Centro de Investigación en Matemáticas (CIMAT). Unidad Monterrey

Análisis de datos complejos: Tarea 4

Jorge Luis Ramos Zavaleta

6 de junio de 2019

1. Ejercicio

Considera el conjunto de datos Paraná de la libreria `geoR` (`data(parana)`, `help(parana)`)

1. ¿Qué preguntas científicas son interesantes de contestar? ¿Incluyen estimación, predicción o pruebas de hipótesis?
2. Identifica la región de interés, el diseño, la respuesta y las covariables, si las hay.
3. ¿Cuál es la *señal* subyacente?

1.1. Solución

Algunas preguntas que pueden plantearse con el conjunto de datos son: ¿En qué regiones se dan las mayores y menores concentraciones de lluvia en el estado de Parana durante la temporada de sequía? Las concentraciones de lluvia en este periodo se dan cerca de asentamientos urbanos? ¿Qué tan intensa puede ser la lluvia en las áreas donde no existen estaciones? ¿Existe una variación espacio-temporal con respecto a la intensidad de lluvia a lo largo de los años que representan los datos?

La primera pregunta se puede responder fácilmente agrupando los datos colocados en percentiles por lo necesita ningún método pesado de análisis. Aunque de igual forma se puede considerar generar clusters espaciales para ello y lograr un resultado más preciso. La segunda pregunta igualmente se puede responder fácilmente dibujando sobre el mapa del estado

de Parana los poligonos de los asentamientos urbanos sobre los que se quiera conocer la cantidad de lluvia que cae en ellos, aparte de considerar alguna medida de magnitud que indique la medida de lluvia en cada estación cercana a dichos poligonos, aunque para obtener una posible mejor estimación sería recomendable hacer predicción usando kriging. La tercera pregunta ciertamente debe ser resuelta haciendo uso de predicción debido a que las areas de las que queremos conocer su intensidad pluvial no tienen asignado ningun valor por lo que se debe estimar. La última pregunta puede conseguirse una posible respuesta haciendo uso de clusters espacio-temporales pero para esto deben usarse los datos originales no solo los promedios que tenemos.

La región de interés es el estado de Paraná en Brasil, la variable de respuesta es el promedio de lluvia durante la temporada de sequia en varios años, y las unicas covariables presentes son las coordenadas de las estaciones donde se realizo la medición. Para el caso del diseño se considero tomar el promedio de densidad pluvial durante varios años en varias estaciones ubicadas en el estado de Parana en Brasil.

Para obtener la señal subyacente se probaron diferentes modelos para ajustar el semivariograma de las observaciones. En la figura 1.1 se puede observar el ajuste de dos de los mencionados modelos. El modelo ajustado en azul corresponde a un modelo exponencial o matern con kappa=0.5, y modelo ajustado en rojo corresponde a un modelo matern con kappa=1.5. En este sentido podemos decir que la señal subyacente corresponde con un modelo matern con kappa=1.5 debido a que genera un buen ajuste.

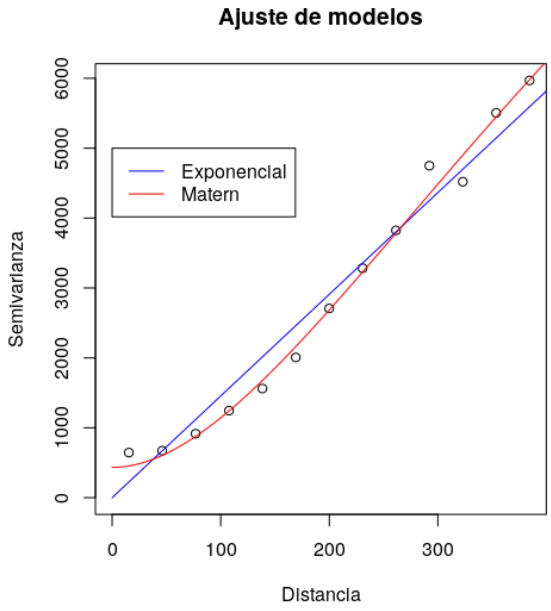


Figura 1.1: Ajuste de modelos exponencial (kappa=0.5) y Matern (kappa=1.5) al semivariograma generado por los datos

2. Ejercicio

Considera los siguientes dos modelos para un conjunto de respuestas Y_i, $i = 1, \ldots, n$, asociados con una secuencia de posiciones x_i a lo largo de un eje espacial de una dimensión x.

