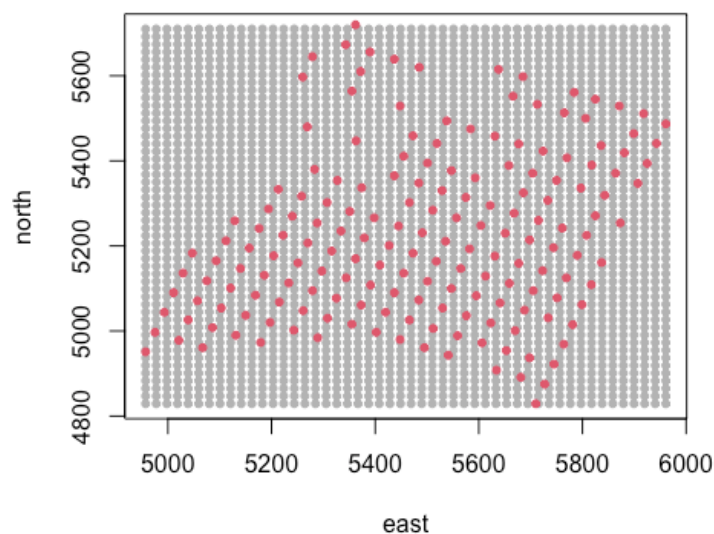


Figura 11: Grilla de puntos junto con los datos de magnesio (color rosado)

A continuación podemos efectuar el ajuste para un kriging. En este caso, el modelo escogido fue el exponencial considerando una media constante. Por ello, con el siguiente código, podemos calcular el kriging para este escenario:

```
# kriging ordinario  
  
okc <- krige.conv(mg020,  
                  locations=grilla_mg,  
                  krige=krige.control(obj.model=lik1,trend.d = "cte",trend.l = "cte"))
```

los campos “trend” indican que tanto para los valores entregados como argumento (los de magnesio) como para los predecidos, considere también una media constante. Esta decisión es para dar continuidad respecto a la elección del modelo con mejor AIC.

Los resultados de la predicción se muestran en las siguientes figuras:

### Predicciones para el valor de magnesio

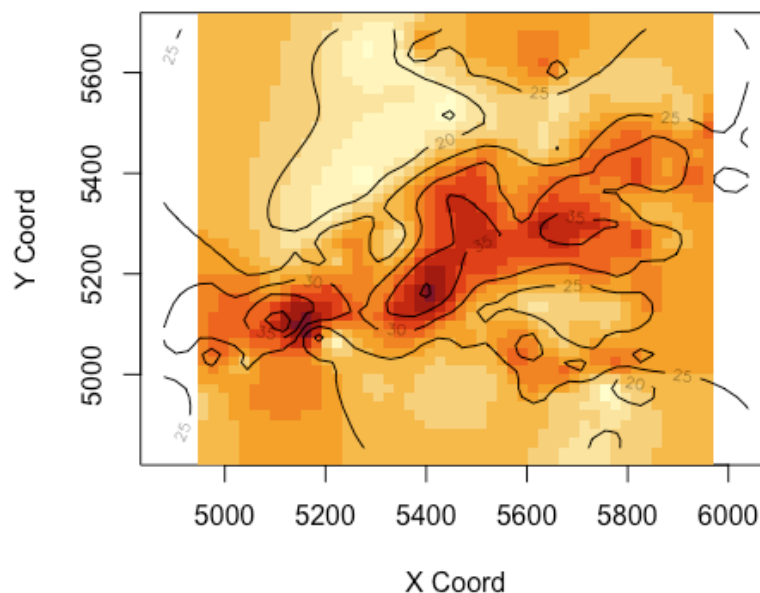


Figura 12: Valores predecidos para el magnesio acorde a las coordenadas

Podemos ver, al comparar visualmente con la figura 11, que se predicen valores altos de magnesio para las coordenadas que están en los extremos de la “X Coord”, lo que corresponde al norte. Es decir, mucho más al norte desde el centro y mucho más al sur desde esa posición, es posible encontrar zonas de alto valor de magnesio (color más oscuro). En cambio en el sentido de la “Y Coord”, más al este, existe una de las zonas de color más claro, que indica que existe muy bajos niveles de magnesio.

