



UNIVERSIDAD NACIONAL AUTÓNOMA  
DE MÉXICO

---

---

FACULTAD DE CIENCIAS

MÉTODOS GEOESTADÍSTICOS DE  
ESTIMACIÓN PARA LA  
MODELACIÓN DE DATOS  
CLIMATOLÓGICOS

T E S I S  
QUE PARA OBTENER EL TÍTULO DE:  
MATEMÁTICO

P R E S E N T A:  
FERNANDO JOSAFATH AÑORVE LÓPEZ

DIRECTOR DE TESIS:

DR. JONATHAN MONTALVO URQUIZO



Ciudad Universitaria, Cd. Mx., 2018



Universidad Nacional  
Autónoma de México



**UNAM – Dirección General de Bibliotecas**  
**Tesis Digitales**  
**Restricciones de uso**

**DERECHOS RESERVADOS ©**  
**PROHIBIDA SU REPRODUCCIÓN TOTAL O PARCIAL**

Todo el material contenido en esta tesis esta protegido por la Ley Federal del Derecho de Autor (LFDA) de los Estados Unidos Mexicanos (México).

El uso de imágenes, fragmentos de videos, y demás material que sea objeto de protección de los derechos de autor, será exclusivamente para fines educativos e informativos y deberá citar la fuente donde la obtuvo mencionando el autor o autores. Cualquier uso distinto como el lucro, reproducción, edición o modificación, será perseguido y sancionado por el respectivo titular de los Derechos de Autor.

## Hoja de datos del jurado

**1. Datos del alumno**

Añorve  
López  
Fernando Josafath  
(81) 11 58 49 80  
Universidad Nacional Autónoma de México  
Facultad de Ciencias  
Matemáticas  
414007213

**2. Datos del tutor**

Dr.  
Jonathan  
Montalvo  
Urquizo

**3. Datos del sinodal 1**

Dra.  
Lizbeth  
Naranjo  
Albarrán

**4. Datos del sinodal 2**

M. en C.  
José Salvador  
Zamora  
Muñoz

**5. Datos del sinodal 3**

M. en C.  
Graciela  
Martínez  
Sánchez

**6. Datos del sinodal 4**

Dr.  
Jesús  
López  
Estrada

**7. Datos del trabajo escrito**

Métodos geoestadísticos de estimación para la modelación de datos  
climatológicos  
133 p.  
2018

*A mis padres y hermanos.*

## *Agradecimientos*

*Agradezco infinitamente al Dr. Jonathan Montalvo Urquizo, por su tiempo y paciencia.*

*Agradezco mis padres, por su confianza y apoyo. A mis hermanos, por ser parte de mi alegría*

*Agradezco también a Emilio, Tamara, Ana Luisa y Dolores, por su enorme hospitalidad y por no dejarme solo.*

*Finalmente, a Ro, por estar ahí desde el inicio.*

## Resumen

El presente trabajo tiene como objetivo establecer un modelo geostatístico de interpolación que prediga los valores de temperatura, humedad relativa y presión atmosférica en los puntos del Área Metropolitana de Monterrey (AMM) a partir de un conjunto de registros disponibles de ciertas estaciones climatológicas en la metrópoli obtenidas de la *Base de datos climatológicos, de tráfico y de contaminación de Monterrey*, del Centro de Investigación en Matemáticas Unidad Monterrey. Asimismo, se buscaba identificar diferencias entre la efectividad de este método aplicado para distintas variables en al menos dos periodos del día.

Bajo el supuesto de que están distribuidas espacialmente de forma continua, a las variables climatológicas mencionadas se les asoció con una función aleatoria  $Z(\mathbf{x})$  que arroja una variable aleatoria según la posición  $\mathbf{x}$ . Cada lectura en una posición determinada se considera como una realización de la variable aleatoria correspondiente a dicha posición. Posteriormente, se estudió la correlación espacial de dicha función calculando su variograma utilizando tres modelos de ajuste: exponencial, esférico y gaussiano. El modelo de variograma se utilizó para determinar los pesos que tiene cada una de las mediciones conocidas sobre el estimador kriging, definido como el estimador lineal óptimo para valores de una función aleatoria. Durante el proceso, se excluyeron algunas mediciones y estaciones que dificultaban el estudio de la correlación espacial.

Al estudiar las tres variables climatológicas, tanto la varianza nugget como el umbral y el rango de los modelos de variograma aumentaron considerablemente al cambiar de las mediciones obtenidas en el día a las de la noche, sobre todo al utilizar el modelo exponencial. De los tres modelos estudiados, el gaussiano ajustó mejor a los datos en tres casos, seguido del modelo exponencial en dos casos y del esférico con un único caso.

# Índice

Índice	VII
<b>1 Introducción</b>	<b>1</b>
1.1. Objetivos . . . . .	1
1.2. Zonas importantes del AMM . . . . .	2
1.3. Presentación de los datos . . . . .	2
<b>2 Modelos Geoestadísticos</b>	<b>5</b>
2.1. Introducción . . . . .	5
2.2. Conceptos básicos y definiciones . . . . .	6
2.3. Etapas del estudio geoestadístico . . . . .	11
2.4. Determinación de las distancias . . . . .	12
2.5. Análisis exploratorio de los datos . . . . .	13
2.6. Estimación y modelación del variograma. . . . .	14
2.7. Kriging . . . . .	23
2.8. Metodología de la aplicación del modelo . . . . .	34
<b>3 Kriging en datos de temperatura</b>	<b>37</b>
3.1. Análisis del 1 al 5 mayo de 2017, de 17:00 a 17:15 . . . . .	37
3.2. Análisis del 1 al 5 mayo de 2017, de 00:00 a 00:15 . . . . .	49
<b>4 Kriging en datos de humedad relativa</b>	<b>59</b>
4.1. Análisis del 1 al 5 mayo de 2017, de 17:00 a 17:15 . . . . .	60
4.2. Análisis del 1 al 5 mayo de 2017, de 00:00 a 00:15 . . . . .	70
<b>5 Kriging en datos de presión atmosférica</b>	<b>79</b>
5.1. Análisis del 1 al 5 de mayo de 2017, de 17:00 a 17:15 . . . . .	79
5.2. Análisis del 1 al 5 de mayo de 2017, de 00:00 a 00:15 . . . . .	90
<b>6 Conclusión</b>	<b>101</b>
<b>Apéndice A Funciones definidas en los códigos</b>	<b>103</b>
<b>Apéndice B Código de R para el análisis preliminar</b>	<b>117</b>
<b>Apéndice C Código de R para el kriging</b>	<b>119</b>

VIII

*ÍNDICE*

**Bibliografía**

**125**

# Capítulo 1

## Introducción

Los altos registros de contaminación registrados en el Área Metropolitana de Monterrey (AMM) por el Sistema Integral de Monitoreo Ambiental (SI-MA) hacen necesario un estudio detallado de los factores que influyen en los niveles de contaminación de la ciudad. Es por ello que el Centro de Investigación en Matemáticas, Unidad Monterrey, creó la “Monterrey Weather, Traffic and Pollution Database” (MWTP) (Montalvo-Urquizo, Marroquín, Hernández-Castillo, y Hernández-González, 2017), con el fin de hacer un análisis de las variables climatológicas y de tráfico que influyen en la concentración y dispersión de los contaminantes en el AMM.

Dentro del proyecto general existe la necesidad de conocer valores de las variables en cualquier punto perteneciente al AMM, lo cual requiere la construcción de estimadores geoestadísticos de las variables de interés, como la temperatura, la presión atmosférica, o los niveles de los contaminantes más importantes para la calidad del aire. En este contexto, este trabajo está planteado como una contribución a generar una metodología que permita automatizar los estimadores para ser calculados ya sea con datos históricos o con datos medidos en tiempo real, como los que la MWTP recopila constantemente.

### 1.1. Objetivos

Para este trabajo se plantearon los siguientes objetivos:

- Establecer y evaluar el método de interpolación kriging para estimar distintas variables climatológicas a lo largo del AMM: temperatura, humedad relativa y presión atmosférica.
- Identificar diferencias entre la efectividad de este método aplicado para distintas variables en al menos dos momentos del día.
- Implementar las metodologías en el paquete estadístico R que permita hacer kriging, que pueda adaptarse fácilmente a distintas variables de

interés y que durante el proceso permita realizar acciones como detectar y procesar valores atípicos. Al mismo tiempo, este código de R debería poder utilizarse para estudios posteriores.

## 1.2. Zonas importantes del AMM

En este estudio fue importante tener en cuenta la geografía del AMM, por lo que seis puntos importantes de la metrópoli se muestran en la Figura 1.1:

- A. Sierra Madre Oriental, ubicada al sur del AMM
- B. Cerro de las Mitras, en la zona poniente del AMM
- C. Parque Ecológico Chipinque, ubicado en el municipio de San Pedro Garza García, en la zona sur del AMM.
- D. Zona de la Huasteca (Santa Catarina), en la zona sudoeste.
- E. Aeropuerto Internacional Mariano Escobedo, en la zona noreste.
- F. Cerro del Obispado, cerca del centro del AMM

## 1.3. Presentación de los datos

Los datos utilizados para el análisis fueron obtenidos de la *Base de datos climatológicos, de tráfico y de contaminación de Monterrey*, o MWTP por sus siglas en inglés (Montalvo-Urquizo et al., 2017). Como su nombre lo establece, esta base de datos contiene información sobre tres rubros principales. En particular, la información climatológica contenida en esta base de datos fue descargada y catalogada desde el portal Weather Underground, que recolecta datos sobre el clima de más de 200,000 estaciones climatológicas personales (ECP) y aproximadamente 12,000 estaciones públicas alrededor del mundo. Dichas ECP se localizan en lugares diversos como oficinas, casas, patios, etc., y reportan las condiciones del clima periódicamente en intervalos que van de los 5 a los 60 minutos, dependiendo de cada estación.

El AMM cuenta actualmente con 30 ECP y dos estaciones públicas ubicadas en sus dos aeropuertos. Las condiciones climatológicas registradas por estas estaciones incluyen temperatura, presión atmosférica, humedad, tasa de precipitación, dirección y velocidad del viento, entre otros. Cabe señalar que la MWTP contiene datos históricos registrados desde 1973 por las estaciones públicas, mientras que para algunas ECP contiene registros diarios desde 2008. Estos registros superaban en febrero de 2017 los cuatro millones y, en suma, contenían un total de cerca de 40 millones de datos climatológicos puntuales.



Figura 1.1: Mapa del Área Metropolitana de Monterrey y sus zonas periféricas

Adicionalmente, la MWTP contiene datos sobre la calidad del aire registrados por las diez estaciones del Sistema Integral de Monitoreo Ambiental, e información obtenida en tiempo real sobre la concentración de tráfico a través de ciertas rutas del AMM. En la Figura 1.2 se observan en los puntos rojos las localizaciones de las ECP; en los puntos magenta, las estaciones públicas ubicadas en los aeropuertos; en los puntos verdes, las diez estaciones del SIMA; y en líneas azules, las rutas vehiculares que son monitoreadas para medir el tráfico en tiempo real.

En este trabajo se utilizaron los datos relativos a la temperatura, presión atmosférica, y humedad relativa. Dichas variables climatológicas están expresadas en grados Celsius ( $^{\circ}\text{C}$ ), hectopascales (hPa) y porcentaje (%), respectivamente. La Tabla 1.1 indica la descripción de las variables climatológicas que fueron utilizadas y la forma en que éstas se hallan en la base de datos.

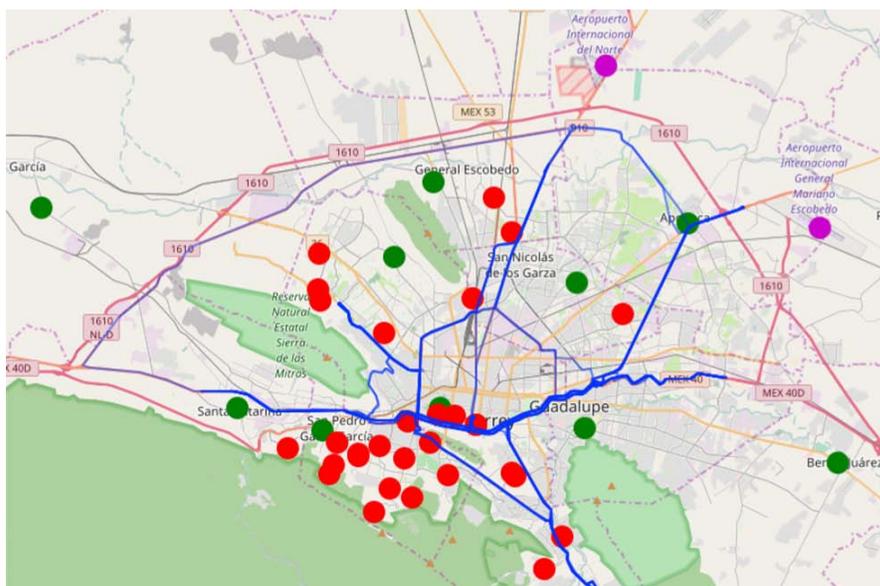


Figura 1.2: Mapa del AMM con las ubicaciones de las ECP (puntos rojos), las estaciones climatológicas en los aeropuertos (puntos magenta), las estaciones del SIMA (puntos verdes) y las rutas de transporte monitoreadas (líneas azules)

Tabla 1.1: Descripción de las variables encontradas en la base de datos MWTP que fueron usadas en este trabajo.

Nombre de la variable	Descripción	Formato/Unidades
TIMESTAMP	Fecha y hora de la medición	aaaa-mm-dd hh:mm:ss
TEMP	Temperatura	°C
HUM	Humedad	%
PRESSURE	Presión atmosférica	hPa

## Capítulo 2

# Modelos Geoestadísticos

### 2.1. Introducción

En áreas como la minería, el estudio ambiental y la ciencia en general, se necesita establecer, en una región determinada, un modelo que describa la manera en que se manifiesta un fenómeno a través del espacio. Por ejemplo, en la agricultura puede ser importante estudiar la distribución de los componentes de la tierra en un área, o bien, en la industria minera podría ser relevante conocer la concentración de cierto mineral en el subsuelo para saber en qué áreas es costoso explorar y en cuáles no. Generalmente, para conocer dichas distribuciones se cuenta únicamente con un conjunto limitado de mediciones del fenómeno de interés, por lo que es necesario establecer un método que prediga los valores asociados a dicho fenómeno en los puntos que no fueron muestreados (Soto Rodríguez, 2015).

El método utilizado en este estudio para interpolar en los puntos del AMM donde no existe medición tiene fundamento en la geoestadística. Según Díaz-Viera, Herrera-Zamarrón, y Valdés-Manzanilla (2009), la geoestadística es una rama de la estadística espacial donde se modela la correlación entre datos de fenómenos regionalizados para, a través de ella, hacer predicciones sobre su comportamiento. Se conoce como *fenómeno regionalizado* a un fenómeno que se manifiesta a través del espacio en un dominio determinado, y la herramienta que permite modelar la correlación entre sus datos es el *variograma*.

A lo largo de este capítulo se establece la metodología a seguirse para el estudio geoestadístico. En la primera sección se mencionan y definen algunos conceptos básicos utilizados a lo largo del estudio geoestadístico. En general, estas definiciones están basadas en el trabajo de Zavaleta Sánchez (2014), salvo cuando se indica lo contrario. Más adelante, se describen las etapas en las que se divide la metodología del trabajo. Finalmente, se describe a detalle en qué consiste cada una de dichas etapas.

## 2.2. Conceptos básicos y definiciones

Antes de presentar la metodología a seguirse a lo largo de la investigación, es necesario definir algunos conceptos teóricos que sirven como fundamento para el estudio geoestadístico.

Llamaremos  $\mathfrak{D} \subset \mathbb{R}^2$ , un conjunto acotado, al dominio en el que se distribuye el fenómeno regionalizado. Este dominio es una representación en  $\mathbb{R}^2$  de la región geográfica analizada.

Para el estudio geoestadístico se cuenta con un conjunto de *datos geoestadísticos*, que están dados como una colección finita  $(\mathbf{x}_i, y_i)$  con  $i = 1, 2, \dots, n$ , donde para cualquier  $i \in 1, 2, \dots, n$  se tiene que  $\mathbf{x}_i \in \mathfrak{D}$  y  $y_i \in \mathbb{R}$  es un escalar, llamado *variable de respuesta*, asociado a la posición  $\mathbf{x}_i$  en el dominio geográfico. Cabe señalar que la variable de respuesta debe estar definida continuamente a lo largo de todo  $\mathfrak{D}$ . Es decir, que el conjunto  $\{y_i\}$  es generado por una función real continua  $f : \mathfrak{D} \subset \mathbb{R}^2 \rightarrow \mathbb{R}$  al ser evaluado en los puntos del dominio  $\{\mathbf{x}_i\}$ . Por medio del kriging se pretende encontrar una aproximación de  $f(x)$  para todo  $x$  en  $\mathfrak{D}$ .

La propiedad que se desea medir del fenómeno regionalizado toma valores dependiendo del punto en el que se toma la medición. Dado un punto  $\mathbf{x} \in \mathfrak{D}$  con dos coordenadas  $(\theta, \phi)$ , al rango de posibles valores que puede tomar la propiedad en dicho punto  $x$  se le asocia una variable aleatoria que depende de él. Para modelar esta situación se utiliza una *función aleatoria*  $Z(\mathbf{x})$ , también conocida como *variable regionalizada* que, tal y como la define Díaz-Viera et al. (2009), es un proceso estocástico que arroja una variable aleatoria de acuerdo a la posición  $\mathbf{x}$  cuando ésta se mueve a través del dominio.

Para estudiar la región de interés se cuenta con una cantidad finita de puntos  $\mathbf{x}_i$  en  $\mathfrak{D}$ , y en cada uno de ellos se observa un valor aleatorio  $Y_i$ , donde  $i = 1, 2, \dots, n$ . Cada  $y_i$  puede interpretarse como una realización de la variable aleatoria  $Y_i$ . Este valor brinda información del fenómeno regionalizado, ya sea porque lo mide directamente o porque está correlacionado con el fenómeno y permite analizarlo de forma indirecta. Un *modelo geoestadístico* consta de dos componentes principales: un proceso estocástico que toma valores reales  $\{Z(\mathbf{x}) : \mathbf{x} \in \mathfrak{D}\}$  y el vector aleatorio  $Y = (Y_1, Y_2, \dots, Y_n)$  con distribución multivariada condicionada a  $Z(\mathbf{x})$ .