1. $Y_i = \alpha + \beta x_i + \epsilon_i$, donde α y β son parámetros y los ϵ_i son mutuamente independientes con media cero y varianza σ^2_ϵ.

2. $Y_i = A + B x_i + \epsilon_i$, donde los ϵ_i son como en el inciso anterior pero A y B son variables aleatorias, independientes entre sí y entre los ϵ_i, cada una con media cero y varianzas σ^2_A y σ^2_B.

Para cada uno de esos modelos, encuentra el promedio y varianza de Y_i, y la covarianza entre Y_i, Y_j, para cualquier $j \neq i$. Dado una sola realización de cualquier modelo, sería posible distinguirlos?

2.1. Solución

Para el caso de las esperanzas tenemos

$$E(\alpha + \beta x_i + \epsilon_i) = \alpha + \beta x_i$$

y

$$E(A + Bx_i + \epsilon_i) = E(A) + E(B)x_i + E(\epsilon_i) = 0$$

Dado esto, tenemos que las varianzas son

$$Var(\alpha + \beta x_i + \epsilon_i) = E(\alpha + \beta x_i + \epsilon_i - (\alpha + \beta x_i))^2 = E(\epsilon_i)^2 = \sigma_\epsilon^2$$

y

$$\begin{aligned} Var(A + Bx_i + \epsilon_i) &= E(A + Bx_i + \epsilon_i)^2 = \\ &= E(A^2 + 2ABx_i + 2Bx_i\epsilon_i + 2A\epsilon_i + (Bx_i)^2 + \epsilon_i^2) = \\ &= E(A^2) + E(B^2)x_i^2 + E(\epsilon^2) = \sigma_A^2 + \sigma_B^2 x_i^2 + \sigma_\epsilon^2 \end{aligned}$$

Para el caso de las covarianzas se tiene

$$\begin{aligned} Cov(\alpha + \beta x_i + \epsilon_i, \alpha + \beta x_j + \epsilon_j) &= E((\alpha + \beta x_i + \epsilon_i)(\alpha + \beta x_j + \epsilon_j)) - (\alpha + \beta x_i)(\alpha + \beta x_j) = \\ &= \alpha^2 + \alpha\beta x_i + \alpha\beta x_j + \beta^2 x_i x_j - (\alpha^2 + \alpha\beta x_i + \alpha\beta x_j + \beta^2 x_i x_j) \\ &= 0 \end{aligned}$$

y

$$\begin{aligned} Cov(A + Bx_i + \epsilon_i, A + Bx_j + \epsilon_j) &= E((A + Bx_i + \epsilon_i)(A + Bx_j + \epsilon_j)) = \\ &= E(A^2 + ABx_i + A\epsilon_i + ABx_j + B^2 x_i x_j + B\epsilon_i x_j + A\epsilon_j + B\epsilon_j x_i + \epsilon_i\epsilon_j) \\ &= E(A^2 + B^2 x_i x_j) = \sigma_A^2 + \sigma_B^2 x_i x_j \end{aligned}$$

Debido a que solo consideramos una realización no podemos hacer uso de los resultados de la covarianza, los cuales nos permitirían distinguir mas claramente la realización de un modelo de otro. Cabe observar que la media del segundo modelo tiene media 0 y el otro una media variable que se mueve alrededor de una recta con intercepto α, lo cual podría permitirnos identificar realizaciones de cada modelo siempre que las varianza de cada modelo fuera suficientemente pequeña para no cruzarse ya que esto nos impediría reconocer a que modelo pertenece cada realización.