Para complementar el análisis se pueden graficar las varianzas del kriging:

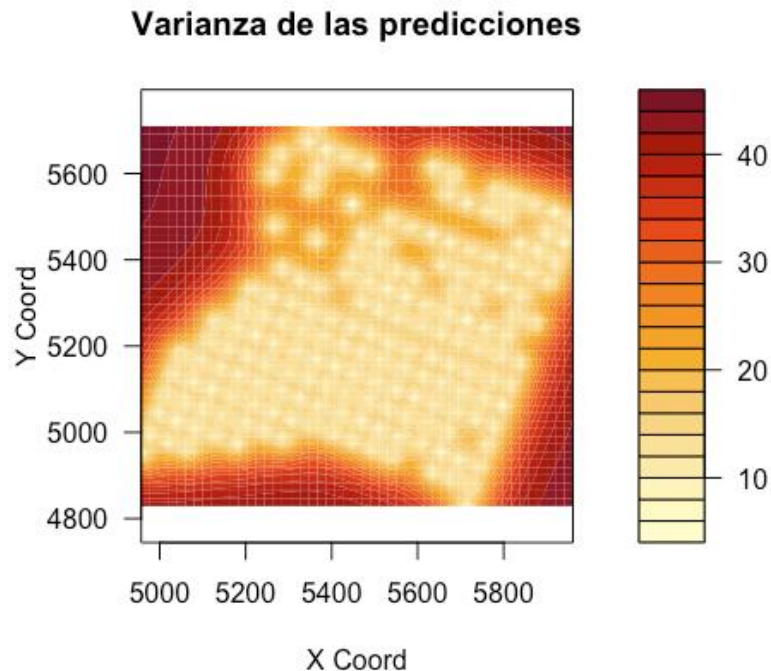


Figura 13: Varianza para los valores de magnesio predecidos

Al analizar en conjunto la figura 13 con la figura 12, podemos ver que en las zonas de color oscuro (donde se predicen altos valores de magnesio), la varianza es baja, lo que nos da una medida de confianza respecto a las predicciones. Las zonas ya más alejadas tienen una varianza sumamente alta, y en parte se debe a la falta de puntos para poder estimar dichos valores, algo que es esperable dentro del modelo y que se podría mejorar considerando más puntos dentro de la base.

## 2. Problema : Datos de área

Considere el conjunto de datos sobre suicidios en 32 municipios de Londres (excluyendo City London) en el periodo de 1089-1993 para hombres y mujeres combinados. Las variables que se registran son:

- Número de suicidios observados en el periodo estudiado ( $y$ )
- Número de casos de suicidio esperados ( $E$ )
- Índice de privación social ( $x_1$ )
- Índice de fragmentación social ( $x_2$ ), que refleja la falta de conexiones sociales y de sentido de comunidad.

## 2.1. Pregunta 1

- Cargue los archivos “LDNSuicides.shp” (polígono espacial) y “LondonSuicides.RData” (datos). Luego, unir en el polígono espacial el data.frame que contiene las variables.

```
# Pregunta 2: Datos de área - London Suicides -----  
  
# Cargar librerías necesarias  
library(sf)           # shapefiles  
library(spdep)        # análisis espacial y matrices de pesos  
library(spatialreg)   # Para modelos espaciales  
  
# Cargar el shapefile  
london_shape <- st_read("LDNSuicides.shp")  
str(london_shape)  
names(london_shape)  
  
# Cargar los datos de suicidios  
load("/cloud/project/LondonSuicides.RData")  
ls()  
  
# Crear un data frame con las variables cargadas  
LondonSuicides <- data.frame(y = y, E = E, x1 = x1, x2 = x2) # datos de los suicidios  
summary(LondonSuicides)  
  
# Verificar número de filas  
nrow(LondonSuicides)  
nrow(london_shape)  
  
# Unir y Añadir las columnas de los datos LondonSuicides a london_shape (polígono espacial)  
london_shape$y <- LondonSuicides$y  
london_shape$E <- LondonSuicides$E  
london_shape$x1 <- LondonSuicides$x1  
london_shape$x2 <- LondonSuicides$x2
```

Cargamos y unimos los datos con el polígono en un dataframe en la variable LondonSuicides

## 2.2. Pregunta 2



- Aplique los test de Moran y Geary para la variable SMR. Utilice la matriz W estandarizada por fila.
- ¿Se justifica un modelo espacial?