Conviene señalar que para  $1 \leq i \leq n$ , la variable  $Z(\mathbf{x}_i)$  modela el fenómeno regionalizado de manera ideal en el punto  $\mathbf{x}_i$ , mientras que  $Y_i$  expresa una cantidad relacionada con  $Z(\mathbf{x}_i)$  que se puede observar y medir, con algún margen de error. Sin embargo, para fines prácticos podemos considerar en este trabajo que  $Y_i = Z(\mathbf{x}_i)$ .

La *función media* de  $Z(\mathbf{x})$  se obtiene calculando el valor esperado de la

función aleatoria evaluada en el punto  $\mathbf{x}$ , esto es:

$$m(\mathbf{x}) = \mathbb{E}[Z(\mathbf{x})].$$

La *función de covarianza*  $\sigma : \mathfrak{D}^2 \rightarrow \mathbb{R}$  está dada por:

$$\sigma(\mathbf{x}, \mathbf{x}') = \text{Cov}\{Z(\mathbf{x}), Z(\mathbf{x}')\} = \mathbb{E}[(Z(\mathbf{x}) - m(\mathbf{x}))(Z(\mathbf{x}') - m(\mathbf{x}'))],$$

donde  $\mathbf{x}, \mathbf{x}' \in \mathfrak{D}^2$ .

Cuando la función de covarianza se evalúa en el caso particular  $\mathbf{x} = \mathbf{x}'$ , se obtiene la varianza de la función aleatoria en  $\mathbf{x}$

$$\sigma(\mathbf{x}, \mathbf{x}) = \text{Var}(Z(\mathbf{x})).$$

Estos momentos pueden no existir.

## Estacionariedad

La efectividad de la estadística se basa en la estacionariedad del fenómeno regionalizado. Se entiende por *estacionariedad* a la hipótesis de que la función aleatoria tiene el mismo grado de variación en una región de interés. Físicamente, la estacionariedad de una función significa que el fenómeno es homogéneo en el espacio y que, por así decirlo, se repite a sí mismo en todo el espacio. (Zavaleta Sánchez, 2014)

Una consecuencia de la estacionariedad de una función aleatoria es que si existen sus momentos, éstos son invariantes bajo traslaciones. Esto significa que si se toman en cuenta la media y la covarianza, existen  $\mu \in \mathbb{R}$  y  $C : \mathbb{R}^2 \rightarrow \mathbb{R}$  tales que para cualesquiera  $\mathbf{x}, \mathbf{x} + \mathbf{h} \in \mathfrak{D}$ ,

$$\begin{aligned} m(\mathbf{x}) &= \mu \\ \sigma(\mathbf{x}, \mathbf{x} + \mathbf{h}) &= C(\mathbf{h}) \end{aligned} \tag{2.1}$$

Es decir, la media es constante a lo largo de la región y la covarianza  $C$  entre dos puntos  $\mathbf{x}, \mathbf{x} + \mathbf{h}$  depende únicamente del vector de separación  $\mathbf{h}$ . Estas dos propiedades conforman la hipótesis de estacionariedad de segundo orden, que es más débil que la hipótesis de estacionariedad, y se define según Armstrong (1998) como sigue:

**Definición 2.2.1.** Se dice que una función  $Z(\mathbf{x})$  es *estacionaria de segundo orden* o *débilmente estacionaria* si cumple las condiciones establecidas en 2.1.

No obstante, en la vida real es común encontrar situaciones en las que no es posible dar por hecho que la media es constante, y en cuando ello ocurre, no se cumplen las hipótesis anteriores de estacionariedad. En estas situaciones se puede buscar una hipótesis más débil, la estacionariedad intrínseca. Esta hipótesis se basa en el comportamiento de la media y la varianza de los incrementos  $Z(\mathbf{x}) - Z(\mathbf{x} + \mathbf{h})$ .

**Definición 2.2.2.** Se dice que una función  $Z(\mathbf{x})$  es una *función aleatoria intrínseca* si satisface que para cualquier traslación de un vector  $\mathbf{h}$  de forma que  $\mathbf{x}, \mathbf{x} + \mathbf{h} \in \mathfrak{D}$  se cumplen las siguientes condiciones:

$$\begin{aligned}\mathbb{E}[Z(\mathbf{x} + \mathbf{h}) - Z(\mathbf{x})] &= 0 \\ \mathbb{E}[(Z(\mathbf{x} + \mathbf{h}) - Z(\mathbf{x}))^2] &= 2\gamma(\mathbf{h})\end{aligned}\quad (2.2)$$

A la función  $\gamma(\mathbf{h})$  se le llama *variograma*, mientras que al valor que toma el variograma en algún vector en particular  $\mathbf{h}_0 \in \mathbb{R}^2$  se le conoce como semivarianza en el vector de separación  $\mathbf{h}_0$ . Es conveniente notar que la hipótesis de estacionariedad de segundo orden implica la estacionariedad intrínseca, que es menos estricta, mientras que esta última no necesariamente implica la primera. Este hecho lo ilustra Myers (1989) con un ejemplo que se muestra a continuación.

Cabe señalar que, aunque el ejemplo utiliza una función con dominio discreto en los números naturales, se puede considerar un caso particular de dominio en  $\mathbb{R}^2$  si se toma en cuenta

$$\mathfrak{D} = \{(0, 0), (1, 0), \dots, (n, 0)\}.$$

Sea  $Z(n)$  un proceso estocástico puntual para  $n \in \mathbb{N}$ , donde  $Z(0) = 0$ . Definimos los incrementos independientes

$$\begin{aligned}Y(i) &= +1 && \text{con probabilidad } 0.5 \\ Y(i) &= -1 && \text{con probabilidad } 0.5\end{aligned}$$

A partir de ellos, para cada  $n \geq 1$  se define  $Z(n) = \sum_{i=1}^n Y(i)$ . Se comprueba que la función cumple la hipótesis de estacionariedad intrínseca:

$$\begin{aligned}\mathbb{E}[Z(n+m) - Z(n)] &= \sum_{i=n+1}^{n+m} \mathbb{E}[Y(i)] = 0 \\ \mathbb{E}[(Z(n+m) - Z(n))^2] &= \text{Var}\{Z(n+m) - Z(n)\} \\ &= \sum_{i=n+1}^{n+m} \text{Var}\{Y(i)\} = m.\end{aligned}$$

Es decir, la media de los incrementos  $Z(n+m) - Z(n)$  es nula, mientras que su varianza existe y está dada en función de  $m$ .

Se puede verificar que  $Z(n)$  no cumple con la hipótesis de estacionariedad de segundo orden ya que si se considera una separación de  $m = 0$ , de acuerdo con la hipótesis de estacionariedad de segundo orden, debería ocurrir que

$$\sigma(n, n) = \sigma(n, n+0) = C(0).$$

Por lo tanto, la covarianza no dependería de  $n$ . No obstante,

$$\sigma(n, n) = \text{Var}\{Z(n)\} = \sum_{i=1}^n \text{Var}\{Y(i)\} = n.$$

### Funciones definidas positivas

Se había comentado anteriormente que, bajo la hipótesis de estacionariedad de segundo orden, dada una función aleatoria  $Z(\mathbf{x})$  existe su función de covarianza  $C : \mathbb{R}^2 \rightarrow \mathbb{R}$  y su función media es igual a una constante  $\mu$ . Tomando esto en cuenta, considérese la combinación lineal

$$Z^* = \sum_{i=1}^n \lambda_i Z(\mathbf{x}_i),$$

donde  $n$  es la cantidad de mediciones,  $\lambda_i$  son los pesos de la combinación y  $\mathbf{x}_i$  son los sitios de medición.

Dado que la media de  $Z(\mathbf{x})$  es  $\mu$ , resulta que

$$\mathbb{E}[Z^*] = \mathbb{E}\left[\sum_{i=1}^n \lambda_i Z(\mathbf{x}_i)\right] = \mu \sum_{i=1}^n \lambda_i.$$

También se tiene que la varianza está definida como

$$\text{Var}\{Z^*\} = \mathbb{E}[Z^* - \mathbb{E}[Z^*]]^2.$$

Desarrollando la varianza se obtendría entonces

$$\begin{aligned} \text{Var}\{Z^*\} &= \mathbb{E}\left[\sum_{i=1}^n \lambda_i Z(\mathbf{x}_i) - \mu \sum_{i=1}^n \lambda_i\right]^2 \\ &= \mathbb{E}\left[\sum_{i=1}^n \lambda_i (Z(\mathbf{x}_i) - \mu)\right]^2 \\ &= \left(\sum_{i=1}^n \lambda_i \mathbb{E}[Z(\mathbf{x}_i) - \mu]\right)^2 \\ &= \sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j \mathbb{E}[Z(\mathbf{x}_i) - \mu] \mathbb{E}[Z(\mathbf{x}_j) - \mu] \\ &= \sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j C(\mathbf{x}_i - \mathbf{x}_j) \end{aligned}$$

Por lo tanto,

$$\sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j C(\mathbf{x}_i - \mathbf{x}_j) \geq 0 \quad (2.3)$$

Una función  $C(\mathbf{h})$  que satisface la condición (2.3) para cualquier elección de pesos  $\lambda_i$  y para cualesquiera sitios muestrales  $\mathbf{x}_i$  se dice que es *definida positiva*

(Armstrong, 1998).

Cuando la función  $Z(\mathbf{x})$  es en cambio solamente intrínseca, la situación cambia ligeramente. En este caso, Armstrong (1998) define una combinación lineal “admisibles” como aquella cuyos pesos  $\lambda_i$  suman cero. Dada una combinación admisible

$$Z^* = \sum_{i=1}^n \lambda_i Z(\mathbf{x}_i),$$

se puede elegir un punto arbitrario  $\mathbf{x}_0$  para obtener

$$Z^* = \sum_{i=1}^n \lambda_i [Z(\mathbf{x}_i) - Z(\mathbf{x}_0)].$$

De esta forma se reescribe  $Z^*$  como combinación lineal de incrementos, lo que implica, según la definición de estacionariedad intrínseca, que cada uno tiene esperanza nula y por consiguiente  $\mathbb{E}[Z^*] = 0$ . Por lo tanto:

$$\begin{aligned} \text{Var}\{Z^*\} &= \mathbb{E}\left[\left(\sum_{i=1}^n \lambda_i [Z(\mathbf{x}_i) - Z(\mathbf{x}_0)]\right)^2\right] \\ &= \sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j \mathbb{E}[(Z(\mathbf{x}_i) - Z(\mathbf{x}_0))(Z(\mathbf{x}_j) - Z(\mathbf{x}_0))] \end{aligned} \quad (2.4)$$

Para desarrollar este resultado, utilizamos la identidad:

$$\begin{aligned} (Z(\mathbf{x}_i) - Z(\mathbf{x}_j))^2 &= (Z(\mathbf{x}_i) - Z(\mathbf{x}_0) + Z(\mathbf{x}_0) - Z(\mathbf{x}_j))^2 \\ &= (Z(\mathbf{x}_i) - Z(\mathbf{x}_0))^2 + (Z(\mathbf{x}_j) - Z(\mathbf{x}_0))^2 \\ &\quad - 2(Z(\mathbf{x}_i) - Z(\mathbf{x}_0))(Z(\mathbf{x}_j) - Z(\mathbf{x}_0)) \end{aligned}$$

Como consecuencia se tiene que

$$\begin{aligned} 2\mathbb{E}[(Z(\mathbf{x}_i) - Z(\mathbf{x}_0))(Z(\mathbf{x}_j) - Z(\mathbf{x}_0))] &= \mathbb{E}[(Z(\mathbf{x}_i) - Z(\mathbf{x}_0))^2] \\ &\quad + \mathbb{E}[(Z(\mathbf{x}_j) - Z(\mathbf{x}_0))^2] \\ &\quad - \mathbb{E}[(Z(\mathbf{x}_i) - Z(\mathbf{x}_j))^2] \end{aligned}$$

De nuevo, dada la hipótesis de estacionariedad intrínseca, se tiene que existe el variograma  $\gamma(\mathbf{h})$  de la función aleatoria. En consecuencia:

$$\mathbb{E}[(Z(\mathbf{x}_i) - Z(\mathbf{x}_0))(Z(\mathbf{x}_j) - Z(\mathbf{x}_0))] = \gamma(\mathbf{x}_i - \mathbf{x}_0) + \gamma(\mathbf{x}_j - \mathbf{x}_0) - \gamma(\mathbf{x}_i - \mathbf{x}_j)$$

Por último, sustituyendo en la ecuación (2.4) se obtiene que

$$\begin{aligned}
 \text{Var}\{Z^*\} &= \sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j (\gamma(\mathbf{x}_i - \mathbf{x}_0) + \gamma(\mathbf{x}_j - \mathbf{x}_0) - \gamma(\mathbf{x}_i - \mathbf{x}_j)) \\
 &= \left( \sum_{i=1}^n \lambda_i \gamma(\mathbf{x}_i - \mathbf{x}_0) \right) \left( \sum_{j=1}^n \lambda_j \right) + \left( \sum_{j=1}^n \lambda_j \gamma(\mathbf{x}_j - \mathbf{x}_0) \right) \left( \sum_{i=1}^n \lambda_i \right) \\
 &\quad - \sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j \gamma(\mathbf{x}_i - \mathbf{x}_j) \\
 &= - \sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j \gamma(\mathbf{x}_i - \mathbf{x}_j).
 \end{aligned}$$

De forma que

$$- \sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j \gamma(\mathbf{x}_i - \mathbf{x}_j) \geq 0$$

sin importar la elección de los pesos  $\lambda_i$  y los sitios muestrales  $x_i$ . De hecho, se dice que la función  $-\gamma$  es condicionalmente definida positiva si cumple esta propiedad.

### 2.3. Etapas del estudio geoestadístico

El estudio geoestadístico se divide en las siguientes tres etapas:

#### 1. Análisis preliminar de los datos.

En el kriging es importante apreciar, a grandes rasgos, el comportamiento de los datos a través del espacio para detectar, por ejemplo, posibles valores atípicos en relación a otros valores cercanos, ya que los variogramas son sensibles a la aparición de valores atípicos, y por lo tanto afectan en la estimación de la función  $C(\mathbf{h})$  o  $\gamma(\mathbf{h})$ , según el caso. Por lo tanto, la primera etapa del análisis geoestadístico (de la que se habla en la sección 2.5) consiste en un análisis exploratorio de las mediciones. En esta fase se utilizan herramientas que comúnmente se hallan en análisis estadísticos sencillos como histogramas o box-plot.

#### 2. Estimación y modelación del variograma.

En esta parte del proceso se pretende estimar una función que describa la correlación espacial del fenómeno, es decir, qué tan parecido se distribuye el fenómeno en dos puntos de acuerdo a la distancia que hay entre ellos. Posteriormente, se intenta “suavizar” lo más posible los valores estimados del variograma para encontrar una función analítica que los represente de forma satisfactoria. Este proceso se describe en la sección 2.6.

### 3. Interpolación espacial usando kriging.

Los parámetros obtenidos en los pasos anteriores se usaron para, finalmente, llevar a cabo el kriging, el método de interpolación de la geostatística. A partir del modelo de variograma obtenido, se construye un estimador insesgado de varianza mínima para cada punto no muestreado. La manera de construir dicho estimador se detalla en la sección 2.7.

## 2.4. Determinación de las distancias

Al modelar correlaciones espaciales, es importante calcular distancias de forma conveniente, por lo que antes de realizar un análisis geoespacial es necesario determinar la métrica que será utilizada (Diggle y Ribeiro, 2007).

### Distancia euclidiana

En algunos casos es posible utilizar por defecto la euclidiana entre dos puntos de  $\mathbb{R}^2$  para este rubro.

**Definición 2.4.1.** Sea  $n > 0$  un número natural y sean  $\mathbf{X} = (x_1, x_2, \dots, x_n)$  y  $\mathbf{Y} = (y_1, y_2, \dots, y_n)$  dos puntos de  $\mathbb{R}^n$ . La distancia euclidiana entre los puntos  $\mathbf{X}$  y  $\mathbf{Y}$  se define como

$$d_E(\mathbf{X}, \mathbf{Y}) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \dots + (x_n - y_n)^2} = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$$

No obstante, existen casos donde resulta más conveniente el uso de métricas no euclidianas para realizar análisis geoespacial. Rathbun (1998), por ejemplo, afirma que la distancia euclidiana no es la mejor opción para analizar regiones no convexas con forma irregular.

Es importante señalar que los grados de la longitud y la latitud sólo representan distancias aproximadamente iguales en puntos cercanos al ecuador. En la base de datos MWTP las coordenadas están dadas en grados de longitud y latitud, por lo cual no es posible usar directamente la distancia euclidiana sobre las coordenadas.

### Métrica del Gran Círculo

Cuando la región a estudiar es suficientemente extensa, la métrica del gran círculo resulta más útil para no omitir la geodesia de la Tierra (Diggle y Ribeiro, 2007). Bajo el supuesto de que la Tierra es una esfera de radio  $r = 6378$  kilómetros, dados dos puntos  $x_1$  y  $x_2$  representados por las coordenadas  $(\theta_1, \phi_1)$  y  $(\theta_2, \phi_2)$ , respectivamente, la distancia entre  $x_1$  y  $x_2$  respecto al gran círculo es

$$r \cos^{-1}\{\sin \phi_1 \sin \phi_2 + \cos \phi_1 \cos \phi_2(\theta_1 - \theta_2)\},$$

como puede leerse también en Diggle y Ribeiro (2007).