3. Ejercicio

Un estudio piloto para estudiar la obesidad infantil fue llevado a cabo en algunas escuelas de los municipios de San Pedro y Santa Catarina en el estado de Nuevo León.

En este estudio, se recopilaron datos a través de encuestas sobre las características familiares y demográficas de niños en edad preescolar en busca de variables relacionadas con problemas potenciales de obesidad temprana. Luego de un pre-proceso, que incluyó la georeferenciación de los domicilios, se tienen las ubicaciones mostradas en la Figura mostrada en la tarea, donde se señala con círculos de diferente diámetro, el peso en kilogramos de los niños.

La información del estudio se encuentra en `obesidad.zip`, y contiene los archivos vectoriales `ESRI-shapefile`.

1. Realiza una estimación espacial mediante Kriging usando como variable de interés (por separado): `Peso`, `Tallacms` y `Circintu`. Compara visualmente los resultados y escribe tus comentarios y conclusiones.

   Describe las características y supuestos de tu modelo, incluyendo la estructura de covarianzas y los parámetros ajustados.

2. Si el objetivo del estudio es analizar y caracterizar el fenómeno del aumento de peso infantil (no precisamente obesidad) en la región metropolitana de Monterrey, qué tipo de información (de preferencia de acceso libre) sugerirías que se incorporara para tal objetivo?

3.1. Solución

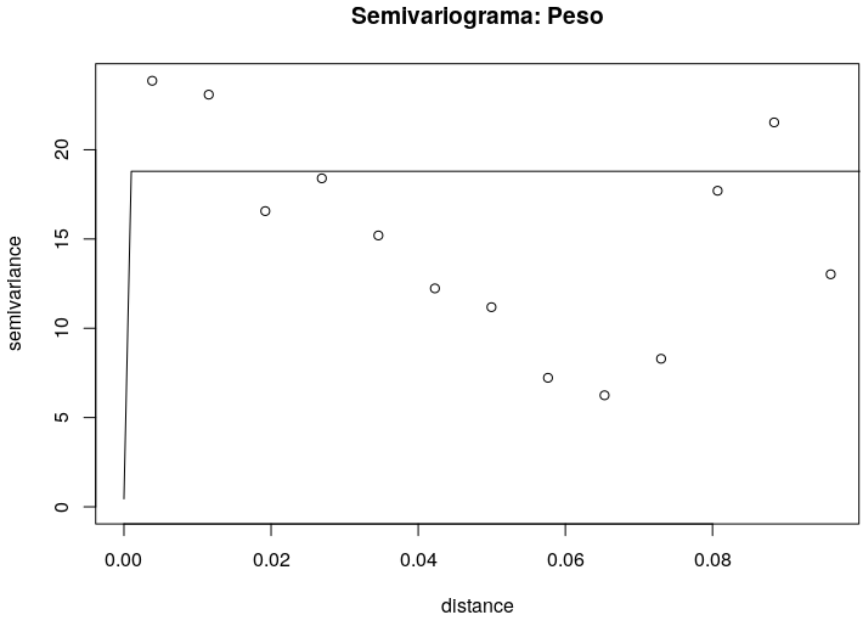
Para realizar el ejercicio de interpolación usando kriging para los datos, primero se debe tener una idea de que modelo puede ser el que sigue el semivariograma de cada una de las variables en cuestión. En las figuras 3.1, 3.2 y 3.3 se pueden observar dichos semivariogramas, algo que debe observarse es que la semivarianza no tiende a generar un crecimiento monotóno por lo que difícilmente se podra encontrar un modelo de covarianza que empate completamente con lo observado en los semivariogramas debido a que no parecen cumplir las hipótesis de estacionariedad. En particular, en el semivariograma de la variable `Peso` se puede observar un patrón de tipo U al cual no se le puede ajustar ningun modelo por completo que permita una buena interpolación de los datos, a menos que se considere no usar supuestos estacionarios.