Se crea la variable SMR calculando la razón de mortalidad estandarizada.

```
# Crear variable SMR -----  
  
## La razon de mortalidad estandarizada se calcula como  
## RME = muertes observadas (y) / muertes esperadas (E)  
  
london_shape$SMR <- london_shape$y / london_shape$E
```

Creamos la matriz estandarizada (W) y la matriz binaria (B).

```
# Crear la matriz de pesos W -----  
  
library(spdep)  
adyacencia <- poly2nb(london_shape)  
  
# Matriz (W) estandarizada  
W <- nb2listw(adyacencia, style = "W", zero.policy = TRUE) ## Utilizar Modelo SAR  
  
# Matriz binaria (B)  
W_binaria <- nb2listw(adyacencia, style = "B", zero.policy = TRUE) ## Utilizar Modelo CAR
```

Realizamos el test de moran y geary con la matriz estandarizada (W).

```
# Test de Moran  
test_moran <- moran.test(london_shape$SMR, W)  
  
# Test de Geary  
test_geary <- geary.test(london_shape$SMR, W)  
  
# Resultados  
test_moran  
test_geary
```

```
> # Test de Moran
> test_moran <- moran.test(london_shape$SMR, W)
> # Resultados
> test_moran
```

Moran I test under randomisation

data: london\_shape\$SMR  
weights: W

Moran I statistic standard deviate = 1.4207, p-value = 0.0777  
alternative hypothesis: greater  
sample estimates:

Moran I statistic	Expectation	Variance
0.12215435	-0.03225806	0.01181287

En el resultado del Test de Moran se puede observar que el valor de moran es de 0.122 indicando una débil correlación espacial. A su vez, el valor esperado es de -0.032 siendo menor al índice de moran, indicando que tenemos una débil correlación espacial positiva. Por otro lado, el p valor es de 0.077 indicando que la autocorrelación espacial no es significativa, sin embargo el valor se encuentra cercano a la significancia. Por lo tanto, el test de moran nos estaría sugiriendo que no hay una evidencia fuerte de que los valores de SMR estén especialmente correlacionados.

```
> # Test de Geary
> test_geary <- geary.test(london_shape$SMR, W)
> test_geary
```

Geary C test under randomisation

data: london\_shape\$SMR  
weights: W

Geary C statistic standard deviate = 1.5149, p-value = 0.06489  
alternative hypothesis: Expectation greater than statistic  
sample estimates:

Geary C statistic	Expectation	Variance
0.83461314	1.00000000	0.01191808

Por otro lado, el test Geary al ser menor a 1 (aunque por muy poco) nos indica una débil correlación positiva. Pero al igual que en el test de moran, esta correlación no es significativa ya que el p-value es 0.05 superior al valor de significancia.

En conclusión, el resultado de ambos test tanto el de Moran como el de Geary sugieren que la correlación espacial no es significativa, por lo que no hay suficiente evidencia para justificar un modelo espacial.

### 2.3. Pregunta 3

- Considere los siguientes modelos para  $SMR = y/E$ :

(M1) :  $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \eta$ ,  $\eta \sim N(0, \sigma^2)$ ;

(M2) :  $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \eta$ ,  $\eta \sim SAR(\rho)$ ;

(M3) :  $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \eta$ ,  $\eta \sim CAR(\rho)$

Ajuste cada uno de estos modelos utilizando las funciones apropiadas de la librería "spatialreg". En el caso del modelo SAR usar el error arlm.

Realizamos los modelos Lineal, SAR y CAR:

```
# Ajuste de modelos espaciales para SMR -----  
# Modelo 1: Lineal  
model_1 <- lm(SMR ~ x1 + x2 + I(x1*x2), data = london_shape)  
summary(model_1)  
  
# Modelo 2: SAR -----  
model_2 <- errorsarlm(SMR ~ x1 + x2 + I(x1*x2), data = london_shape, listw = W)  
summary(model_2)  
  
# Modelo 3: CAR -----  
model_3 <- spautolm(SMR ~ x1 + x2 + I(x1*x2), data = london_shape, listw = W_binaria, family = "CAR")  
summary(model_3)
```

A. En base al AIC ¿Cuál de los 3 modelos escogería usted?

- **Modelo** **1:** **Lineal**

```
> # Modelo 1: Lineal
> model_1 <- lm(SMR ~ x1 + x2 + I(x1*x2), data = london_shape)
> summary(model_1)
```

Call:  
lm(formula = SMR ~ x1 + x2 + I(x1 \* x2), data = london\_shape)

Residuals:

Min	1Q	Median	3Q	Max
-0.24599	-0.08633	0.01131	0.07790	0.31742

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	1.06864	0.03007	35.535	< 2e-16 ***
x1	0.11340	0.02750	4.124	0.000301 ***
x2	0.21062	0.02654	7.937	1.21e-08 ***
I(x1 * x2)	0.06862	0.04901	1.400	0.172436

---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1371 on 28 degrees of freedom  
Multiple R-squared: 0.8077, Adjusted R-squared: 0.7871  
F-statistic: 39.2 on 3 and 28 DF, p-value: 3.714e-10