### Sistema de referencia WGS-84

En este análisis se utilizó el sistema de referencia WGS-84 (World Geodetic System 1984), que toma en cuenta que la forma de la Tierra no es perfectamente esférica, sino que tiene forma de elipsoide. El citado sistema se define a partir de cuatro parámetros (National Imagery and Mapping Agency, 2000) mostrados en la **Tabla 2.1**

Tabla 2.1: Parámetros que definen el sistema de referencia WGS-84

Parámetro	Notación	Valor
Recíproco del Achatamiento	f	298.257223563
Semieje Mayor	a	6378137.0m
Velocidad Angular	$\omega$	$7292115.0 \times 10^{-11} \text{ rad/s}$
Producto de la Constante Gravitacional y la Masa de la Tierra	GM	$3986004.418 \times 10^8 \text{ m}^3/\text{s}^2$

En R se puede usar esta métrica a través de la función `spDistsN1`, definida en la librería `sp` (Pebesma y Bivand, 2017). Dadas las coordenadas  $(\theta_1, \phi_1)$  y  $(\theta_2, \phi_2)$  de dos locaciones terrestres, la función `spDistsN1` permite decidir entre usar la distancia euclídea o la distancia según la métrica del sistema de referencia WGS-84, ambas expresadas en kilómetros.

Aunque la región considerada no es tan grande y se pudo haber usado la distancia euclídea, se usó la distancia del WGS84 por ser más precisa y por ser el estándar en estudios geoestadísticos.

## 2.5. Análisis exploratorio de los datos

En el estudio geoestadístico, el análisis exploratorio preliminar permite construir un entendimiento básico de la información del modelo geoestadístico. En esta sección comúnmente se elaboran pruebas y gráficos característicos de la estadística descriptiva, como histogramas y box-plots. Éstos permiten conocer el comportamiento global de los datos y detectar valores atípicos que puedan causar una interpretación errónea del fenómeno en estudio. Para descartar automáticamente este tipo de datos se tomó en cuenta el rango intercuartílico como criterio. Consideremos el conjunto de datos  $\{y_i\}$ , y respecto a este conjunto, sea  $Q_1$  su primer cuartil,  $Q_3$  su tercer cuartil,  $I_Q = Q_3 - Q_1$  su rango

intercuartílico, y  $r$  un real positivo (generalmente se toma  $r$  como un número entre 1.5 y 3). Se catalogarán como datos atípicos aquellos fuera del rango

$$[Q_1 - r \cdot I_Q, Q_3 + r \cdot I_Q].$$

Es posible descartar del estudio los datos fuera de rango de forma definitiva, o bien omitirlos únicamente durante la estimación y modelación del variograma para incluirlos nuevamente en la etapa de kriging. (Oliver y Webster, 2015) Es importante estudiar también la posible causa de un valor atípico, que puede deberse a errores en la medición, o bien a accidentes geográficos.

Cabe señalar también que, aunque el análisis geoestadístico no requiere necesariamente que los datos obedezcan a una distribución normal, un variograma está compuesto de secuencias de varianzas, las cuales pueden llegar a ser inestables si los datos del modelo contienen valores atípicos o son considerablemente asimétricos.

Cuando los datos no tienen una distribución parecida a la normal y tienen un coeficiente de asimetría fuera de los límites  $\pm 1$ , puede ser conveniente transformarlos y analizar si los nuevos datos se aproximan mejor a una distribución cercana a la normal, por ejemplo, tomando sus logaritmos y examinando el variograma resultante tanto de los datos crudos y de los transformados. En algunos casos, los datos transformados resultarán en un variograma sustancialmente distinto, mientras que en otros casos no. Es necesario tomar esto en cuenta para saber si será necesario trabajar con los datos transformados, ya que si no cambian en gran medida el variograma, es más conveniente utilizar los datos originales (Oliver y Webster, 2015).

En este estudio, aunque en algunos casos no se cumplía con el rango de asimetría buscado, la transformación de los datos en ningún caso aportó un cambio significativo y por lo tanto se utilizó la información original.

## 2.6. Estimación y modelación del variograma.

Para encontrar el estimador lineal óptimo para valores de una función aleatoria se necesita modelar la correlación espacial de las mediciones conocidas. La herramienta que se usa para ello es el variograma.

Se hace una distinción entre los tres tipos de variogramas que utilizan Diggle y Ribeiro (2007):

- *Variograma teórico.* Es la función que, idealmente, modela la correlación espacial de los datos. Esta función se desea aproximar usando los datos empíricos.

- *Variograma empírico.* Representa una primera aproximación del variograma teórico. Para obtenerlo, se utilizan estimadores insesgados de las semivarianzas.
- *Variograma muestral.* Es una versión suavizada del variograma empírico donde las semivarianzas se agrupan y se promedian para aproximar mejor el variograma teórico.

### Variograma Teórico

El variograma es la herramienta fundamental para la interpretación estructural del fenómeno regionalizado y, como se había mencionado, modela la correlación espacial de los datos. En un proceso estocástico espacial, el variograma teórico está dado por la función

$$V(\mathbf{x}, \mathbf{x}') = \frac{1}{2} \text{Var}\{Z(\mathbf{x}) - Z(\mathbf{x}')\}.$$

**Proposición 2.6.1.** En el caso particular donde la función  $Z(\mathbf{x})$  cumple la hipótesis de estacionariedad intrínseca se obtiene el variograma definido en (2.2).

*Demostración.* Sea  $\mathbf{h} = \mathbf{x}' - \mathbf{x}$ .

$$\begin{aligned} V(\mathbf{x}, \mathbf{x} + \mathbf{h}) &= \frac{1}{2} \text{Var}\{Z(\mathbf{x}) - Z(\mathbf{x} + \mathbf{h})\} \\ &= \frac{1}{2} \mathbb{E} [(Z(\mathbf{x}) - Z(\mathbf{x} + \mathbf{h}))^2] - \frac{1}{2} \mathbb{E} [Z(\mathbf{x}) - Z(\mathbf{x} + \mathbf{h})]^2 \\ &= \frac{1}{2} \mathbb{E} [(Z(\mathbf{x}) - Z(\mathbf{x} + \mathbf{h}))^2] \\ &= \gamma(\mathbf{h}) \end{aligned}$$

□

Consecuentemente, el variograma de un fenómeno intrínseco queda definido como

$$\gamma(\mathbf{h}) = \frac{1}{2} \mathbb{E} [(Z(\mathbf{x}) - Z(\mathbf{x} + \mathbf{h}))^2].$$

### Propiedades principales del variograma

Existen algunos parámetros y características que definen al variograma, descritos a continuación y mostrados en la Figura 2.1.

- Como  $Z(\mathbf{x} + \mathbf{h}) = Z(\mathbf{x})$  cuando  $\mathbf{h} = (0, 0) = \mathbf{0}$ , se define  $\gamma(\mathbf{0}) = 0$ .
- En general, se espera que el valor de  $\gamma(\mathbf{h})$  incremente cuando incrementa  $\mathbf{h}$ . La pendiente de dicho incremento muestra el cambio de la dependencia espacial entre los puntos muestrales conforme crece la distancia entre ellos.

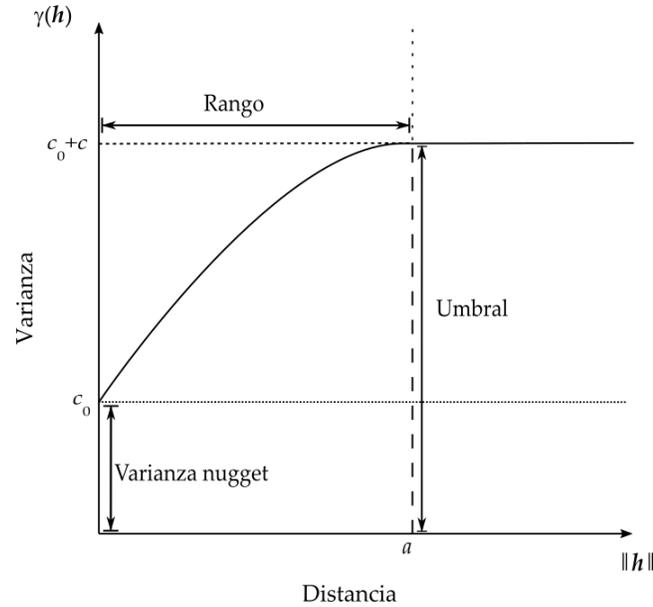


Figura 2.1: Parámetros del variograma

- El *umbral* o *varianza sill*. Si el proceso es estacionario de segundo orden, el variograma se aproxima a una cota superior, conocida como el *umbral*, conforme aumenta  $\mathbf{h}$ . Algunos variogramas alcanzan su umbral de forma estricta, mientras que otros se acercan a él de forma asintótica.
- El *rango* de correlación espacial. Si el variograma alcanza su umbral a una distancia finita, se dice que éste tiene un *rango*, que es la distancia a partir de la cual las observaciones del fenómeno estudiado no están correlacionadas entre sí. Es decir, si los puntos  $\mathbf{x}_i$  y  $\mathbf{x}_j$  están a una distancia mayor que el rango del variograma, las observaciones  $Y_i$  y  $Y_j$  son variables independientes.

Cuando el variograma alcanza su umbral asintóticamente, no posee un rango definido. Para estos casos se utiliza un rango efectivo en el que el variograma alcanza el 0.95 del valor de su umbral.

- Un variograma no acotado describe un proceso que no es estacionario de segundo orden, y que por lo tanto no tiene covarianza. En este caso, no obstante, el variograma existe cuando cumple la hipótesis intrínseca.
- *Varianza nugget* o *pepita*.<sup>1</sup> Aunque, según la teoría, el variograma  $\gamma(\mathbf{h})$  debería aproximarse al origen conforme  $\|\mathbf{h}\| \rightarrow 0$ , en la práctica es más

<sup>1</sup>Este término fue acuñado en la minería. Las pepitas de oro (*gold nuggets*, en inglés) parecían estar distribuidas independientemente una de la otra y por lo tanto representaban una discontinuidad al origen en la variación. (Oliver y Webster, 2015)

común que el variograma se aproxime a un valor positivo  $c_0$  conocido como la *varianza nugget*. La varianza nugget puede representar una discontinuidad en la variación o bien un error de medición.

- **Anisotropía.** Cuando el variograma es calculado en diferentes direcciones, a veces se comporta distinto según la dirección. A este fenómeno se le conoce como *anisotropía*. De no ser así, entonces el variograma es isotrópico y depende únicamente de la magnitud de la distancia entre los puntos. En este caso se dice que el variograma es *omnidireccional*.

### Comportamiento cerca del origen

En el apartado anterior se describieron algunos parámetros que describen del variograma para valores grandes de  $\mathbf{h}$ . No obstante, también es importante analizar el comportamiento del variograma para los valores pequeños ya que está relacionado con la continuidad y la regularidad espacial de la variable regionalizada. Armstrong (1998) describe cuatro tipos de comportamiento cerca del origen:

- **Cuadrático.** Este tipo de comportamiento indica que la variable regionalizada es continua y diferenciable.
- **Lineal.** En este caso la variable regionalizada es continua mas no diferenciable, por lo que es menos regular que en el caso anterior.
- **Discontinua en el origen.** Esto implica que  $\gamma(\mathbf{h})$  no tiende a cero cuando  $\mathbf{h}$  tiende a cero, lo que significa que la variable es altamente irregular a distancias cortas.
- **Plana.** Esto corresponde a cuando las variables  $Z(\mathbf{x}+\mathbf{h})$  y  $Z(\mathbf{x})$  no están correlacionadas para todos los valores de  $\mathbf{h}$  sin importar qué tan cerca estén.

En la Figura 2.2 se muestran los cuatro tipos de comportamientos.

### Variograma Empírico

Basándonos en los datos geostatísticos  $\{(\mathbf{x}_i, y_i) : i = 1, 2, \dots, n\}$ , es necesario dar una primera aproximación del variograma  $\gamma(\mathbf{h})$ . Para dar esta aproximación se admite la hipótesis intrínseca sobre la función aleatoria  $Z(\mathbf{x})$ .

Considérese el vector de diferencia entre dos puntos muestrales del dominio  $\mathbf{u}_{ij} = \mathbf{x}_j - \mathbf{x}_i \in \mathbb{R}^2$ , donde se toma  $1 \leq i < j \leq n$  para considerar los pares de puntos muestrales sin repeticiones. Por definición,

$$\gamma(\mathbf{u}_{ij}) = \frac{1}{2} \mathbb{E} [(Z(\mathbf{x}) - Z(\mathbf{x} + \mathbf{u}_{ij}))^2].$$

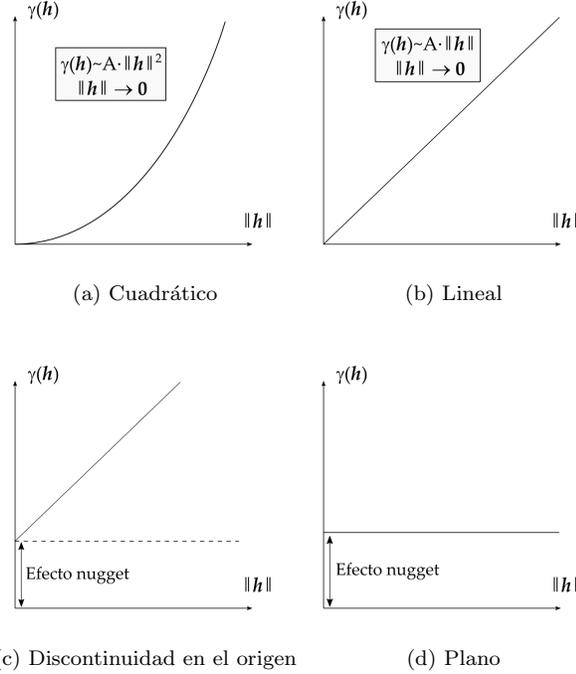


Figura 2.2: Tipos de comportamiento del variograma para valores de  $\mathbf{h}$  cerca de cero.

No obstante, la hipótesis intrínseca permite sustituir  $\mathbf{x}$  por cualquier punto del dominio  $\mathcal{D}$ . En particular, al sustituir  $\mathbf{x} = \mathbf{x}_i$  se obtiene que:

$$\begin{aligned}
 \gamma(\mathbf{u}_{ij}) &= \frac{1}{2} \mathbb{E} [(Z(\mathbf{x}_i) - Z(\mathbf{x}_i + \mathbf{u}_{ij}))^2] \\
 &= \frac{1}{2} \mathbb{E} [(Z(\mathbf{x}_i) - Z(\mathbf{x}_j))^2] \\
 &= \frac{1}{2} \mathbb{E} [(Y_i - Y_j)^2].
 \end{aligned}$$

A partir de este hecho, se define el estimador  $v_{ij} = \frac{1}{2}(Y_i - Y_j)^2$ . Se conoce como *variograma empírico* a la colección de pares ordenados

$$\{(\|\mathbf{u}_{ij}\|, v_{ij}) : 1 \leq i < j \leq n\}.$$

Diggle y Ribeiro (2007) hacen la observación de que, bajo el supuesto de la hipótesis intrínseca, el variograma teórico  $\gamma(\mathbf{h})$  queda definido en  $\mathbf{u}_{ij}$  como

$$\gamma(\mathbf{u}_{ij}) = \mathbb{E} [v_{ij}].$$

Esto implica que las cantidades observadas  $v_{ij} = \frac{1}{2}(Y_i - Y_j)^2$  son estimaciones insesgadas de la semivarianza  $\gamma(\mathbf{u}_{ij})$  del variograma teórico.

De esta forma, considerando que hay 32 estaciones climatológicas, y si se toma en cuenta solamente una lectura por estación, puede obtenerse un máximo de

$$\binom{32}{2} = \frac{32 \cdot 31}{2} = 496$$

estimaciones. Sin embargo, como se menciona más adelante en la sección 2.8, el número de lecturas disponibles fue menor en la práctica.

A manera de ejemplo ilustrativo, las parejas ordenadas que conforman el variograma correspondiente a las temperaturas registradas el 3 de agosto de 2016 de las 18:00 a las 18:15 hrs. se pueden apreciar en un gráfico en la Figura 2.3.

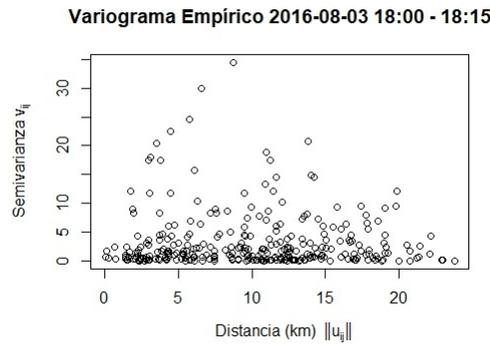


Figura 2.3: Variograma empírico de las medias de las temperaturas registradas el 3 de agosto de 2016 de las 18:00 a las 18:15 hrs.

### Variograma Muestral

Conviene recordar que se espera que  $\gamma(\mathbf{h})$  sea una función suave que depende del vector  $\mathbf{h}$ , por lo que se requiere presentar una versión “suavizada” del variograma empírico que permita aproximarla mejor. Una primer alternativa consiste en tomar en cuenta el promedio de las estimaciones cuyas distancias se repiten.

Sea  $\mathbf{h} = \mathbf{u}_{ij}$  para ciertas  $1 \leq i < j \leq n$ . Para aproximar  $\gamma(\mathbf{h})$  se toma el promedio de las estimaciones de los pares de  $\{\mathbf{x}_k\}_{k=1}^n$  que se hallan a distancia  $\mathbf{h}$ . Es decir, si se define  $s(\mathbf{h})$  como el número de parejas cuyas distancias entre sí es  $\mathbf{h}$  y se supone que  $\mathbf{x}_i, \mathbf{x}_i + \mathbf{h} \in \{\mathbf{x}_k\}_{k=1}^n$  para cada  $i = 1, 2, \dots, s(\mathbf{h})$ , el estimador del variograma evaluado en  $\mathbf{h}$  sería

$$\hat{\gamma}(\mathbf{h}) = \frac{1}{2s(\mathbf{h})} \sum_{i=1}^{s(\mathbf{h})} \{Z(\mathbf{x}_i) - Z(\mathbf{x}_i + \mathbf{h})\}^2. \quad (2.5)$$

**Observación:** el estimador (2.5) también es insesgado puesto que

$$\begin{aligned} \mathbb{E}[\hat{\gamma}(\mathbf{h})] &= \mathbb{E} \left[ \frac{1}{2s(\mathbf{h})} \sum_{i=1}^{s(\mathbf{h})} \{Z(\mathbf{x}_i) - Z(\mathbf{x}_i + \mathbf{h})\}^2 \right] \\ &= \frac{1}{s(\mathbf{h})} \sum_{i=1}^{s(\mathbf{h})} \frac{1}{2} \mathbb{E}[(Z(\mathbf{x}_i) - Z(\mathbf{x}_i + \mathbf{h}))^2] \\ &= \frac{1}{s(\mathbf{h})} \sum_{i=1}^{s(\mathbf{h})} \gamma(\mathbf{h}) \\ &= \gamma(\mathbf{h}) \end{aligned}$$

Cuando los datos geoestadísticos se obtienen de una retícula regular, se puede esperar que haya suficientes repeticiones para llevar a cabo este procedimiento. No obstante, si los puntos  $\{\mathbf{x}_i\}$  están dispuestos de forma irregular, es poco probable encontrar distintas parejas de puntos a la misma distancia. En este último caso, se procede a aproximar  $\gamma(\mathbf{u})$  para cierto  $\mathbf{u} \in \mathfrak{D}$  tomando el promedio de los valores  $v_{ij}$  cuyas distancias correspondientes  $\mathbf{u}_{ij}$  entren en un intervalo  $(\|\mathbf{u}\| - \frac{d}{2}, \|\mathbf{u}\| + \frac{d}{2}]$  de longitud  $d > 0$ .