En los semivariogramas presentados también se muestran los modelos que mejor ajuste mostraron para las variables `Tallacms` y `Circintu` que son un Matern con $\kappa = 0{,}5$ y un Matern con $\kappa = 2$, en el caso de la variables `Peso` se muestra el ajuste de un modelo Matern con $\kappa = 1{,}5$ y como puede observarse no logra un ajuste suficientemente bueno para considerarlo. Debido a esto último se considero hacer kriging con varios parámetros para el caso del modelo de covarianzas Matern y realizar un kriging con modelo de covarianzas Gaussiano. Los resultados de dicha interpolación usando kriging para la variables `Peso` se

muestran en las figuras 3.4 y 3.5.

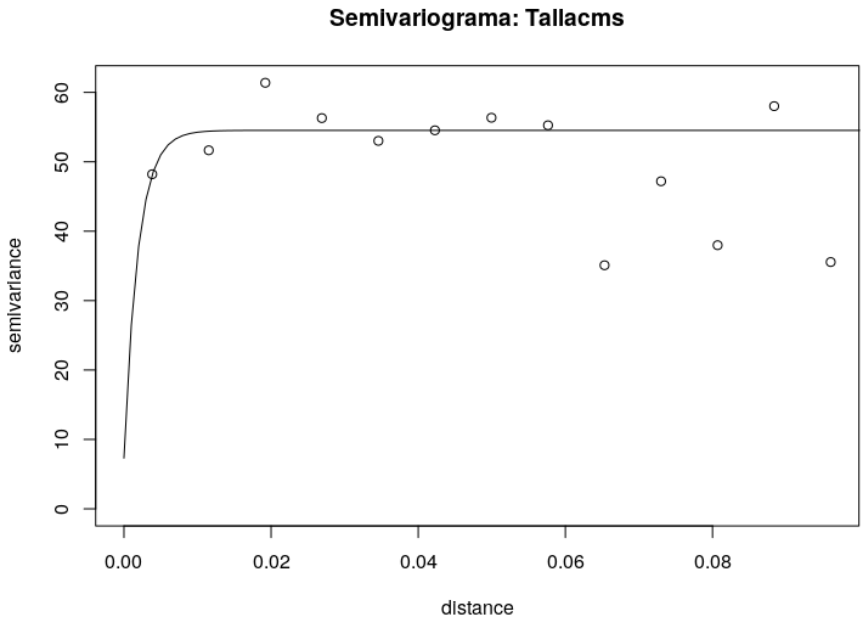
En el caso de dichas interpolaciones puede observarse que el ajuste con respecto de los datos es similar para todos los casos, pero recordando la forma del semivariograma podemos indicar que la interpolación no es necesariamente certera en ninguno de los casos debido al mal ajuste que se tiene con los modelos de covarianzas con respecto del semivariograma.

En las figuras 3.6 y 3.7 se pueden observar las interpolaciones logradas para el caso de las variables `Tallacms` y `Circintu` respectivamente. De acuerdo al ajuste que se logro en el semivariograma con los respectivos modelos de covarianzas usados se espera que la interpolación nos ofrezca una buena predicción.



: Figura 3.1
Semivariograma de la variable Peso, con un modelo ajustado con modelo de covarianza Matern con paramétro $\kappa = 1{,}5$

Si el objetivo del estudio fuese analizar y caracterizar el aumento de peso infantil una opción sería incorporar información del número de establecimientos que venden comida no "saludable."[a]lrededor de escuelas digamos a un par de kilometros, aunque se debería considerar una distancia un poco mas amplia debido al crecimiento del parque vehicular en los últimos años en Monterrey y su zona metropolitana. La información de dicho numero de