- **Modelo** **2:** **SAR**

```
> # Modelo 2: SAR -----
> model_2 <- errorsarlm(SMR ~ x1 + x2 + I(x1*x2), data = london_shape, listw = W)
> summary(model_2)
```

Call:errorsarlm(formula = SMR ~ x1 + x2 + I(x1 \* x2), data = london\_shape, listw = W)

Residuals:

Min	1Q	Median	3Q	Max
-0.219816	-0.093984	0.014620	0.079028	0.304376

Type: error  
Coefficients: (asymptotic standard errors)

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	1.073742	0.021416	50.1368	< 2.2e-16
x1	0.099666	0.025091	3.9722	7.122e-05
x2	0.220213	0.023171	9.5038	< 2.2e-16
I(x1 * x2)	0.063088	0.040902	1.5424	0.123

Lambda: -0.43339, LR test value: 2.1416, p-value: 0.14335  
Asymptotic standard error: 0.28275  
z-value: -1.5328, p-value: 0.12533  
Wald statistic: 2.3494, p-value: 0.12533

Log likelihood: 21.39551 for error model  
ML residual variance (sigma squared): 0.014804, (sigma: 0.12167)  
Number of observations: 32  
Number of parameters estimated: 6  
AIC: -30.791, (AIC for lm: -30.649)

- **Modelo**

**3:**

**CAR**

```
> # Modelo 3: CAR -----
> model_3 <- spautolm(SMR ~ x1 + x2 + I(x1*x2), data = london_shape, listw = W_binaria, family = "CAR")
> summary(model_3)

Call: spautolm(formula = SMR ~ x1 + x2 + I(x1 * x2), data = london_shape, listw = W_binaria, family = "CAR")

Residuals:
    Min       1Q   Median       3Q      Max
-0.223953 -0.072318  0.010553  0.091970  0.276398

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.075066   0.020723  51.8782 < 2.2e-16
x1           0.099904   0.024325   4.1070 4.008e-05
x2           0.219134   0.022358   9.8012 < 2.2e-16
I(x1 * x2)   0.058603   0.039676   1.4770  0.1397

Lambda: -0.22093 LR test value: 2.9781 p-value: 0.0844
Numerical Hessian standard error of lambda: 0.10729

Log likelihood: 21.81372
ML residual variance (sigma squared): 0.013424, (sigma: 0.11586)
Number of observations: 32
Number of parameters estimated: 6
AIC: -31.627
```

- Comparamos AIC de los tres modelos para determinar el que mejor se ajusta.

```
# Comparacion AIC -----

aic_model1 <- AIC(model_1)
aic_model2 <- AIC(model_2)
aic_model3 <- AIC(model_3)
```

```
> aic_model1 #LINEAL
[1] -30.64937
> aic_model2 #SAR
[1] -30.79101
> aic_model3 #CAR
[1] -31.62745
```

Modelo	AIC
Lineal clásico	-30.64937
SAR	-30.79101
CAR	-31.62745

Al comparar los valores para seleccionar el mejor modelo basado en el AIC, podemos observar que en base a los tres modelos ajustados, el modelo CAR es el que tiene un valor más bajo (-31.62745), lo cual nos indica que es el modelo con mejor equilibrio entre ajuste y complejidad, considerando la estructura espacial de los datos.

- B. Para el modelo seleccionado, según los p-valores ¿son todas las covariables significativas?

En el modelo CAR, el efecto de la interacción entre  $x_1$  y  $x_2$  no es significativa (p-valor  $\approx 0.1397$ ), lo cual indica que la interacción no contribuye de manera significativa al ajuste del modelo. Sin embargo, el intercepto  $x_1$  y  $x_2$  son estadísticamente significativos, dado que sus p-valores son menores a 0.05.

- C. En caso de ser necesario, reduzca (en el sentido de eliminar alguna covariable) el modelo anterior.