De esta forma, siendo  $k$  un entero positivo, se define la ordenada  $V_k$  del *variograma muestral* con intervalos de longitud  $d$  como la media de los valores  $v_{ij}$  que satisfacen  $(k-1)d < \|\mathbf{u}_{ij}\| \leq kd$ . Si se considera el punto medio del intervalo  $u_k = (k - \frac{1}{2})d$ , entonces  $V_k$  es aproximadamente una estimación insesgada de  $\gamma(u_k)$ .

Para este procedimiento generalmente se considera una cantidad mínima de diez intervalos, mientras que la distancia a tomar en cuenta en el variograma muestral va de un tercio de la distancia máxima de separación

$$d_{\text{máx}} = \max_{1 \leq i < j \leq n} \{\mathbf{u}_{ij}\}$$

a la mitad de ésta. Es decir, el tamaño de los intervalos debe ser aproximadamente  $h = d_{\text{máx}}/20$ . Por un lado, una cantidad menor de intervalos dificulta la modelación del variograma y, por otro lado, si se utilizan distancias superiores a  $d_{\text{máx}}/2$  para modelar el variograma, se corre el riesgo de obtener una estimación errónea y sesgada debido a la pequeña cantidad de valores  $v_{ij}$  a esa distancia  $\mathbf{u}_{ij}$  de separación (Díaz-Viera et al., 2009).

Dentro de la totalidad de las estaciones, la distancia máxima entre ellas es de 35.390 km, dada entre la estación en el Aeropuerto Internacional Mariano

Escobedo y otra ubicada en la zona de la Huasteca en Santa Catarina. Sin embargo, la distancia máxima en cada variograma puede variar puesto que durante ciertos periodos no todas las estaciones registran lecturas.

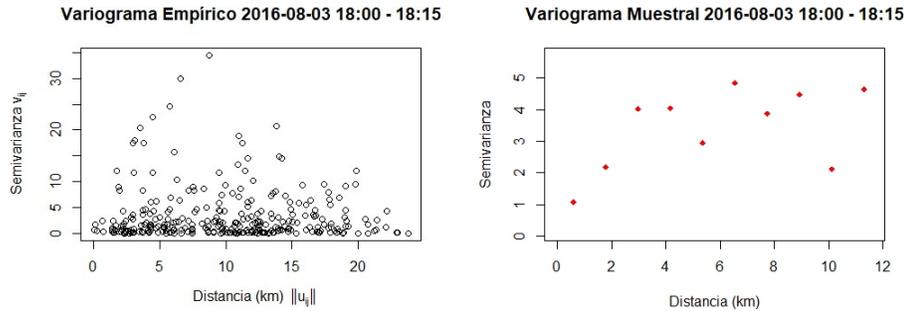


Figura 2.4: Variogramas empírico (izq.) y muestral (der.) de las medias de las temperaturas registradas el 3 de agosto de 2016 de las 18:00 a las 18:15 hrs.

En la Figura 2.4 aparece el mismo variograma de la Figura 2.3, con su variograma muestral correspondiente. Es a partir de esta última estimación que se puede hacer un modelo de ajuste que aproxime al variograma teórico. En este ejemplo, la distancia máxima registrada entre estaciones activas fue de  $d_{\text{máx}} = 23.810$  km, por lo que el variograma muestral abarca hasta una separación de  $\frac{d_{\text{máx}}}{2} = 11.905$  km.

### Modelos admisibles de variograma

Como siguiente paso es necesario elegir un modelo de ajuste admisible de variograma que cumpla con las características ya mencionadas del variograma teórico, dado que es a partir de este modelo de ajuste con el que se hacen estimaciones cuando se lleva a cabo el kriging. Armstrong (1998) resalta que, para que un modelo de variograma sea admisible para la modelación geoestadística, debe de ser condicionalmente definido positivo, así como una función de covarianza debe de ser definida positiva (ver la sección 2.2). Más adelante se describen algunos modelos admisibles comunes de variograma.

Cabe señalar que existen dos tipos principales de modelos de variograma, acotados y no acotados, y a continuación se describen los más utilizados. En ellos,  $a$  denota el rango,  $c$  el umbral del variograma y  $c_0$  es la varianza nugget.

### Modelos acotados

- **Modelo de efecto nugget.** Este modelo indica que el fenómeno regionalizado no tiene correlación especial. Generalmente se utiliza en combinación con algún otro modelo cuando la el fenómeno presenta discontinuidades.

$$\gamma(\mathbf{h}) = \begin{cases} 0 & \text{si } \|\mathbf{h}\| = 0 \\ c_0 & \text{si } \|\mathbf{h}\| > 0 \end{cases}$$

- **Modelo esférico.** Este modelo es uno de los más utilizados. Se caracteriza porque alcanza el umbral en la distancia  $\|\mathbf{h}\| = a$ , por tener comportamiento lineal en el origen y porque describe fenómenos continuos, mas no es derivable para todo  $\mathbf{h}$ .

$$\gamma(\mathbf{h}) = \begin{cases} 0 & \text{si } \|\mathbf{h}\| = 0 \\ c_0 + c \left[ \frac{3\|\mathbf{h}\|}{2a} - \frac{1}{2} \left( \frac{\|\mathbf{h}\|}{a} \right)^3 \right] & \text{si } 0 < \|\mathbf{h}\| \leq a \\ c_0 + c & \text{si } \|\mathbf{h}\| > a \end{cases}$$

- **Modelo exponencial.** Esta función se aproxima a su umbral de manera asintótica y por lo tanto no tiene un rango estricto. Para fines prácticos se le suele atribuir un *rango efectivo*  $a'$  equivalente a  $3a$  (Armstrong, 1998). Al igual que el esférico, este modelo es lineal en el origen.

$$\gamma(\mathbf{h}) = c_0 + c \left[ 1 - \exp \left( -\frac{\|\mathbf{h}\|}{a} \right) \right].$$

- **Modelo gaussiano.** Este modelo representa un fenómeno continuo y se comporta de forma cuadrática en el origen. Para fines prácticos se le suele atribuir un rango efectivo  $a'$  equivalente a  $1.73a$  (Armstrong, 1998).

$$\gamma(\mathbf{h}) = c_0 + c \left[ 1 - \exp \left( -\frac{\|\mathbf{h}\|^2}{a^2} \right) \right].$$

La Figura 2.5 muestra una gráfica con algunos de los modelos antes descritos.

### Modelos no acotados

- **Modelos de funciones potencia.** En este modelo no acotado se utilizan los parámetros  $g$  y  $\beta$ . El primero describe la intensidad de la variación y el segundo describe la curvatura. La Figura 2.6 muestra una gráfica con algunos valores distintos de  $\beta$ .

$$\gamma(\mathbf{h}) = g\|\mathbf{h}\|^\beta, \quad 0 < \beta < 2.$$

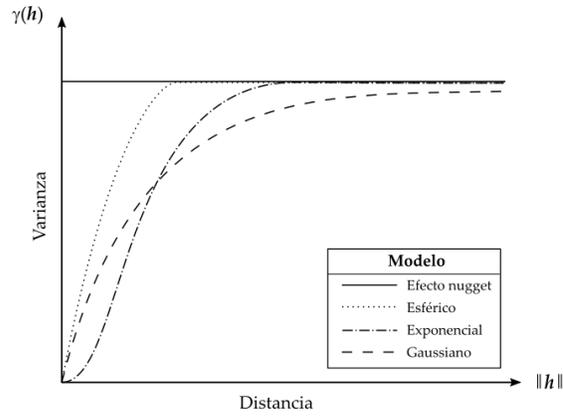
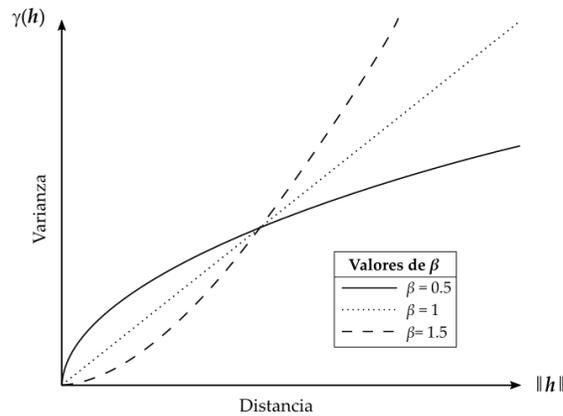


Figura 2.5: Tipos de modelos admisibles acotados.

Figura 2.6: Modelos admisibles de funciones potencia para distintos valores de  $\beta$ .

## 2.7. Kriging

Se conoce como **kriging** al estimado lineal insesgado de varianza mínima para valores de una función aleatoria. La estimación de la función aleatoria se basa en la información conocida de los valores muestrados y de la correlación espacial entre los puntos del dominio. La idea esencial del kriging para estimar un valor desconocido en un punto, está basada en el promedio de las observaciones de los  $n$  valores muestrales conocidos alrededor de él, ponderado de acuerdo a la correlación espacial entre los puntos.

La exactitud de la estimación depende de una variedad de factores entre los

que se incluyen:

- El número de muestras conocidas y la calidad de los datos obtenidos en cada una.
- La distribución de los puntos de muestra.
- La continuidad espacial de la función aleatoria.
- La distancia entre los sitios de muestra y el punto que se desea estimar.

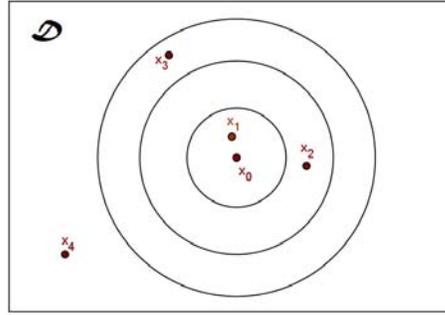


Figura 2.7: Ejemplo de muestras de datos

Por ejemplo, supóngase que en la Figura 2.7 se desea estimar la función aleatoria en el punto  $\mathbf{x}_0$  a partir de las mediciones observadas en los sitios  $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4$ . Al momento de estimar  $Z(\mathbf{x}_0)$ , las muestras de puntos más cercanos a éste brindan más información sobre él que las muestras obtenidas en sitios lejanos. Por lo tanto, en virtud de las distancias desde  $\mathbf{x}_0$  al resto de los puntos con valores conocidos, el estimador de  $Z(\mathbf{x}_0)$  sería

$$Z^*(\mathbf{x}_0) = \lambda_1 Z(\mathbf{x}_1) + \lambda_2 Z(\mathbf{x}_2) + \lambda_3 Z(\mathbf{x}_3) + \lambda_4 Z(\mathbf{x}_4),$$

donde los  $\lambda_i$  son las ponderaciones de los sumandos. En este caso cabría esperarse que

$$\lambda_1 > \lambda_2 > \lambda_3 > \lambda_4.$$

Los parámetros  $\lambda_i$  se definen de acuerdo a los siguientes criterios:

- El estimador  $Z^*(\mathbf{x}_0)$  es insesgado.
- $\text{Var}\{Z^*(\mathbf{x}_0) - Z(\mathbf{x}_0)\}$  es mínima.

Existen diferentes variantes del kriging. En particular, este trabajo se describen el kriging simple, kriging ordinario, y kriging universal. La diferencia entre los tres radica en que la función aleatoria cumpla o no con la hipótesis intrínseca o con las de estacionariedad.

Tabla 2.2: Descripción de los distintos modelos de kriging.

Tipo de kriging	Media	Prerrequisito mínimo
Kriging simple (KS)	Constante y conocida	Covarianza
Kriging ordinario (KO)	Constante y desconocida	Variograma
Kriging universal (KU)	Variante y desconocida	Variograma

### Kriging simple

El kriging simple se basa en la hipótesis de que la variable  $Z(\mathbf{x})$  es estacionaria y tiene media conocida  $m(\mathbf{x})$ . A partir de dichas hipótesis se debe estimar el valor desconocido  $Z(\mathbf{x}_0)$  en la posición  $\mathbf{x}_0$  con base en las observaciones  $Z(\mathbf{x}_1), \dots, Z(\mathbf{x}_n)$  usando el estimador  $Z^*$  definido como:

$$Z^*(\mathbf{x}) = \sum_{i=1}^n \lambda_i Z(\mathbf{x}_i) + \alpha.$$

Tanto los pesos  $\lambda_i$  como el valor de  $\alpha$  varían dependiendo de la posición  $\mathbf{x}$  donde se haga la estimación (i.e. se pueden expresar como  $\lambda_i(\mathbf{x})$  y  $\alpha(\mathbf{x})$ ) y se eligen tratando de minimizar el error cuadrático medio del estimador

$$\mathbb{E} [(Z^*(\mathbf{x}_0) - Z(\mathbf{x}_0))^2] = \text{Var}\{Z^*(\mathbf{x}_0) - Z(\mathbf{x}_0)\} + \mathbb{E} [(Z^*(\mathbf{x}_0) - Z(\mathbf{x}_0))]^2.$$

Como se había mencionado, se busca que el estimador  $Z^*(\mathbf{x}_0)$  sea insesgado, i.e.  $\mathbb{E} [(Z^*(\mathbf{x}_0) - Z(\mathbf{x}_0))] = 0$ . Si ello ocurre, dado que no hay sesgo, el error cuadrático medio se reduce al minimizar la varianza  $\text{Var}\{Z^*(\mathbf{x}_0) - Z(\mathbf{x}_0)\}$ .

Para eliminar el sesgo hay que considerar

$$\begin{aligned} \mathbb{E} [(Z^*(\mathbf{x}_0) - Z(\mathbf{x}_0))] &= 0 \\ \mathbb{E} \left[ \sum_{i=1}^n \lambda_i Z(\mathbf{x}_i) \right] + \alpha - m(\mathbf{x}_0) &= 0 \\ \sum_{i=1}^n \lambda_i m(\mathbf{x}_i) + \alpha - m(\mathbf{x}_0) &= 0 \end{aligned}$$

A partir de ahí, la constante  $\alpha$  puede expresarse en función de la media, conocida por hipótesis, y los parámetros  $\lambda_i$

$$\alpha = m(\mathbf{x}_0) - \sum_{i=1}^n \lambda_i m(\mathbf{x}_i).$$

El estimador puede expresarse ahora como

$$Z^*(\mathbf{x}_0) = m(\mathbf{x}_0) + \sum_{i=1}^n \lambda_i [Z(\mathbf{x}_i) - m(\mathbf{x}_i)]. \quad (2.6)$$

Y considerando  $\lambda_0 = -1$  se obtiene que

$$\begin{aligned} Z^*(\mathbf{x}_0) - Z(\mathbf{x}_0) &= -(Z(\mathbf{x}_0) - m(\mathbf{x}_0)) + \sum_{i=1}^n \lambda_i [Z(\mathbf{x}_i) - m(\mathbf{x}_i)] \\ Z^*(\mathbf{x}_0) - Z(\mathbf{x}_0) &= \sum_{i=0}^n \lambda_i [Z(\mathbf{x}_i) - m(\mathbf{x}_i)]. \end{aligned}$$

En consecuencia, es posible desarrollar el error cuadrático medio de la siguiente forma:

$$\begin{aligned} \mathbb{E} [(Z^*(\mathbf{x}_0) - Z(\mathbf{x}_0))^2] &= \mathbb{E} \left[ \left( \sum_{i=0}^n \lambda_i [Z(\mathbf{x}_i) - m(\mathbf{x}_i)] \right)^2 \right] \\ &= \sum_{i=0}^n \sum_{j=0}^n \lambda_i \lambda_j \mathbb{E} [(Z(\mathbf{x}_i) - m(\mathbf{x}_i)) \cdot (Z(\mathbf{x}_j) - m(\mathbf{x}_j))] \\ &= \sum_{i=0}^n \sum_{j=0}^n \lambda_i \lambda_j \sigma(\mathbf{x}_i, \mathbf{x}_j) \\ &= \sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j \sigma(\mathbf{x}_i, \mathbf{x}_j) - 2 \sum_{i=1}^n \lambda_i \sigma(\mathbf{x}_i, \mathbf{x}_0) + \sigma(\mathbf{x}_0, \mathbf{x}_0) \end{aligned} \tag{2.7}$$

Para minimizarlo, se calculan las derivadas parciales respecto a los parámetros  $\lambda_i$  y se igualan a cero:

$$\frac{\partial}{\partial \lambda_i} \mathbb{E} [(Z^*(\mathbf{x}_0) - Z(\mathbf{x}_0))^2] = 2 \sum_{j=1}^n \lambda_j \sigma(\mathbf{x}_i, \mathbf{x}_j) - 2 \sigma(\mathbf{x}_i, \mathbf{x}_0) = 0$$

De dicha ecuación se deriva el siguiente sistema de ecuaciones, también conocido como *sistema kriging simple*,

$$\sum_{j=1}^n \lambda_j \sigma(\mathbf{x}_i, \mathbf{x}_j) = \sigma(\mathbf{x}_i, \mathbf{x}_0), \quad i = 1, \dots, n.$$

Este sistema de ecuaciones también se puede expresar de forma matricial. Consideremos  $\Sigma$  una matriz, llamada *matriz de covarianzas*, de dimensiones  $n \times n$  con entradas  $\Sigma_{ij} = \sigma_{ij} = \sigma(\mathbf{x}_i, \mathbf{x}_j)$ ,  $\boldsymbol{\sigma}_0$  un vector de tamaño  $n$  donde la  $i$ -ésima entrada representa la covarianza  $\sigma(\mathbf{x}_i, \mathbf{x}_0)$  entre las variables  $Z(\mathbf{x}_i)$  y  $Z(\mathbf{x}_0)$ , y  $\boldsymbol{\lambda}$  un vector de tamaño  $n$  cuya  $i$ -ésima entrada es el peso  $\lambda_i$  de la suma ponderada. De esta forma, el sistema kriging simple puede reescribirse en forma matricial como

$$\Sigma \boldsymbol{\lambda} = \boldsymbol{\sigma}_0.$$

Es decir,

$$\begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & & \sigma_{2n} \\ \vdots & & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_{nn} \end{bmatrix} \cdot \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_n \end{bmatrix} = \begin{bmatrix} \sigma(\mathbf{x}_1, \mathbf{x}_0) \\ \sigma(\mathbf{x}_2, \mathbf{x}_0) \\ \vdots \\ \sigma(\mathbf{x}_n, \mathbf{x}_0) \end{bmatrix}$$

Este sistema de ecuaciones tiene una solución única si la matriz  $\Sigma$  no es singular. Cabe señalar que como la función covarianza  $\sigma$  es positiva definida, y si ningún punto de muestra aparece múltiples veces en el modelo entonces  $\Sigma^{-1}$  existe y por lo tanto el sistema sería soluble.