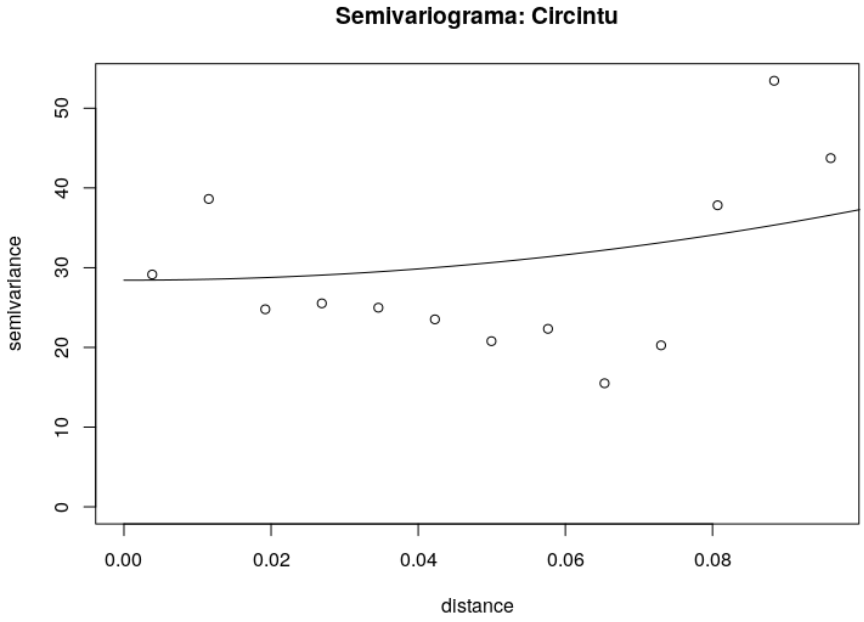


: Figura 3.2
Semivariograma de la variable Tallacms, con un modelo ajustado con modelo de covarianza Matern con paramétro $\kappa = 0{,}5$

establecimientos puede conseguirse haciendo uso del DENUE, inclusive puede limitarse la búsqueda al numero de establecimientos de comida rapida en un cierto radio alrededor de las escuelas.

Se sugiere usar las escuelas como punto de referencia debido a que regularmente las familias tienden a llevar a sus hijos a escuelas cerca de sus residencias e incluso en algunos estados (desconozco el caso de Nuevo León) las escuelas discriminan la inscripción con respecto a la cercania de la residencia del infante con respecto de la escuela.

Otra opción no tan de acceso libre sería considerar la razón entre infantes y parque vehicular por unidad de area, pero habría que discriminar del parque vehicular vehiculos que son de uso para transporte de mercancia como trailers. Esta información sobre el parque vehicular debería estar disponible por los departamentos de transito y sino pueden ser aproximados usando la información contenida en las licencias de conducir.



: Figura 3.3
Semivariograma de la variable Circintu, con un modelo ajustado con modelo de covarianza Matern con paramétro $\kappa = 2$