Se analizó el modelo CAR obtenido previamente, en donde se observan que existen dos variables predictoras  $x_1$  y  $x_2$ , ambas con p-valores muy bajos, lo cual indican que para el modelo son estadísticamente significativas. Por otro lado, la interacción entre las variables  $x_1$  y  $x_2$  presentan un p-valor de 0.1397 mayor al 0.05, lo cual nos sugiere que esta interacción no es estadísticamente significativa, por lo que esta interacción podría no ser relevante para el modelo. Por lo tanto, se realiza la prueba eliminando la interacción de  $x_1$  y  $x_2$  como se observa a continuación:

```
> model_car <- spautolm(SMR ~ x1 + x2, data = london_shape, listw = W_binaria, family = "CAR")
> summary(model_car)
```

Call: spautolm(formula = SMR ~ x1 + x2, data = london\_shape, listw = W\_binaria, family = "CAR")

Residuals:

Min	1Q	Median	3Q	Max
-0.224919	-0.085652	0.024064	0.082134	0.283804

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	1.096836	0.014980	73.2219	< 2.2e-16
$x_1$	0.088932	0.023975	3.7093	0.0002078
$x_2$	0.222529	0.022945	9.6984	< 2.2e-16

Lambda: -0.22602 LR test value: 3.0355 p-value: 0.081461  
Numerical Hessian standard error of lambda: 0.10689

Log likelihood: 20.75959  
ML residual variance (sigma squared): 0.014259, (sigma: 0.11941)  
Number of observations: 32  
Number of parameters estimated: 5  
AIC: -31.519

Como se puede observar, al eliminar la interacción de  $x_1$  y  $x_2$ , baja la complejidad del modelo, y se contempla que el AIC del modelo nuevo es ligeramente menor al AIC del modelo anterior, pero no existe un cambio notorio entre ambos AIC.

## 2.4. Pregunta 4

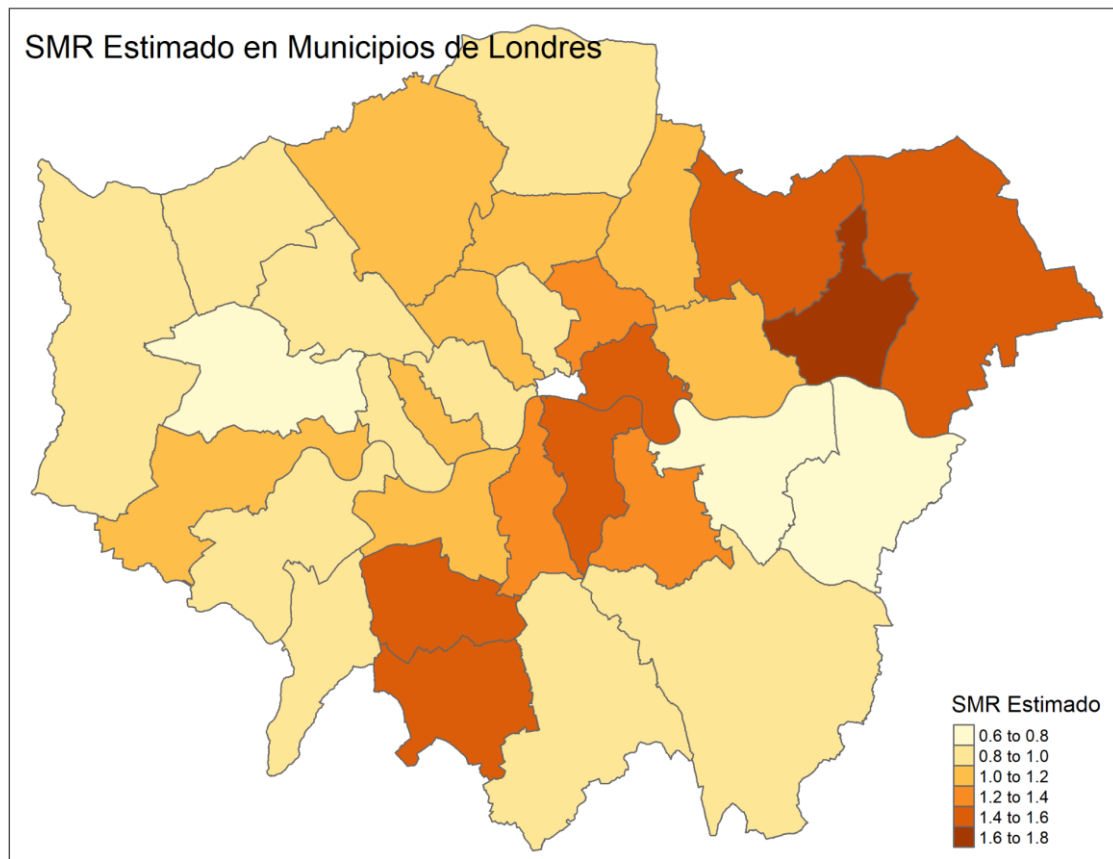
Para el modelo final, graficar el SMR estimado y el real. Para ello guarde en el polígono espacial el ajuste y el residuo del modelo final.

```
london_shape$SMR_est <- fitted(model_3) # modelo final seleccionado
london_shape$residuals <- residuals(model_3)

mapa_srm_estimado <- tm_shape(london_shape) + tm_polygons("SMR_est", title = "SMR Estimado") +
  tm_layout(title = "SMR Estimado en Municipios de Londres") # Graficar SMR estimado
tmap_save(mapa_srm_estimado, "SMR_mapa_estimado.png")

mapa_srm_real <- tm_shape(london_shape) + tm_polygons("SMR", title = "SMR Real") +
  tm_layout(title = "SMR Real en Municipios de Londres") # Graficar SMR real
tmap_save(mapa_srm_real, "SMR_mapa.png")
```