Por último, es necesario calcular la varianza  $\sigma_{KS}^2$  de la estimación  $Z^*$ , también llamada varianza de kriging simple, que es igual al error cuadrático medio y sirve para medir el error asociado con el estimador kriging. Esto se puede hacer expresando el resultado obtenido en (2.7) en forma matricial y sustituyendo el vector  $\boldsymbol{\lambda}$  por  $\Sigma^{-1}\boldsymbol{\sigma}_0$ .

$$\begin{aligned} \sigma_{KS}^2 &= \sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j \sigma(\mathbf{x}_i, \mathbf{x}_j) - 2 \sum_{i=1}^n \lambda_i \sigma(\mathbf{x}_i, \mathbf{x}_0) + \sigma(\mathbf{x}_0, \mathbf{x}_0) \\ &= \boldsymbol{\lambda}^T \Sigma \boldsymbol{\lambda} - 2 \boldsymbol{\lambda}^T \boldsymbol{\sigma}_0 + \sigma(\mathbf{x}_0, \mathbf{x}_0) \\ &= \boldsymbol{\lambda}^T \Sigma \Sigma^{-1} \boldsymbol{\sigma}_0 - 2 \boldsymbol{\lambda}^T \boldsymbol{\sigma}_0 + \sigma(\mathbf{x}_0, \mathbf{x}_0) \\ &= \sigma(\mathbf{x}_0, \mathbf{x}_0) - \boldsymbol{\lambda}^T \boldsymbol{\sigma}_0 \end{aligned}$$

Conviene señalar que el kriging simple es un interpolador exacto. Es decir que si  $\mathbf{x}_0$  coincide con una de las posiciones de los datos, entonces el valor de la estimación es el mismo que el valor de la muestra obtenida en ese punto. Esto se comprueba a continuación:

**Proposición 2.7.1.** Si  $\mathbf{x}_k$  es un punto de muestreo del modelo geoestadístico, entonces  $Z^*(\mathbf{x}_k) = Z(\mathbf{x}_k)$ , donde  $Z^*(\mathbf{x})$  representa el estimador de kriging simple en la posición  $\mathbf{x}$ .

*Demostración.* Como  $\mathbf{x}_0 = \mathbf{x}_k$ ,

$$\begin{aligned} \boldsymbol{\sigma}_0 &= [\sigma(\mathbf{x}_1, \mathbf{x}_k) \quad \sigma(\mathbf{x}_2, \mathbf{x}_k) \quad \cdots \quad \sigma(\mathbf{x}_n, \mathbf{x}_k)]^T \\ &= [\sigma_{1k} \quad \sigma_{2k} \quad \cdots \quad \sigma_{nk}]^T. \end{aligned}$$

Para  $i, j = 1, 2, \dots, n$  sea  $\hat{\sigma}_{ij} = [\Sigma^{-1}]_{ij}$  la entrada  $i, j$  de la matriz  $\Sigma^{-1}$ . Se calcula el parámetro  $\lambda_i$

$$\lambda_i = [\Sigma^{-1}\boldsymbol{\sigma}_0]_i = \sum_{j=1}^n \hat{\sigma}_{ij} \cdot \sigma_{jk} = [\Sigma^{-1}\Sigma]_{ik} = \delta_{ik},$$

donde  $\delta_{ik}$  es la delta de Kronecker. De acuerdo con (2.6), el estimador en  $\mathbf{x}_k$  sería

$$\begin{aligned} Z^*(\mathbf{x}_k) &= m(\mathbf{x}_k) + \sum_{i=1}^n \delta_{ik} [Z(\mathbf{x}_i) - m(\mathbf{x}_i)] \\ &= m(\mathbf{x}_k) + [Z(\mathbf{x}_k) - m(\mathbf{x}_k)] = Z(\mathbf{x}_k) \end{aligned}$$

□

### Kriging ordinario

Salvo que se cuente con una gran cantidad de repeticiones del fenómeno que permitan aproximarla con exactitud, en la práctica es complicado conocer la media de un fenómeno regionalizado. Es por ello que, en los casos donde la media es desconocida, se utiliza otra variante de kriging conocida como *kriging ordinario*.

Para usar el kriging ordinario se da por hecho que la variable regionalizada  $Z(\mathbf{x})$  es una función aleatoria intrínseca con media constante  $m(\mathbf{x}) = \mu$  desconocida, y cuya varianza depende únicamente de  $\mathbf{h}$  y no de la posición absoluta de los puntos en el dominio. Este estimador requiere únicamente que se conozca el variograma para modelar la correlación entre los datos. Con base en estas hipótesis, el problema del kriging ordinario consiste en estimar el valor de  $Z(\mathbf{x}_0)$  en la posición  $\mathbf{x}_0$  a partir de las observaciones  $Z(\mathbf{x}_1), Z(\mathbf{x}_2), \dots, Z(\mathbf{x}_n)$ .

Al igual que en el caso anterior, el kriging ordinario utiliza el estimador

$$Z^*(\mathbf{x}_0) = \sum_{i=1}^n \lambda_i Z(\mathbf{x}_i) + \alpha,$$

donde, al igual que con el kriging simple, los pesos  $\lambda_i$  y la constante  $\alpha$  se calculan con base en la posición  $\mathbf{x}$  donde se hace la estimación, de forma que  $Z^*(\mathbf{x})$  sea insesgado y tratando de minimizar el error cuadrático medio

$$\mathbb{E} [(Z^*(\mathbf{x}_0) - Z(\mathbf{x}_0))^2] = \text{Var}\{Z^*(\mathbf{x}_0) - Z(\mathbf{x}_0)\} + \mathbb{E} [Z^*(\mathbf{x}_0) - Z(\mathbf{x}_0)]^2.$$

Como primer paso, hay que igualar a cero el sesgo del estimador.

$$\begin{aligned} \mathbb{E}[Z^*(\mathbf{x}_0) - Z(\mathbf{x}_0)] &= \mathbb{E} \left[ \sum_{i=1}^n \lambda_i Z(\mathbf{x}_i) + \alpha - Z(\mathbf{x}_0) \right] \\ &= \sum_{i=1}^n \lambda_i \mu + \alpha - \mu \\ &= \mu \left( \sum_{i=1}^n \lambda_i - 1 \right) + \alpha \end{aligned}$$

Por lo tanto, para que el estimador sea insesgado debe cumplirse que

$$\alpha = \mu \left( 1 - \sum_{i=1}^n \lambda_i \right). \quad (2.8)$$

A partir de (2.8) se observa que la relación entre los parámetros  $\lambda_i$  y  $\alpha$  depende del valor de  $\mu$  que, según los supuestos del kriging ordinario, es desconocido. Una primera opción para resolver este problema sería utilizar un estimador de la media  $\hat{\mu}$ . No obstante, ello implicaría que los parámetros del estimador ya no dependerían únicamente de la correlación espacial de  $Z(\mathbf{x})$ , ya que para calcularlos se utilizarían los datos observados. La única forma de hacer que los parámetros  $\lambda_i$  y  $\alpha$  ya no dependan de  $\mu$  sería entonces que

$$\left( 1 - \sum_{i=1}^n \lambda_i \right) = 0.$$

De esta forma, el problema del kriging ordinario consiste ahora en estimar el valor de  $Z(\mathbf{x}_0)$  en la posición  $\mathbf{x}_0$  usando el estimador

$$Z(\mathbf{x}_0) = \sum_{i=1}^n \lambda_i Z(\mathbf{x}_i), \quad \text{sujeto a } \sum_{i=1}^n \lambda_i = 1,$$

donde  $Z(\mathbf{x}_1), Z(\mathbf{x}_2), \dots, Z(\mathbf{x}_n)$  son las observaciones del modelo geoestadístico.

La condición de la suma de los pesos garantiza que el estimador sea insesgado, dado que su sesgo sería

$$\mathbb{E} [Z^*(\mathbf{x}_0) - Z(\mathbf{x}_0)] = \mathbb{E} \left[ \sum_{i=1}^n \lambda_i Z(\mathbf{x}_i) - Z(\mathbf{x}_0) \right] = \mu \sum_{i=1}^n \lambda_i - \mu = 0.$$

Puesto que es insesgado, el error cuadrático medio equivale a la varianza  $\sigma_{KO}^2$  asociada al estimador kriging ordinario:

$$\sigma_{KO}^2 = \text{Var}\{Z^*(\mathbf{x}_0) - Z(\mathbf{x}_0)\} = \mathbb{E} [(Z^*(\mathbf{x}_0) - Z(\mathbf{x}_0))^2].$$

De esta forma, lo siguiente que se hace es elegir los valores de  $\lambda_i$  que minimicen

la varianza del estimador.

$$\begin{aligned}
\sigma_{KO}^2 &= \mathbb{E} \left[ \left( \sum_{i=1}^n \lambda_i [Z(\mathbf{x}_i) - Z(\mathbf{x}_0)] \right)^2 \right] \\
&= \sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j \mathbb{E} [(Z(\mathbf{x}_i) - Z(\mathbf{x}_0))(Z(\mathbf{x}_j) - Z(\mathbf{x}_0))] \\
&= \sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j \left( \frac{1}{2} \mathbb{E} [(Z(\mathbf{x}_0) - Z(\mathbf{x}_i))^2] + \frac{1}{2} \mathbb{E} [(Z(\mathbf{x}_0) - Z(\mathbf{x}_j))^2] \right. \\
&\quad \left. - \frac{1}{2} \mathbb{E} [(Z(\mathbf{x}_i) - Z(\mathbf{x}_j))^2] \right) \\
&= \sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j [\gamma(\mathbf{x}_0 - \mathbf{x}_i) + \gamma(\mathbf{x}_0 - \mathbf{x}_j) - \gamma(\mathbf{x}_i - \mathbf{x}_j)] \\
&= 2 \sum_{i=1}^n \lambda_i \gamma(\mathbf{x}_0 - \mathbf{x}_i) - \sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j \gamma(\mathbf{x}_i - \mathbf{x}_j) \tag{2.9}
\end{aligned}$$

Ahora el problema consiste en minimizar

$$2 \sum_{i=1}^n \lambda_i \gamma(\mathbf{x}_0 - \mathbf{x}_i) - \sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j \gamma(\mathbf{x}_i - \mathbf{x}_j), \quad \text{sujeto a } \sum_{i=1}^n \lambda_i = 1.$$

Para resolverlo usando un multiplicador de Lagrange  $\rho$ , consideramos

$$\begin{aligned}
f(\boldsymbol{\lambda}) &= 2 \sum_{i=1}^n \lambda_i \gamma(\mathbf{x}_0 - \mathbf{x}_i) - \sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j \gamma(\mathbf{x}_i - \mathbf{x}_j) \\
g(\boldsymbol{\lambda}) &= \sum_{i=1}^n \lambda_i - 1
\end{aligned}$$

Sea  $S$  la curva de nivel de  $g$  con valor 0. Para encontrar un punto crítico de  $f$  restringido a  $S$ , hay que hallar un vector  $\boldsymbol{\lambda}$  y un escalar  $\rho$  de modo que

$$\begin{aligned}
\frac{\partial f}{\partial \lambda_i}(\boldsymbol{\lambda}) + 2\rho \frac{\partial g}{\partial \lambda_i}(\boldsymbol{\lambda}) &= 0 \quad \text{para cada } i = 1, 2, \dots, n \\
g(\boldsymbol{\lambda}) &= 0
\end{aligned}$$

Lo cual da lugar al siguiente sistema de ecuaciones, conocido como *sistema kriging ordinario*.

$$\begin{aligned}
\sum_{j=1}^n \lambda_j \gamma(\mathbf{x}_j - \mathbf{x}_i) + \rho &= \gamma(\mathbf{x}_i - \mathbf{x}_0), \quad \text{para } i = 1, 2, \dots, n \tag{2.10} \\
\sum_{j=1}^n \lambda_j &= 1
\end{aligned}$$

Para determinar la varianza del kriging ordinario se multiplica por  $\lambda_i$  las ecuaciones en (2.10) y se suman sobre  $i$  para obtener:

$$-\sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j \gamma(\mathbf{x}_j - \mathbf{x}_i) = \rho - \sum_{i=1}^n \lambda_i \gamma(\mathbf{x}_i - \mathbf{x}_0)$$

Así, retomando el resultado (2.9), se tiene que

$$\begin{aligned} \sigma_{KO}^2 &= 2 \sum_{i=1}^n \lambda_i \gamma(\mathbf{x}_0 - \mathbf{x}_i) - \sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j \gamma(\mathbf{x}_i - \mathbf{x}_j) \\ &= \sum_{i=1}^n \lambda_i \gamma(\mathbf{x}_0 - \mathbf{x}_i) + \rho \end{aligned}$$

El sistema kriging ordinario también puede ser expresado en forma matricial  $\Gamma \cdot \boldsymbol{\lambda} = \boldsymbol{\gamma}_0$ , donde  $\Gamma$  contiene las semivarianzas entre el  $i$ -ésimo y el  $j$ -ésimo punto muestral,  $\boldsymbol{\lambda}$  es un vector contiene los pesos del estimador, y  $\boldsymbol{\gamma}_0$  es un vector con las semivarianzas entre el punto a estimar y cada uno los puntos muestrales. De forma explícita, el sistema kriging simple en forma matricial sería

$$\begin{bmatrix} \gamma(\mathbf{x}_1 - \mathbf{x}_1) & \cdots & \gamma(\mathbf{x}_1 - \mathbf{x}_n) & 1 \\ \vdots & \ddots & \vdots & \vdots \\ \gamma(\mathbf{x}_n - \mathbf{x}_1) & \cdots & \gamma(\mathbf{x}_n - \mathbf{x}_n) & 1 \\ 1 & \cdots & 1 & 0 \end{bmatrix} \cdot \begin{bmatrix} \lambda_1 \\ \vdots \\ \lambda_n \\ \rho \end{bmatrix} = \begin{bmatrix} \gamma(\mathbf{x}_1 - \mathbf{x}_0) \\ \vdots \\ \gamma(\mathbf{x}_n - \mathbf{x}_0) \\ 1 \end{bmatrix}.$$

Así, se invierte la matriz  $\Gamma$  y se obtienen los pesos  $\lambda_i$  y el multiplicador  $\rho$ :

$$\boldsymbol{\lambda} = \Gamma^{-1} \cdot \boldsymbol{\gamma}_0.$$

Por otra parte, la varianza del estimador expresada en forma matricial sería:

$$\sigma_{KO}^2 = \boldsymbol{\gamma}_0^T \cdot \boldsymbol{\lambda}.$$

Al igual que en el caso de kriging simple, el estimador de kriging ordinario es un interpolador exacto. Es decir, que si se desea interpolar en un punto de muestra  $\mathbf{x}_i$ , entonces el estimador es igual a la observación en ese punto. Dicho de otra forma: si  $\mathbf{x}_0 = \mathbf{x}_i$ , entonces  $Z^*(\mathbf{x}_0) = Z(\mathbf{x}_i)$ . Ello se comprueba de forma similar a como se demostró la Proposición 2.7.1.

### Kriging universal

Las hipótesis de estacionariedad, sobre las que se basan los modelos de kriging anteriores, no siempre resulta satisfactoria. Tal es el caso cuando existe una pronunciada tendencia geográfica. Es decir, que la función aleatoria  $Z^*(\mathbf{x}_0)$  depende ahora de la posición absoluta del punto  $\mathbf{x}_0$ .

La función aleatoria puede representarse como

$$Z(\mathbf{x}) = m(\mathbf{x}) + \varepsilon(\mathbf{x}),$$

donde, como en los casos anteriores,  $m(\mathbf{x})$  es la media del proceso aleatorio, y  $\varepsilon(\mathbf{x})$  es una función aleatoria de media cero correlacionada espacialmente, llamada el residual.

Al igual que con kriging ordinario, la media en kriging universal es desconocida. Sin embargo, en este caso, su valor ya no se considera constante sino que está dado por una función determinista suave

$$m(\mathbf{x}) = \sum_{k=1}^M w_k f_k(\mathbf{x}),$$

donde las funciones  $f_k$  son monomios dados a partir de las coordenadas  $(x_1, x_2)$  de  $\mathbf{x}$ . Por ejemplo, para representar una tendencia lineal se utilizan tres funciones con valores

$$f_0 = 1, \quad f_1 = x_1, \quad \text{y} \quad f_2 = x_2,$$

y para representar una tendencia cuadrática se consideran tres funciones adicionales.:

$$f_3 = x_1^2, \quad f_4 = x_1 x_2, \quad \text{y} \quad f_5 = x_2^2.$$

Para estimar  $Z(\mathbf{x}_0)$  se usa nuevamente el estimador lineal

$$Z^*(\mathbf{x}_0) = \sum_{i=1}^n \lambda_i Z(\mathbf{x}_i),$$

buscando primero que éste sea insesgado y, posteriormente, minimizar su error cuadrático medio.