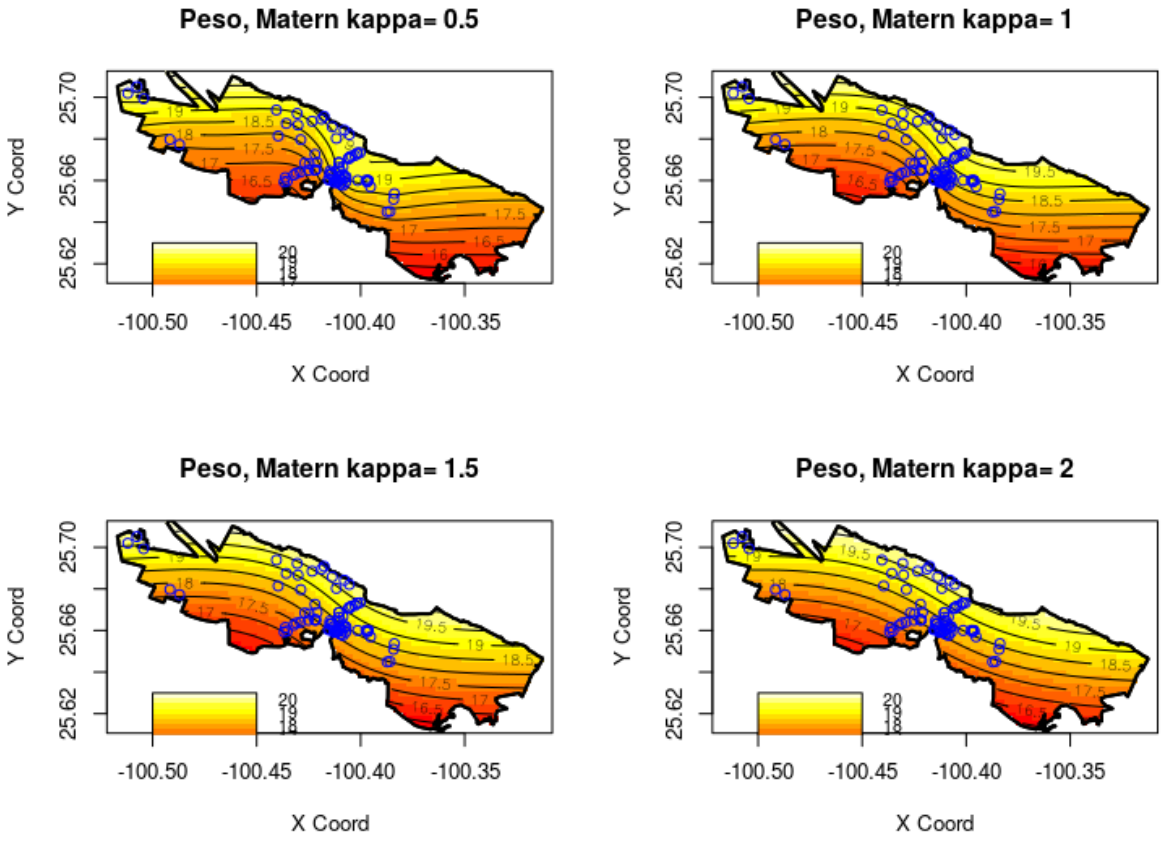


Figura 3.4: Interpolación por kriging con modelo Matern para la variable Peso con diferentes κ