- **SMR estimado en Municipios de Londres:**



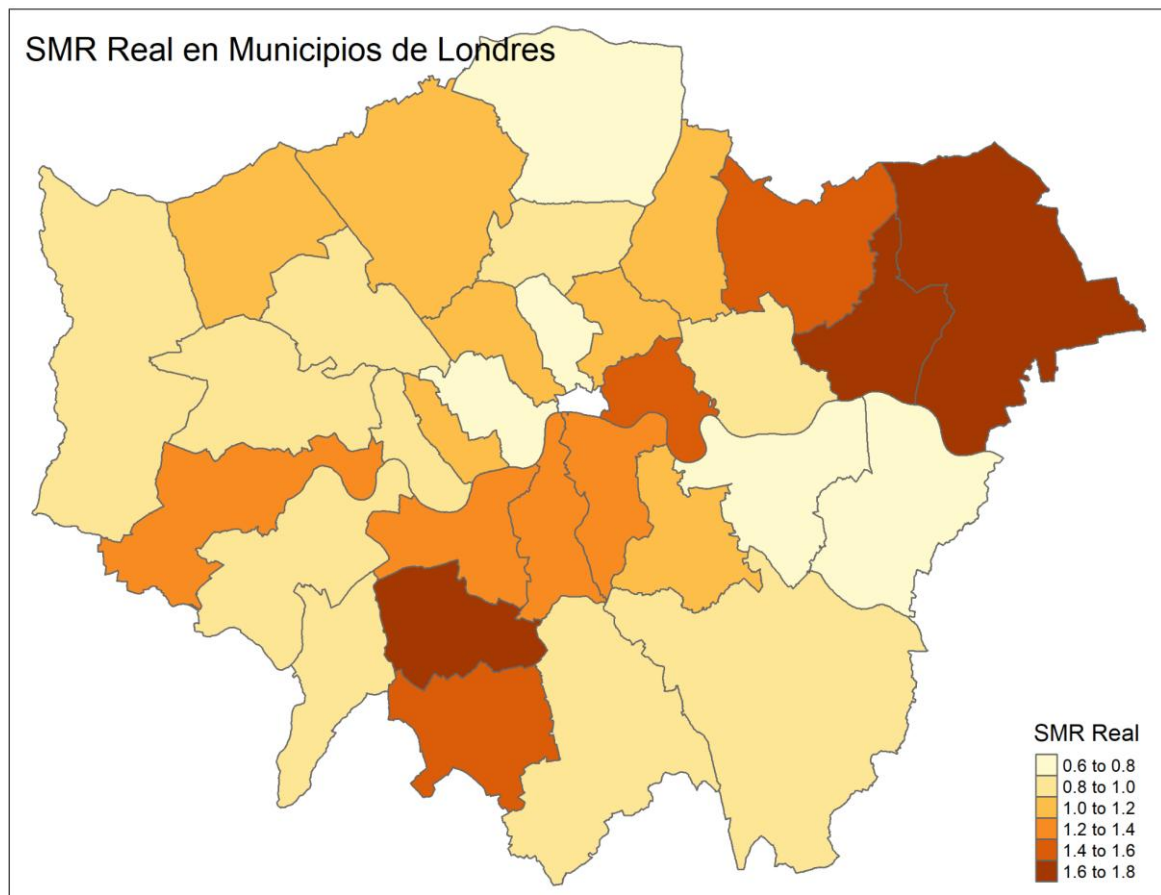
El mapa muestra el SMR estimado, donde las zonas de SMR alto (reflejado en los valores oscuros) son el noreste y sur de Londres. Estas zonas tienen un SMR estimado alto (1.4 a 1.6), según el modelo, se espera que la mortalidad en estas áreas es mayor a la media, lo cual puede indicar factores asociados a una mayor mortalidad.

En cuanto a las zonas de SMR bajo (valores claros), los municipios en el oeste y centro de Londres tienen valores de SMR estimado más bajos (0.6 a 1.0), esto indica que las zonas de mortalidad está por debajo o cerca de la media esperada.