Desarrollando la media  $m(\mathbf{x})$ , el sesgo del estimador se puede expresar como

$$\begin{aligned} \mathbb{E}[Z^*(\mathbf{x}_0) - Z(\mathbf{x}_0)] &= \sum_{i=1}^n \lambda_i m(\mathbf{x}_i) - m(\mathbf{x}_0) \\ &= \sum_{i=1}^n \lambda_i \sum_{k=1}^M w_k f_k(\mathbf{x}_i) - \sum_{k=1}^M w_k f_k(\mathbf{x}_0) \end{aligned}$$

y cambiando el orden de la sumas, se obtiene

$$\mathbb{E}[Z^*(\mathbf{x}_0) - Z(\mathbf{x}_0)] = \sum_{k=1}^M w_k \left( \sum_{i=1}^N \lambda_i f_k(\mathbf{x}_i) - f_k(\mathbf{x}_0) \right). \quad (2.11)$$

Para eliminar el sesgo del error cuadrático medio de  $Z^*(\mathbf{x}_0)$  sin importar los parámetros  $w_k$  que modelan la media del proceso, se elimina el segundo factor en los sumandos de la ecuación (2.11) usando las siguientes  $M$  condiciones:

$$\sum_{i=1}^N \lambda_i f_k(\mathbf{x}_i) = f_k(\mathbf{x}_0), \quad \text{para } k = 1, 2, \dots, M. \quad (2.12)$$

Siendo insesgado el estimador, se tiene que el error cuadrático medio del estimador kriging universal es equivalente a la varianza. Por lo tanto, su expresión es la misma que se obtuvo en (2.9) para el estimador kriging ordinario:

$$\begin{aligned} \mathbb{E}[(Z^*(\mathbf{x}_0) - Z(\mathbf{x}_0))^2] &= \text{Var}\{Z^*(\mathbf{x}_0) - Z(\mathbf{x}_0)\} \\ &= 2 \sum_{i=1}^n \lambda_i \gamma(\mathbf{x}_0 - \mathbf{x}_i) - \sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j \gamma(\mathbf{x}_i - \mathbf{x}_j). \end{aligned} \quad (2.13)$$

El problema se puede formular entonces de la siguiente forma: Encontrar los  $N$  parámetros  $\lambda_i$  que minimicen  $\text{Var}\{Z^*(\mathbf{x}_0) - Z(\mathbf{x}_0)\}$ , expresada en (2.13), sujeto a las restricciones lineales (2.12).

Este problema puede resolverse utilizando nuevamente el método de multiplicadores de Lagrange. Consideramos

$$\begin{aligned} f(\boldsymbol{\lambda}) &= 2 \sum_{i=1}^n \lambda_i \gamma(\mathbf{x}_0 - \mathbf{x}_i) - \sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j \gamma(\mathbf{x}_i - \mathbf{x}_j) \\ g_k(\boldsymbol{\lambda}) &= \sum_{i=1}^n \lambda_i f_k(\mathbf{x}_i) - f_k(\mathbf{x}_0), \quad \text{para } k = 1, 2, \dots, M. \end{aligned}$$

Para determinar un punto crítico  $\boldsymbol{\lambda}_0$  de  $f$  restringida a las condiciones

$$g_k(\boldsymbol{\lambda}) = 0, \quad \text{para } k = 1, 2, \dots, M$$

hay que hallar un vector  $\boldsymbol{\lambda}$  y escalares  $\boldsymbol{\rho} = (\rho_1, \rho_2, \dots, \rho_M)^T$  tales que

$$\frac{\partial}{\partial \lambda_i} \left( f(\boldsymbol{\lambda}) - 2 \sum_{k=1}^M \rho_k g_k(\boldsymbol{\lambda}) \right) = 0 \quad \text{para } i = 1, 2, \dots, n$$

$$g_k(\boldsymbol{\lambda}) = 0 \quad \text{para } k = 1, 2, \dots, M.$$

Esto resulta en el sistema de  $n+M$  ecuaciones con  $n+M$  incógnitas llamado *sistema kriging universal*.

$$\begin{aligned} \sum_{j=1}^n \lambda_j \gamma(\mathbf{x}_i - \mathbf{x}_j) + \sum_{k=1}^M \rho_k f_k(\mathbf{x}_i) &= \gamma(\mathbf{x}_0 - \mathbf{x}_i) \quad \text{para } i = 1, 2, \dots, n \quad (2.14) \\ \sum_{i=1}^n \lambda_i f_k(\mathbf{x}_i) &= f_k(\mathbf{x}_0) \quad \text{para } k = 1, 2, \dots, M. \end{aligned}$$

En forma matricial, el sistema kriging universal se puede expresar como

$$\begin{bmatrix} \Gamma & F \\ F^T & O \end{bmatrix} \cdot \begin{bmatrix} \boldsymbol{\lambda} \\ \boldsymbol{\rho} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\gamma}_0 \\ \mathbf{f}_0 \end{bmatrix},$$

donde  $O$  es una matriz de tamaño  $M \times M$  con entradas iguales a cero;  $\Gamma$  es la matriz de las semivarianzas con entradas  $\Gamma_{ij} = \gamma(\mathbf{x}_i - \mathbf{x}_j)$  y de tamaño  $n \times n$ ;  $F$  es una matriz de tamaño  $n \times M$  con entradas  $F_{ij} = f_j(\mathbf{x}_i)$ ;  $\boldsymbol{\lambda}$  es el vector de los pesos del estimador;  $\boldsymbol{\rho}$  es el vector de los multiplicadores de Lagrange;  $\boldsymbol{\gamma}_0$  es el vector de las semivarianzas entre el punto a estimar  $\mathbf{x}_0$  y cada uno de los puntos muestrales  $\mathbf{x}_i$ ; por último,  $\mathbf{f}_0$  es de tamaño  $M$  y contiene los monomios  $f_k$  evaluados en el punto  $\mathbf{x}_0$ .

Después de un proceso similar al que se siguió con kriging ordinario, a partir de (2.14) se deriva

$$-\sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j \gamma(\mathbf{x}_i - \mathbf{x}_j) = \sum_{i=1}^n \sum_{k=1}^M \lambda_i \rho_k f_k(\mathbf{x}_i) - \sum_{i=1}^n \lambda_i \gamma(\mathbf{x}_0 - \mathbf{x}_i).$$

Finalmente, sustituyendo en (2.13) se obtiene

$$\begin{aligned} \sigma_{UK}^2 &= \text{Var}\{Z^*(\mathbf{x}_0) - Z(\mathbf{x}_0)\} = \sum_{i=1}^n \lambda_i \gamma(\mathbf{x}_0 - \mathbf{x}_i) + \sum_{k=1}^M \rho_k \sum_{i=1}^n \lambda_i f_k(\mathbf{x}_i) \\ &= \sum_{i=1}^n \lambda_i \gamma(\mathbf{x}_0 - \mathbf{x}_i) + \sum_{k=1}^M \rho_k f_k(\mathbf{x}_0). \end{aligned}$$

## 2.8. Metodología de la aplicación del modelo

Se utilizó la estimación geoestadística para modelar la temperatura, la humedad, y la presión atmosférica del AMM, utilizando los datos disponibles del MWTP. Es importante tener en cuenta que las estaciones no registran información de forma sincronizada ni con la misma frecuencia, y que entonces es necesario tomar todas las mediciones existentes en un intervalo suficientemente grande para aportar una buena cantidad de datos, sin que sea demasiado largo

para que exista mucha variación entre los datos. Para tal propósito se consideró el promedio de las mediciones que registra cada una de las estaciones en un intervalo de quince minutos y a partir de dichos bloques se llevó a cabo el kriging ordinario.

En general, en cada bloque de 15 minutos existen mediciones de alrededor de 20 de las 32 estaciones, por lo que no existen suficientes datos para elaborar un análisis para detectar anisotropía. Por ello se partió de la suposición de que las funciones aleatorias asociadas a cada uno de los parámetros climatológicos cumplen con la hipótesis intrínseca y para el análisis estructural se utilizan variogramas omnidireccionales. Otra suposición importante, motivada por la misma razón, es que el variograma  $\gamma(\mathbf{h})$  es constante en el mismo periodo para un rango suficientemente pequeño de días. Bajo esta suposición se utilizaron las semivarianzas de varios días consecutivos, durante el mismo periodo de quince minutos en cada día, para estimar el variograma.

Durante el análisis preliminar de los datos es importante tener en cuenta la presencia de posibles datos atípicos o fuera de rango que pueden alterar la modelación de la correlación espacial y descartarlos. Este análisis se implementó utilizando un código en R (ver Apéndices A y B). Para establecer un criterio que permita descartar este tipo de información automáticamente se tomó el rango intercuartílico del bloque de datos por día, tal y como se describe en la sección 2.5.

Para el análisis se eligieron dos bloques principales, uno correspondiente a las 17:00 horas y el otro a las 00:00 horas, tomados del 1 al 5 de mayo de 2017, durante la temporada de calor en el AMM. Para cada uno de esos datos se compararon los resultados obtenidos al utilizar los distintos modelos admisibles acotados de variograma que fueron abordados en esta sección: el modelo esférico, el exponencial y el gaussiano. Para aplicar el modelo se desarrolló un código en R que usa las librerías `lubridate` (Spinu, Grolemond, y Wickham, 2017), `sp` (Pebesma y Bivand, 2017), `nlsl` (Mullen y van Stokkum, 2012), `calibrate` (Graffelman, 2013), `ggplot2` (Wickham y Chang, 2016), `ggmap` (Kahle y Wickham, 2016) (ver Apéndices A y C).



## Capítulo 3

# Aplicación de kriging en datos de temperatura

En este capítulo se toman en cuenta los datos recolectados de la temperatura. En la primera sección se incluyen para el análisis preliminar de los datos, sus box-plots y un resumen estadístico, en la segunda se muestra la estimación del variograma del modelo geoestadístico, en la tercera se evalúan los modelos de ajuste y, por último, aparece el resultado del estimador kriging.

Para el análisis preliminar de los datos se elaboraron box-plots y se detectaron valores atípicos en las mediciones. En la sección 3.1, hubo cinco mediciones fuera de rango de un total de 98, mientras que en la sección 3.2 hubo seis de 104. Los criterios para determinar qué valores eran atípicos se menciona en cada sección. Para elaborar un modelo más confiable, éstos se eliminaron provisionalmente de la elaboración de los variogramas, aunque fueron tomados en cuenta más tarde, al momento de realizar el kriging.

Es importante señalar por último que no se tomaron en cuenta las mediciones de la estación ubicada en las coordenadas  $(-100.414^\circ, 25.7415^\circ)$ , al poniente del AMM. La inclusión de esta estación en la elaboración del variograma dificultaba de forma importante la modelación de los datos. En el análisis de datos de temperatura de las 17:00 el modelo de variograma resultaba totalmente en efecto nugget, como lo muestra la Figura 3.1, mientras que en el análisis de datos de temperatura de las 00:00 el modelo de variograma también se ve afectado considerablemente, como aparece en la Figura 3.2.

### 3.1. Análisis del 1 al 5 mayo de 2017, de 17:00 a 17:15

En este análisis se estimaron las temperaturas en cada punto del AMM en el intervalo de las 17:00 a las 17:15 del día primero de mayo de 2017. Para elaborar el modelo de variograma se tomaron en cuenta las semivarianzas del 1 al 5 de mayo de 2017 durante ese mismo bloque de tiempo. Puesto que los

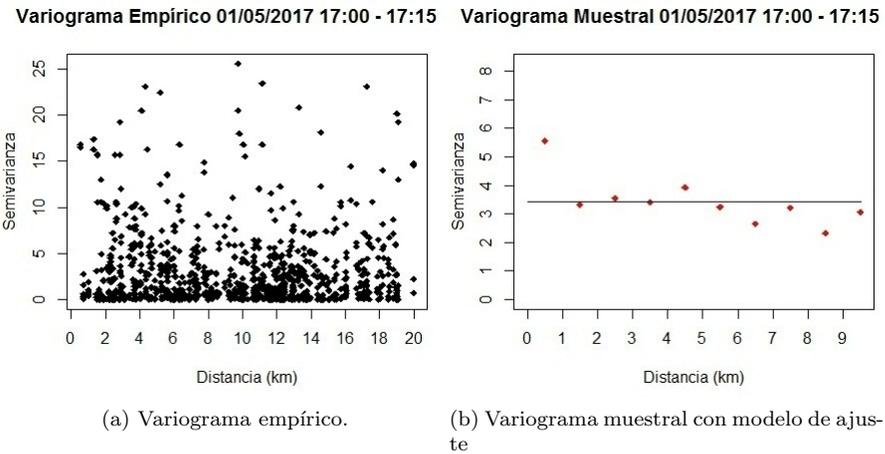


Figura 3.1: Variograma empírico (a) y muestral (b) del 1 al 5 de mayo de 2017, de las 17:00 a las 17:15 horas, incluyendo la estación en las coordenadas  $(-100.414^\circ, 25.7415^\circ)$ .

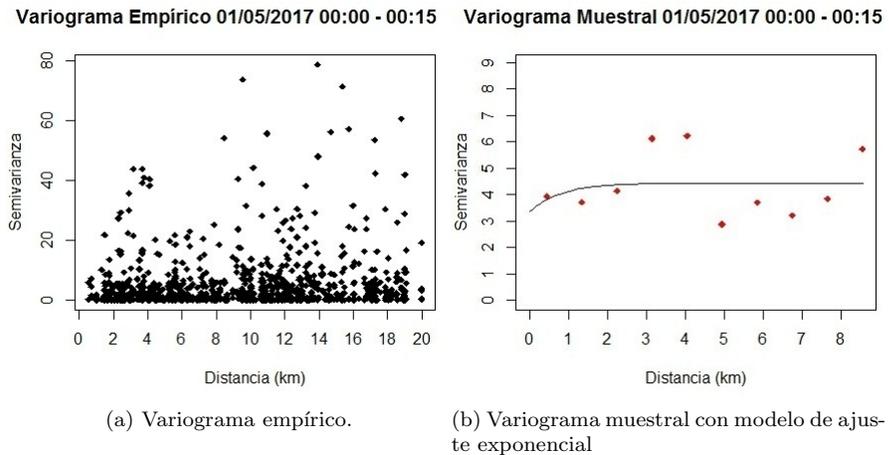


Figura 3.2: Variograma empírico (a) y muestral (b) del 1 al 5 de mayo de 2017, de las 00:00 a las 00:15 horas, incluyendo la estación en las coordenadas  $(-100.414^\circ, 25.7415^\circ)$ .

datos de estos cinco días serán usados en la elaboración del variograma, los cinco bloques serán sujetos al análisis preliminar.

### Análisis preliminar

La Figura 3.3 muestra que los datos extremos de algunos de los días no entran en el margen de 1.5 veces el rango intercuartílico que se estableció para detectar valores atípicos, por lo que en esta primera etapa hay datos que su-

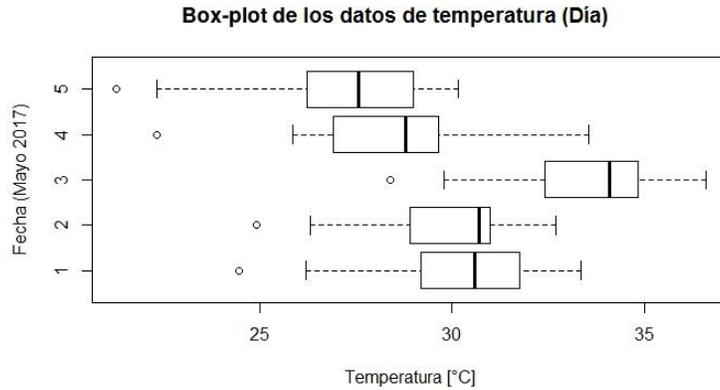


Figura 3.3: Box-plot de las temperaturas registradas del 1 al 5 de mayo de 2017, de las 17:00 a las 17:15 horas.

primir para elaborar el variograma.

Tabla 3.1: Resumen estadístico de las temperaturas registradas del 1 al 5 de mayo de 2017, de las 17:00 a las 17:15 horas.

Mayo 2017 (Día)					
Fecha	1	2	3	4	5
Mínimo	26.20	26.30	29.80	25.85	22.30
Primer cuartil	29.27	29.12	32.69	27.16	26.35
Mediana	30.75	30.75	34.20	28.83	27.60
Media	30.37	30.18	33.80	28.76	27.48
Tercer cuartil	31.80	31.23	34.92	29.64	29.00
Máximo	33.35	32.70	36.60	33.55	30.15
Desviación estándar	1.7757	1.6773	1.8991	2.0967	2.1824
Varianza	3.153014	2.8133	3.6065	4.3961	4.7627
Asimetría	-0.4550	-0.5427	-0.6026	0.6553	-0.8269
Número de datos	20	20	18	16	19

En la Tabla 3.1 aparece el resumen estadístico de los datos de cada día. Se observa que en general el periodo de quince minutos más caluroso fue el del 3 de mayo, en tanto que tiene los registros más altos en los primeros seis rubros del resumen. Es importante también señalar que los coeficientes de asimetría se encuentran todos dentro del rango de  $\pm 1$ , por lo que se procedió al uso de estos datos sin realizarles ningún tipo de transformación adicional.

### Estimación del variograma

La Figura 3.4 muestra el variograma empírico de los datos que se analizan en esta sección, mientras que la Figura 3.5 muestra el variograma muestral que toma en cuenta la cantidad de diez intervalos y que abarca hasta la mitad de la distancia máxima  $d_{\text{máx}} = 20.016$  km, es decir, el variograma muestral toma en cuenta hasta una distancia  $\frac{d_{\text{máx}}}{2} = 10.008$  km. Para elaborar a partir de éste el variograma muestral, se utilizan los datos que aparecen en la Tabla 3.2.

Tabla 3.2: Intervalos de distancia, semivarianzas y conteo de estimaciones de las temperaturas registradas del 1 al 5 de mayo de 2017, de las 17:00 a las 17:15 horas.