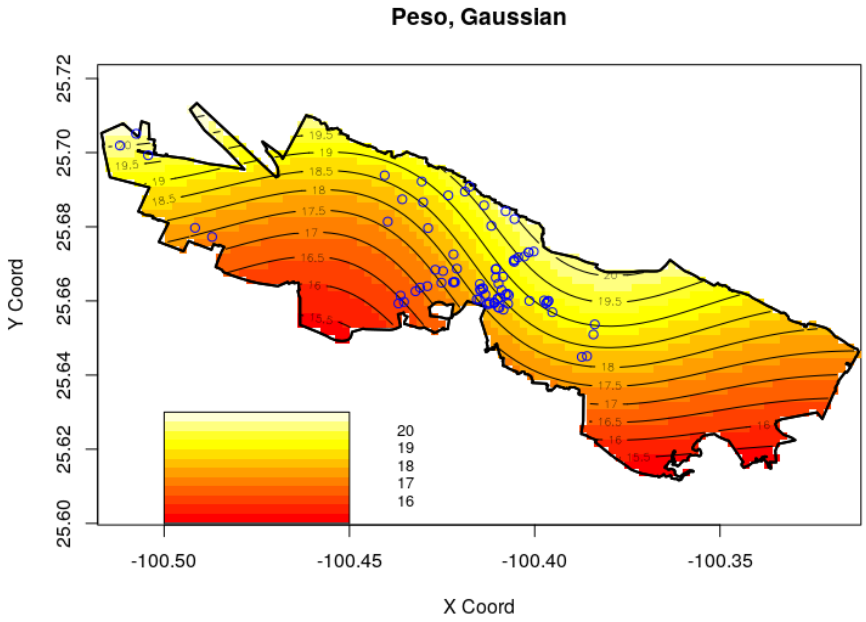


Figura 3.5: Interpolación por kriging con modelo Gaussiano para la variable Peso

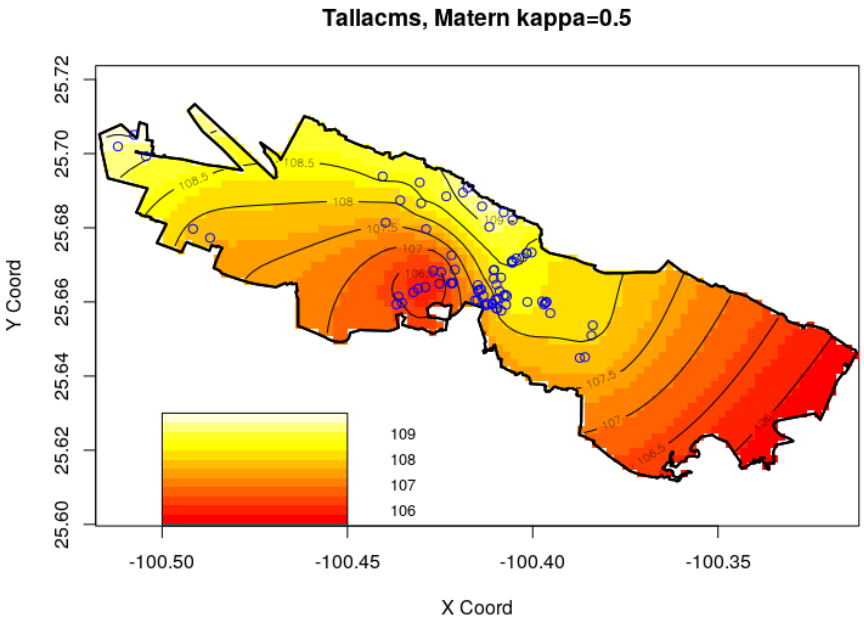


Figura 3.6: Interpolación por kriging con modelo Matern para la variable Tallacms con $\kappa = 0{,}5$

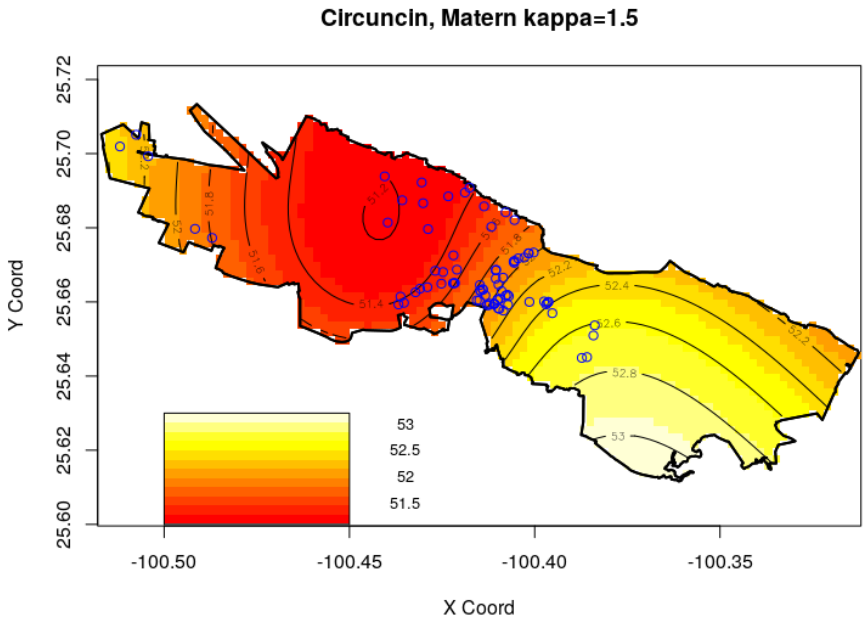


Figura 3.7: Interpolación por kriging con modelo Matern para la variable Circintu con $\kappa = 1{,}5$