Podemos concluir con el mapa de SMR estimado de municipios en Londres que, el modelo estima una distribución de mortalidad desigual, con algunas áreas en el noreste y sur con riesgo de mortalidad más alto y otras con riesgo bajo. Sobre las posibles razones detrás de las diferencias en las variaciones estimadas, pueden estar relacionadas con factores que el modelo considera, como ingresos, educación o acceso a servicios de salud. Asimismo, la existencia de municipios con SMR estimado alto, puede ser debido a factores locales adicionales que no están representados en el modelo, como estilo de vida, ambiente laboral y otros riesgos relacionados a la calidad de vida de los habitantes de Londres.



- SMR real en Municipios de Londres:



En cuanto a SMR real en municipios de Londres, se puede observar en el mapa que en zonas de SMR alto (colores oscuros), al igual que el mapa de SMR estimado, las zonas de noreste y sureste de Londres tiene los valores de SMR real más alto (1.4 a 1.8), esto indica que la mortalidad en estas áreas fue significativamente mayor a la esperada. Ante lo cual, puede estar relacionado con factores específicos de cada municipio, tales como condiciones de vida, acceso a servicios de salud o factores demográficos y socioeconómicos. Más aún, las áreas de alto SMR real son importantes para las intervenciones de salud públicas, ya que señalan zonas con un riesgo mayor de mortalidad que podrían requerir recursos adicionales.

Sumado a lo anterior, las zonas de SMR bajo (colores claros), las áreas en el centro y algunas en el oeste de Londres muestran un SMR real bajo (0.6 a 1.0), lo cual indica que la mortalidad en estas zonas está en línea, e incluso por debajo, de lo esperado. En este sentido, dichas zonas probablemente tengan mejores condiciones de vida o acceso a recursos que ayudan a reducir la mortalidad, presentando características poblacionales asociadas a un menor riesgo de mortalidad.

En conclusión, el análisis muestra que la mortalidad en Londres no se distribuye de manera uniforme. Los mapas revelan que las zonas noreste y sur tienen un riesgo de mortalidad notablemente mayor, mientras que las áreas del centro y algunas en el oeste tienen un riesgo menor. Esto sugiere que en las áreas con mayor riesgo hay factores específicos que podrían estar afectando la salud de las personas, y entender estos factores puede ayudar a mejorar las condiciones de vida en estos municipios de Londres.

## 2.5. Pregunta 5

- Aplique los test de Moran y geary a los residuos del modelo final.
- ¿Existe dependencia espacial?

```
> moran_resid <- moran.test(london_shape$residuals, W)
> print(moran_resid)
```

Moran I test under randomisation

data: london\_shape\$residuals  
weights: W

Moran I statistic standard deviate = 3.4821, p-value = 0.0002488  
alternative hypothesis: greater  
sample estimates:

Moran I statistic	Expectation	Variance
0.34742166	-0.03225806	0.01188936

Al aplicar el test de Moran vemos que el índice I (0.34) es mayor al índice de expectativa (-0.032) lo que indica una correlación positiva, el p valor indica que la correlación espacial es significativa.

```
> geary_resid <- geary.test(london_shape$residuals, W)
> print(geary_resid)
```

Geary C test under randomisation

data: london\_shape\$residuals  
weights: W

Geary C statistic standard deviate = 3.1764, p-value = 0.0007455  
alternative hypothesis: Expectation greater than statistic  
sample estimates:

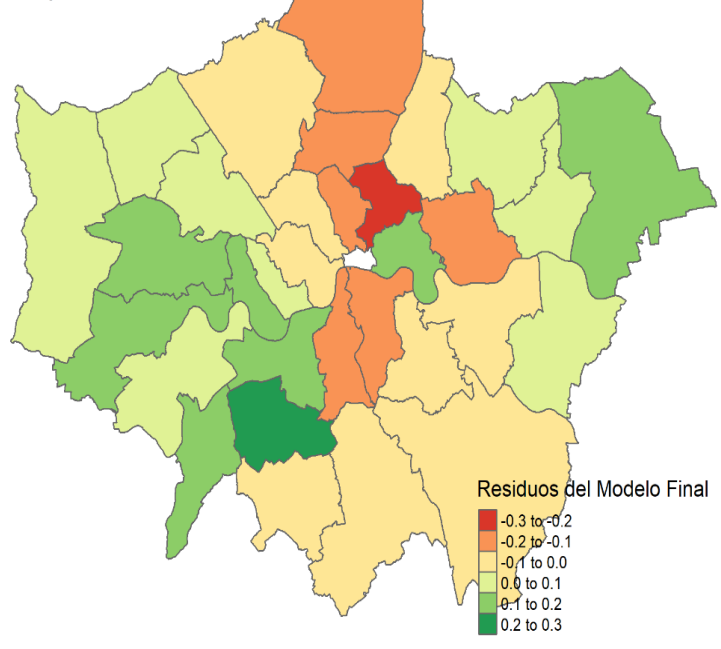
Geary C statistic	Expectation	Variance
0.65362085	1.00000000	0.01189118