Distancia (km)	Semivarianza	Número de estimaciones
0.500	1.629	5
1.501	4.021	31
2.502	3.402	63
3.503	4.535	46
4.504	4.370	48
5.504	3.879	51
6.505	4.646	47
7.506	3.597	40
8.507	4.642	26
9.508	4.072	51

Se observa que la semivarianza del primer intervalo vale alrededor de la mitad de las demás semivarianzas, y que a partir del segundo intervalo, centrado en la distancia  $h = 1.501$  km, las semivarianzas dejan de disminuir, y llegan a estabilizarse. Por lo tanto, es probable que el rango del variograma tenga un valor parecido a éste.

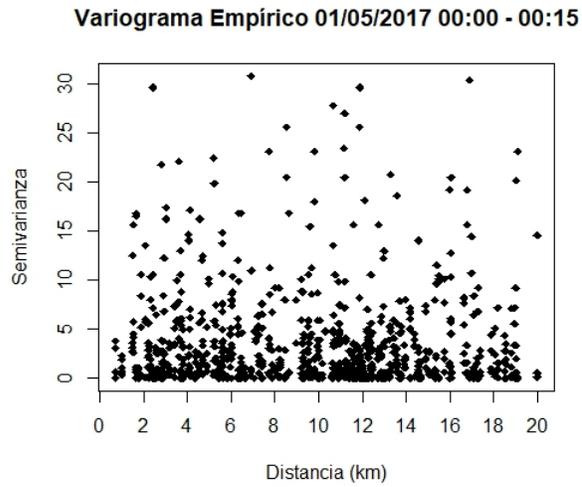


Figura 3.4: Variograma empírico de las temperaturas registradas del 1 al 5 de mayo de 2017, de las 17:00 a las 17:15 horas.

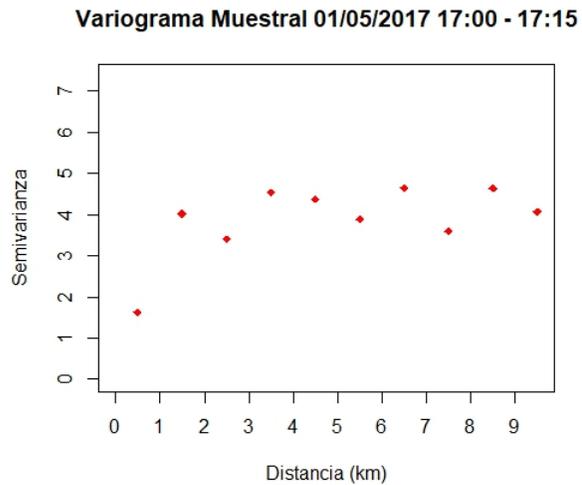


Figura 3.5: Variograma muestral de las temperaturas registradas del 1 al 5 de mayo de 2017 de las 17:00 a las 17:15 horas.

### Modelo de ajuste del variograma

Posteriormente, se aplicaron los tres tipos de modelos de ajuste sobre el variograma muestral para obtener los parámetros mostrados en la Tabla 3.3. Para medir la discrepancia entre el modelo de estimación y los datos se calculó la suma de residuos cuadrados (SRC) y la media de residuos cuadrados (MRC).

Las gráficas de los modelos de ajuste exponencial, esférico y gaussiano pueden apreciarse en las líneas curvas en las las Figuras 3.6a, 3.6b y 3.6c, respectivamente. Cabe mencionar que el comportamiento de los modelos son acotados, por lo que puede decirse que los modelos cumplen con la hipótesis de estacionariedad.

Tabla 3.3: Parámetros de los modelos de ajuste para el variograma del 1 al 5 de mayo de 17:00 a 17:15 horas

Tipo de modelo	Parámetros del modelo			Discrepancia	
	$c_0$	$c$	$a$ (km)	SRC	MRC
Modelo exponencial	0	4.22406	0.845	1.7467	0.17467
Modelo esférico	0	4.14577	1.808	1.6321	0.16321
Modelo gaussiano	0.50352	3.64024	0.822	1.6274	0.16274

Tabla 3.4: Parámetros del variograma para cada modelo de ajuste del 1 al 5 de mayo de 17:00 a 17:15 horas

Tipo de modelo	Varianza nugget	Umbral	Rango efectivo (km)
Modelo exponencial	0	4.22406	2.535
Modelo esférico	0	4.14577	1.808
Modelo gaussiano	0.50352	4.14376	1.423

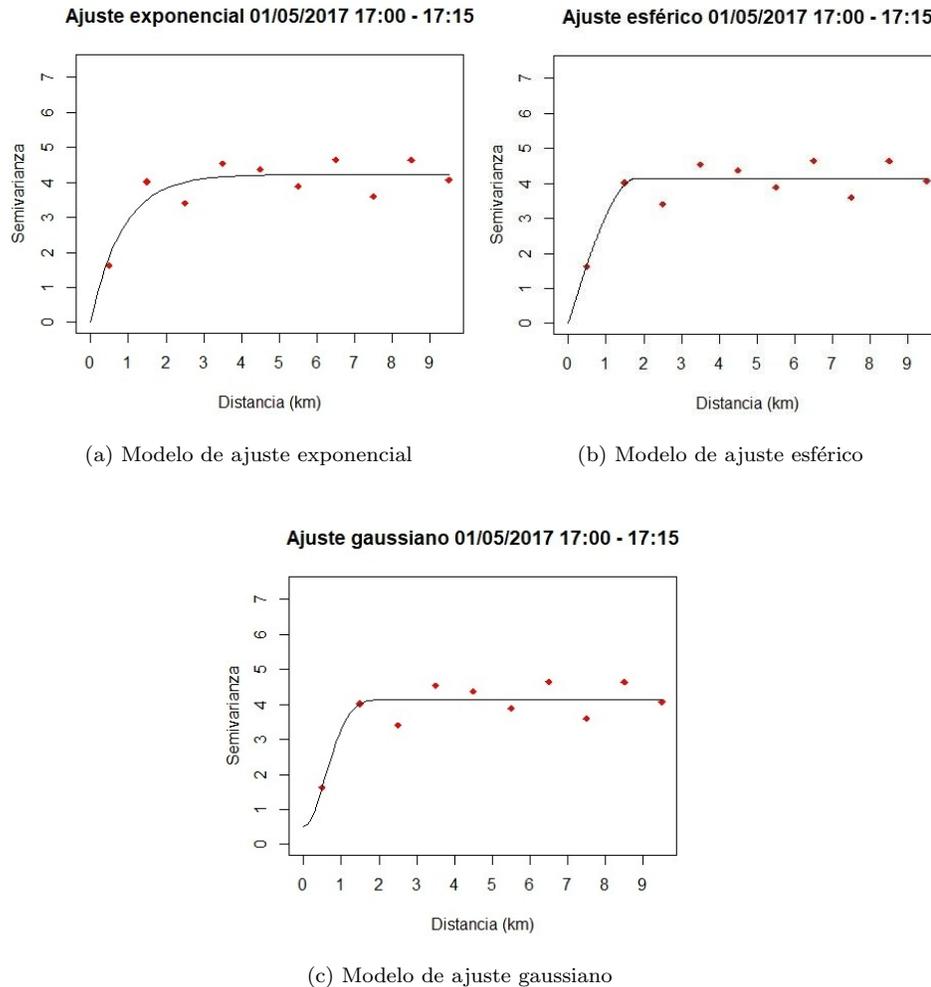


Figura 3.6: Modelo de ajuste exponencial (a), esférico (b) y gaussiano (c) para el variograma del 1 al 5 de mayo de 17:00 a 17:15 horas sobre datos de temperatura.

De las tres estimaciones, según la SRC es el modelo gaussiano el que mejor se ajusta al variograma muestral. Este modelo, aplicado al variograma muestral, tiene un umbral de 4.14376 y un rango efectivo de 1.423 km. Dichos parámetros se muestran en la Tabla 3.4, junto con los correspondientes a los demás modelos.

### Interpolación usando kriging

Una vez obtenidos los modelos de variograma es posible interpolar usando kriging ordinario para obtener las estimaciones de temperatura para todos los

puntos dentro del AMM. En la Figura 3.7 aparecen las mediciones de temperatura por estación, registradas el primero de mayo de 2017. Conviene recordar que no todas las estaciones arrojaron datos durante ese intervalo y que si una estación registró dos o más datos, se tomó en cuenta el promedio de éstos como la medición final de dicha estación.

Es con base en las observaciones que aparecen en la Figura 3.7 que se estiman las temperaturas en los demás puntos de la ciudad. En la Figura 3.8 se muestran los mapas con las estimaciones basadas en el modelo exponencial, en la Figura 3.9 las basadas en el modelo esférico y en la la Figura 3.10 las del modelo gaussiano. Una característica común de las tres estimaciones es que se registran menores temperaturas en las estaciones cercanas a la Sierra Madre Oriental, en particular las que son cercanas al Parque Ecológico Chipinque, y que en las estaciones cercanas a la zona central de Monterrey se registran temperaturas mayores.

Se observa en los mapas con las estimaciones que en las zonas suficientemente alejadas de las estaciones activas las estimaciones de temperatura se vuelven homogéneas, mientras que las estimaciones tienden a variar más conforme la distancia a los puntos de medición disminuye, ya que cuando la distancia sobrepasa el rango efectivo del modelo de variograma, la correlación espacial de la función aleatoria se vuelve inapreciable. Al momento de estimar la temperatura de algún lugar suficientemente alejado del resto de los puntos muestrales, el peso de éstos en el estimador se homogeneizan puesto que la correlación espacial con cada uno de ellos es inapreciable y, por lo tanto, se considera la misma hacia cada uno.

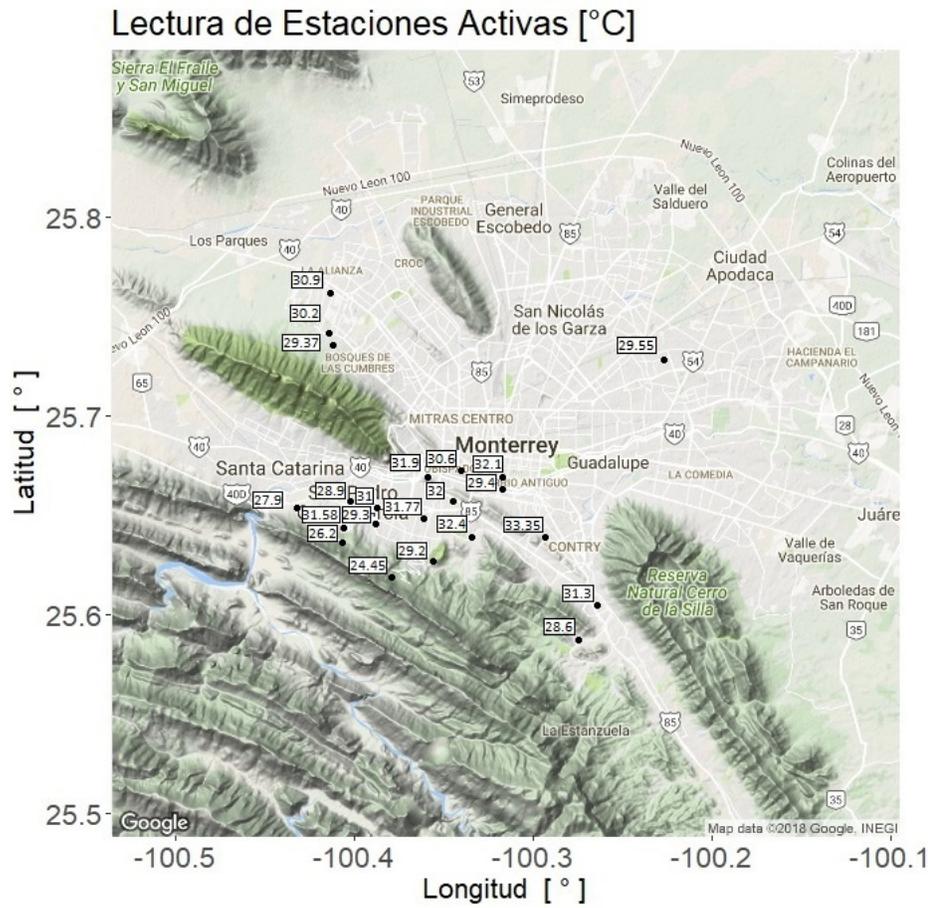


Figura 3.7: Lecturas de temperatura de las estaciones activas del día primero de mayo de 2017, de las 17:00 a las 17:15

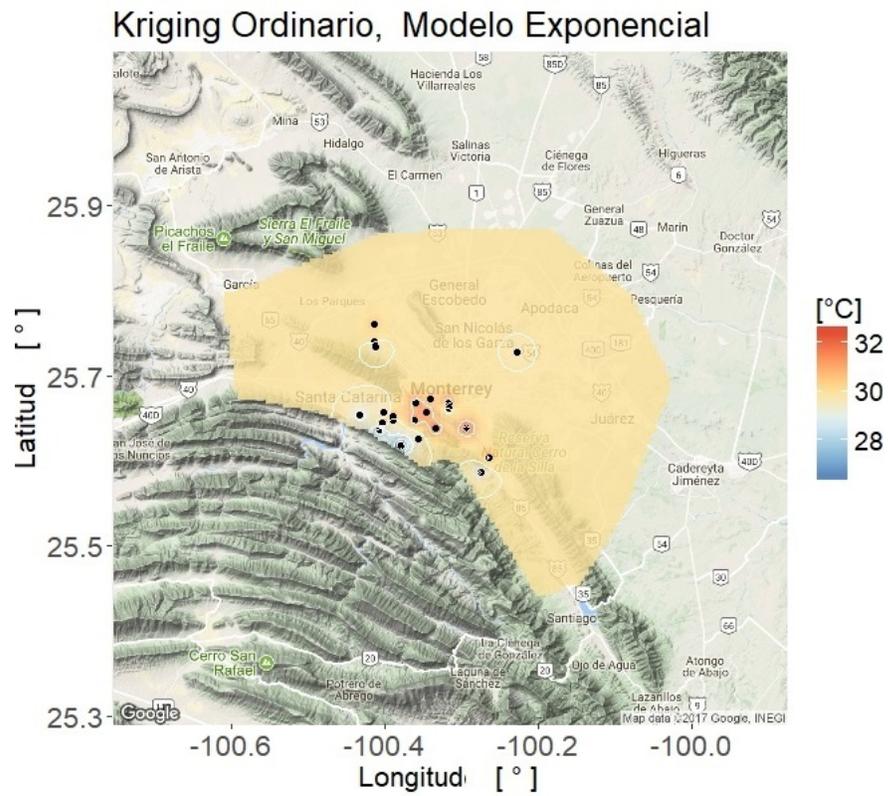


Figura 3.8: Estimación de temperatura del día primero de mayo de 2017, de las 17:00 a las 17:15. Modelo exponencial.

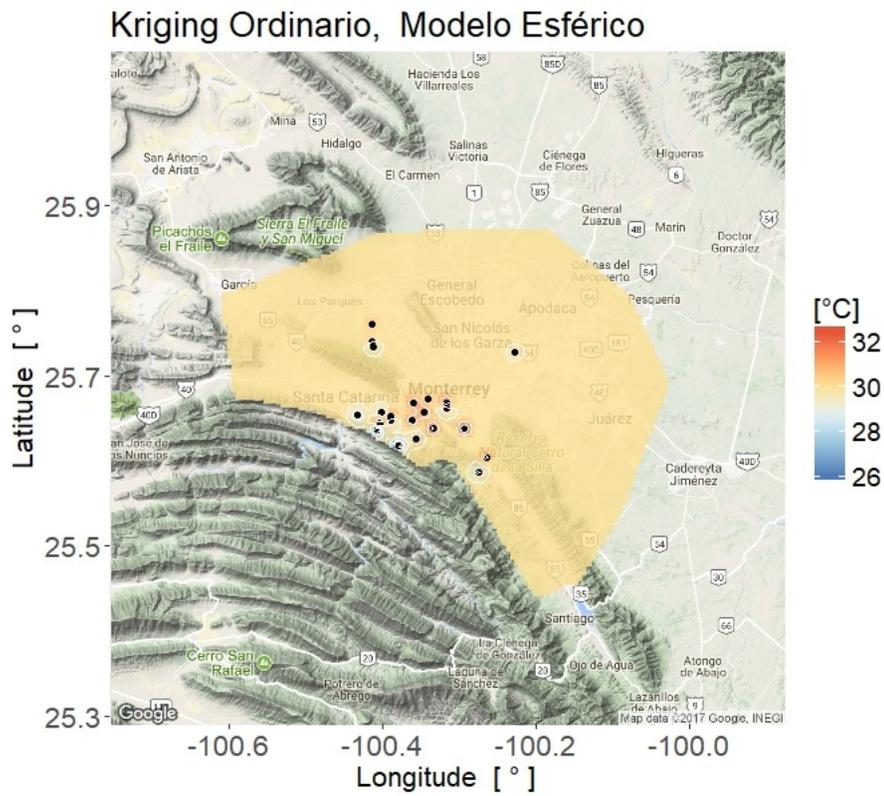


Figura 3.9: Estimación de temperatura del día primero de mayo de 2017, de las 17:00 a las 17:15. Modelo esférico.

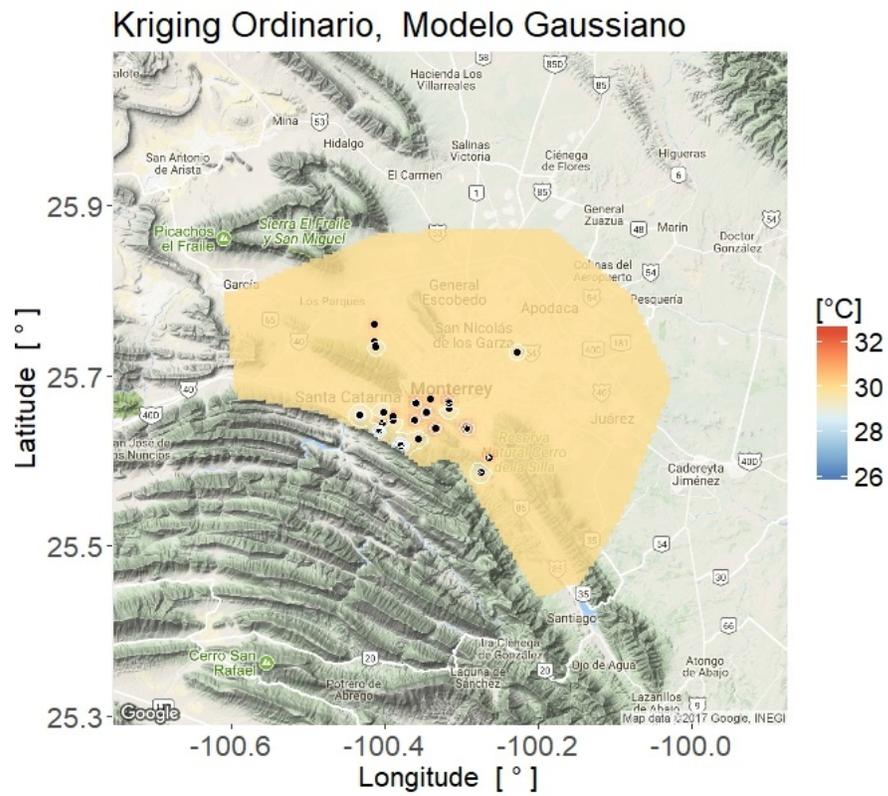


Figura 3.10: Estimación de temperatura del día primero de mayo de 2017, de las 17:00 a las 17:15. Modelo gaussiano.