Al igual que en el test de moran, el índice C (0.65) del test de Geary al ser menor que 1 indica una correlación positiva que es significativa dado un p-value (0.01) tan alto.

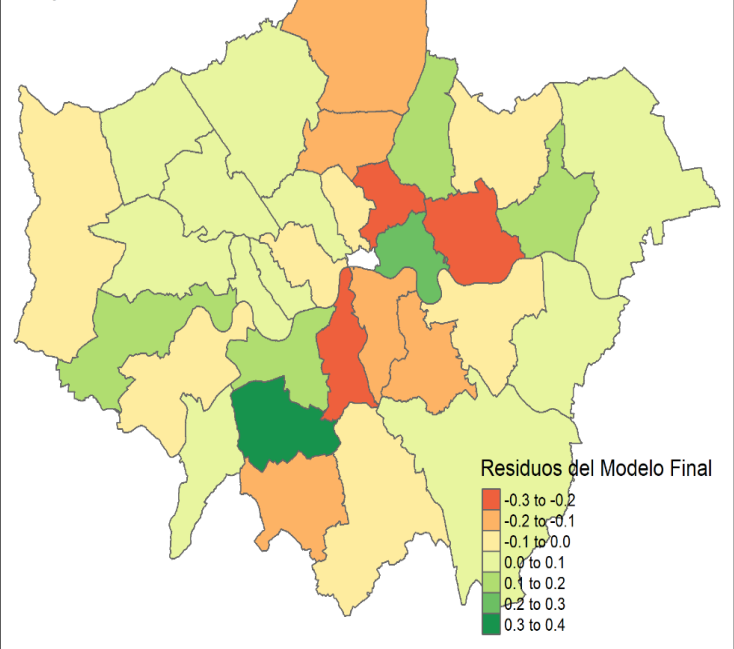
Dado los resultados obtenidos de los Test de Moran y Test de Geary, es posible rechazar la hipótesis nula, ya que ambos test arrojaron valores de p muy bajos. Esto indica que existe una correlación espacial. La presencia de correlación espacial en los residuos sugiere que el modelo actual no captura completamente la estructura espacial de los datos, lo cual podría reflejar que aún existe dependencia espacial en las áreas vecinas que el modelo no está considerando.

- Grafique los residuos de este modelo y los residuos del modelo lineal con las mismas covariables.

Mapa de Residuos Modelo CAR



Mapa de Residuos Modelo Lineal



Podemos observar los dos mapas de residuos, el del modelo CAR que se encuentra a la izquierda y el del modelo lineal a la derecha. En el modelo CAR, se observa una distribución algo homogénea en el sector de Londres, con una buena presencia de zonas en amarillo, lo que indica que el modelo tiene residuos cercanos a cero en varias áreas. Existen algunas zonas en verde y en tonos naranja/rojos, pero estas se encuentran distribuidas de manera menos concentrada que en el modelo lineal. Esto sugiere que el Modelo CAR, captura mejor la variación espacial en los datos, resultando en menos áreas con grandes subestimaciones o sobreestimaciones. Por otro lado, en el modelo lineal, se observa una mayor presencia de residuos extremos, particularmente en verde oscuro y rojo oscuro en los sectores cercanos al centro de Londres donde los valores en el mapa son más variables, lo cual indica que el modelo lineal tiene dificultades para capturar esa variabilidad. La concentración de residuos verdes y rojos es más notable, lo cual sugiere que el modelo lineal tiene un ajuste menos preciso en comparación con el Modelo CAR, especialmente en áreas donde hay dependencia espacial que no logra captar.

En conclusión, la comparación entre los residuos del Modelo CAR y el Modelo Lineal muestra que el Modelo Car tiene un mejor ajuste a los datos, ya que presenta menos áreas con residuos extremos, lo que implica que captura mejor la estructura espacial de los datos. Además su distribución más uniforme de los residuos indica que es más efectivo al ajustar la dependencia espacial, mientras que el modelo lineal muestra más residuos extremos y concentración de colores, sugiriendo que es menos preciso.