### 3.2. Análisis del 1 al 5 mayo de 2017, de 00:00 a 00:15

Como segunda parte de este capítulo se muestra la estimación de las temperaturas de la madrugada del primero de mayo usando kriging universal. Para calcular el variograma se usarán las semivarianzas estimadas del día 1 al 5 de mayo en el mismo intervalo de tiempo, y estos cinco bloques de datos estarán sujetos al análisis preliminar.

#### Análisis preliminar

Para tres de las noches se detectaron valores atípicos, mostrados en la Figura 3.11, que fueron omitidos provisionalmente para la construcción del modelo de variograma. Para la detección de valores atípicos en este caso se utilizó el doble del rango intercuartílico como parámetro.

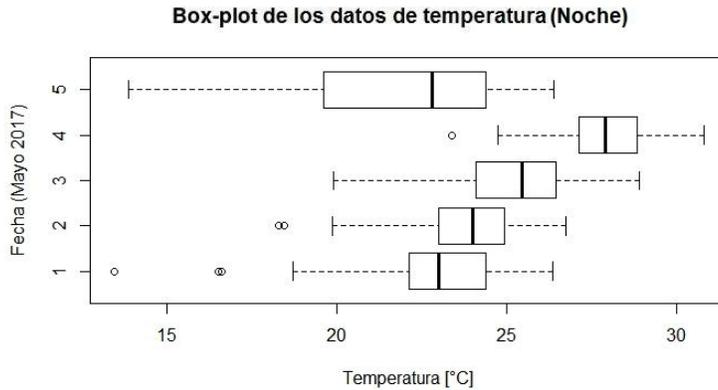


Figura 3.11: Box-plot de las temperaturas registradas del 1 al 5 de mayo de 2017, de las 00:00 a las 00:15 horas.

Habiendo omitido los valores mencionados, se muestra en la Tabla 3.5 el resumen estadístico de los datos modificados. En cada bloque existen alrededor de veinte estaciones con lecturas disponibles. Se puede decir que, en general, la temperatura fue más baja en el periodo que correspondía al 5 de mayo y más alta en el periodo correspondiente al 4 de mayo. También se observa que, en los cinco casos, el coeficiente de asimetría se mantiene nuevamente dentro del rango de  $\pm 1$ .

Tabla 3.5: Resumen estadístico de las temperaturas registradas del 1 al 5 de mayo de 2017, de las 00:00 a las 00:15 horas.

Mayo 2017 (Día)					
Fecha	1	2	3	4	5
Mínimo	18.70	19.87	19.90	24.73	13.85
Primer cuartil	22.50	23.45	24.27	27.66	19.60
Mediana	23.14	24.12	25.48	27.90	22.70
Media	23.64	24.16	25.02	28.04	21.90
Tercer cuartil	25.30	25.31	26.55	28.88	24.40
Máximo	26.35	26.75	28.90	30.80	26.40
Desviación estándar	1.9876	1.7671	2.2576	1.6718	3.7535
Varianza	3.9505	3.1224	5.0967	2.7950	14.0891
Asimetría	-0.4553	-0.7881	-0.6359	-0.3115	-0.8685
Número de datos	18	20	20	18	17

### Estimación del variograma

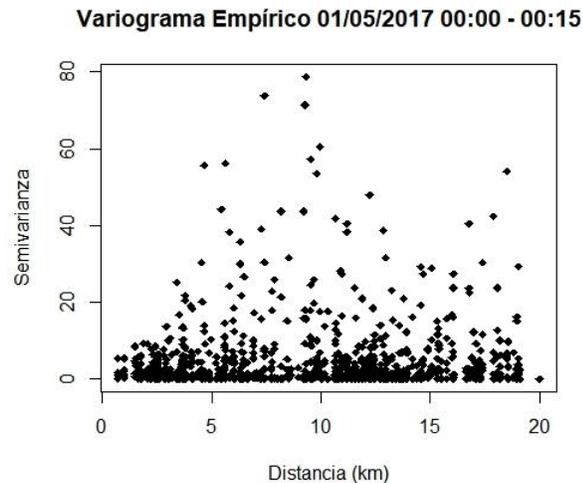


Figura 3.12: Variograma empírico de las temperaturas registradas del 1 al 5 de mayo de 2017, de las 00:00 a las 00:15 horas.

Se muestra en la Figura 3.12 la gráfica del variograma empírico, la primera aproximación al modelo final de variograma. A partir de las semivarianzas que allí se muestran se agruparon en diez intervalos que abarcan hasta una distancia  $0.45 \cdot d_{\text{máx}} = 0.45 \cdot 20.016 \text{ km} = 9.007 \text{ km}$ . Se eligió el factor 0.45 para definir el alcance del variograma muestral sobre el factor 0.5 puesto que al usar el segundo, en el último intervalo, centrado en la distancia  $h = 9.508$ , se obtiene

12.17085 como estimación de la semivarianza, que puede considerarse un valor inusualmente alto si se compara con las semivarianzas vecinas (véase Figura 3.14) y, por lo tanto, afectaría el modelo de ajuste del variograma.

Tabla 3.6: Intervalos de distancia, semivarianzas y conteo de estimaciones de las temperaturas registradas del 1 al 5 de mayo de 2017, de las 00:00 a las 00:15 horas.

Distancia (km)	Semivarianza	Número de estimaciones
0.450	1.68881	5
1.351	2.65779	25
2.252	2.62266	50
3.153	3.82046	38
4.053	5.14339	52
4.954	5.51016	35
5.855	7.02178	52
6.755	5.82723	40
7.656	7.91015	37
8.557	6.31689	21

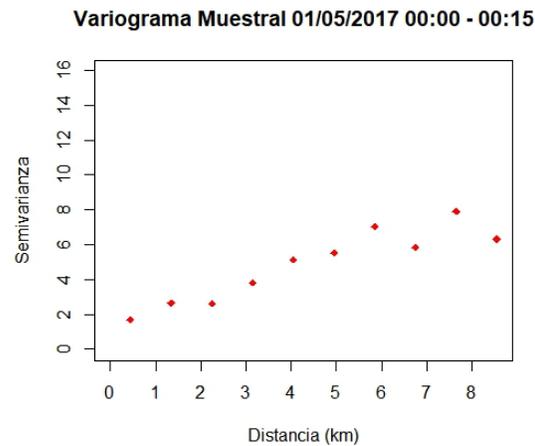


Figura 3.13: Variograma muestral de las temperaturas registradas del 1 al 5 de mayo de 2017 de las 00:00 a las 00:15 horas.

Las estimaciones de las semivarianzas según el intervalo se muestran en la Tabla 3.6. Puede notarse que los valores de semivarianza siguen en aumento

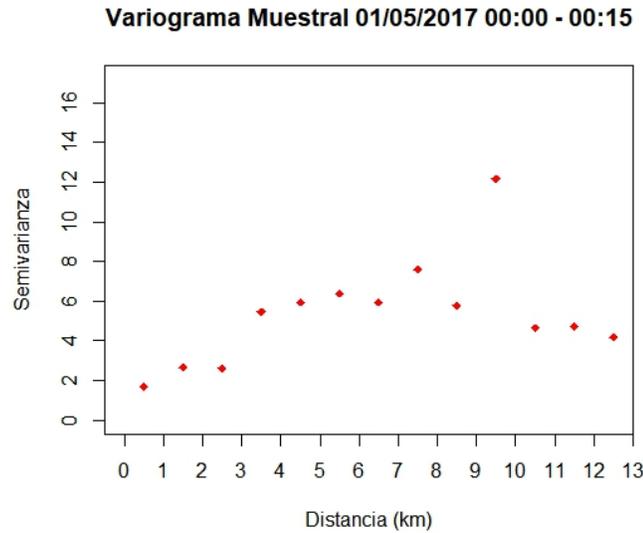


Figura 3.14: Variograma empírico (extendido) de las temperaturas registradas del 1 al 5 de mayo de 2017, de las 00:00 a las 00:15 horas.

hasta una distancia cercana a 6 km, por lo que es posible el rango del variograma adquiera un valor cercano a éste. La gráfica donde se ilustran los datos anteriores aparece en la Figura 3.13. Observando dicha ilustración, no parece tan trivial observar el umbral del variograma, ya que, en las últimas cinco distancias graficadas, el valor de las semivarianzas oscila.

### Modelo de ajuste del variograma

Al variograma muestral se ajustaron los tres modelos sujetos a análisis. Los parámetros obtenidos al hacer el ajuste son los que aparecen en la Tabla 3.7. Las primeras tres columnas incluyen los parámetros de cada modelo, y las últimas dos columnas muestran la SRC y la MRC como indicadores de ajuste entre la estimación y los datos.

Tabla 3.7: Parámetros de los modelos de ajuste para el variograma del 1 al 5 de mayo de 00:00 a 00:15 horas

Tipo de modelo	Parámetros del modelo			Discrepancia	
	$c_0$	$c$	$a$ (km)	SRC	MRC
Modelo exponencial	0.82784	8.72671	6.460	4.2212	0.42212
Modelo esférico	0.92062	6.01140	8.327	3.5408	0.35408
Modelo gaussiano	1.68799	5.37144	4.186	3.2272	0.32272

Tabla 3.8: Parámetros del variograma para cada modelo de ajuste del 1 al 5 de mayo de 00:00 a 00:15 horas

Tipo de modelo	Varianza nugget	Umbral	Rango efectivo (km)
Modelo exponencial	0.82784	9.55455	19.378
Modelo esférico	0.92062	6.93202	8.327
Modelo gaussiano	1.68799	7.05943	7.242

A continuación se muestran las gráficas de cada uno de los modelos aplicados sobre el variograma muestral. La Figura 3.15a representa el modelo exponencial, la Figura 3.15b el modelo esférico y la Figura 3.15c el modelo gaussiano.

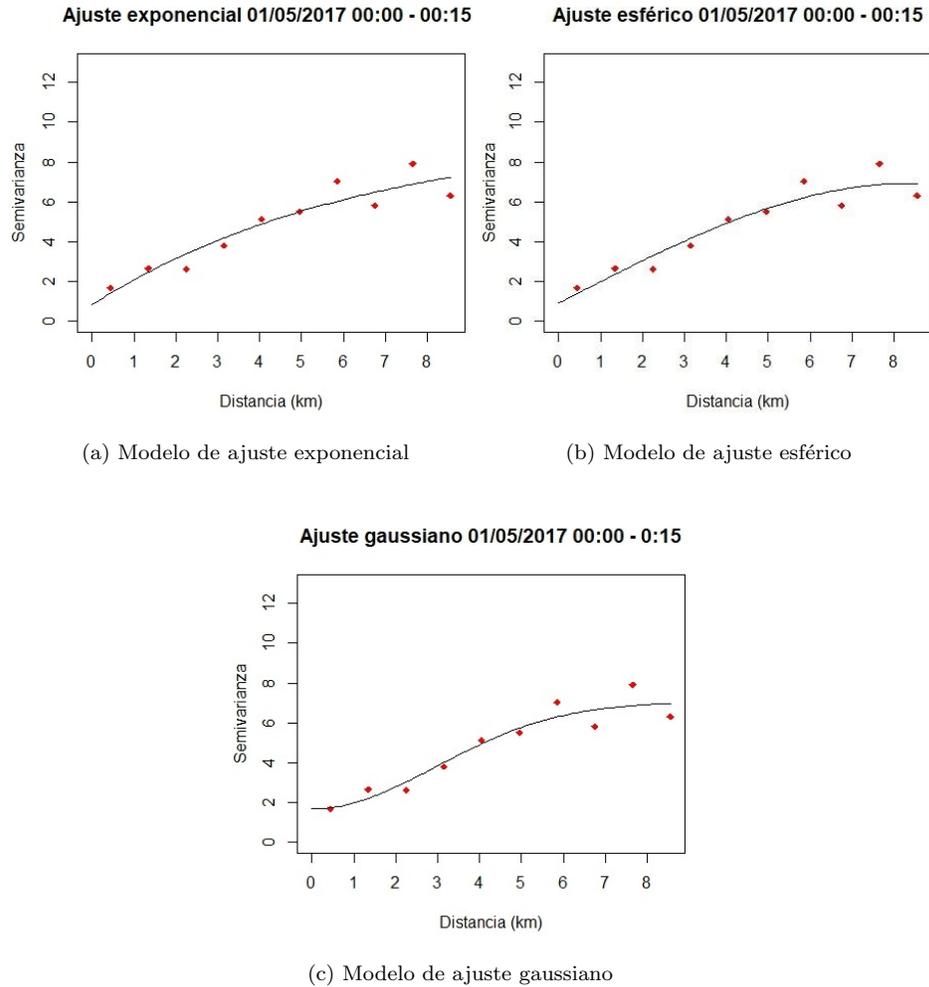


Figura 3.15: Modelo de ajuste exponencial (a), esférico (b) y gaussiano (c) para el variograma del 1 al 5 de mayo de 00:00 a 00:15 horas sobre los datos de temperatura.

El modelo gaussiano es el modelo que mejor ajusta sobre los datos con una media de residuos cuadrados de 0.32272. También es este modelo el que tiene un rango efectivo menor. Es importante señalar que el modelo con mayor rango efectivo es el exponencial, con 19.378 km, que es menor que la distancia máxima por poco menos de un kilómetro. Esto implicará que el variograma esté alejado de su umbral a lo largo de un rango de distancia muy extenso en comparación de las distancias entre las estaciones. Por lo tanto, podría esperarse que el área del mapa donde las estimaciones varían sea más extensa cuando se utiliza como base el modelo exponencial que cuando se implementa el kriging con los demás

modelos.

### Interpolación usando kriging

El promedio de las mediciones registradas por estación activa de las 00:00 a las 00:15 se muestran en la Figura 3.16. A partir de ellos se realiza el kriging ordinario para interpolar la temperatura en cada punto del AMM el día primero de mayo en el bloque de tiempo mencionado.

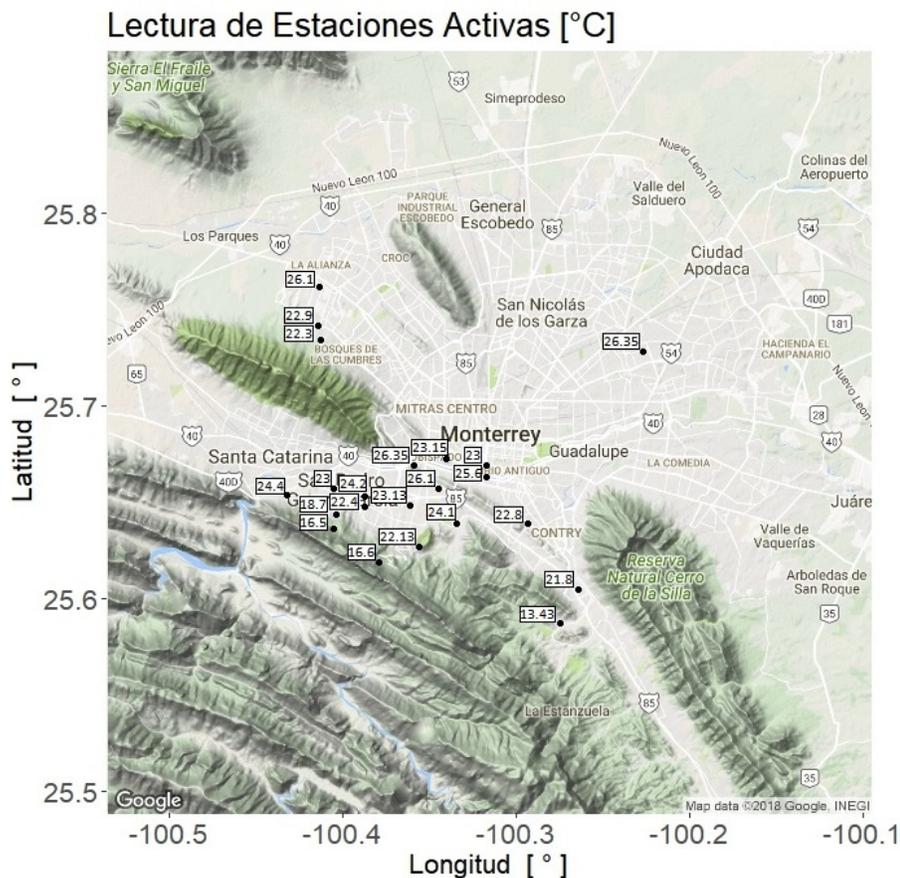


Figura 3.16: Lecturas de las estaciones activas del día primero de mayo de 2017, de las 00:00 a las 00:15

La interpolación basada en el modelo exponencial se muestra en la Figura 3.17, en la Figura 3.18 está el mapa de estimaciones producido por el modelo esférico y en la Figura 3.19 aparece la estimación a partir del modelo gaussiano. Notamos que las temperaturas más bajas se concentran en la zona sur del AMM, cerca de las estaciones ubicadas en el Parque Ecológico Chipinque, mientras que las temperaturas más altas se concentran en la zona centro y

norte del AMM.

En general, y sobre todo con el modelo exponencial, el área en que la estimación varía es mayor que en el caso de las temperaturas en la sección anterior. Esto es entendible dado que el rango del variograma en este caso tiene una magnitud cercana a la distancia máxima entre estaciones.

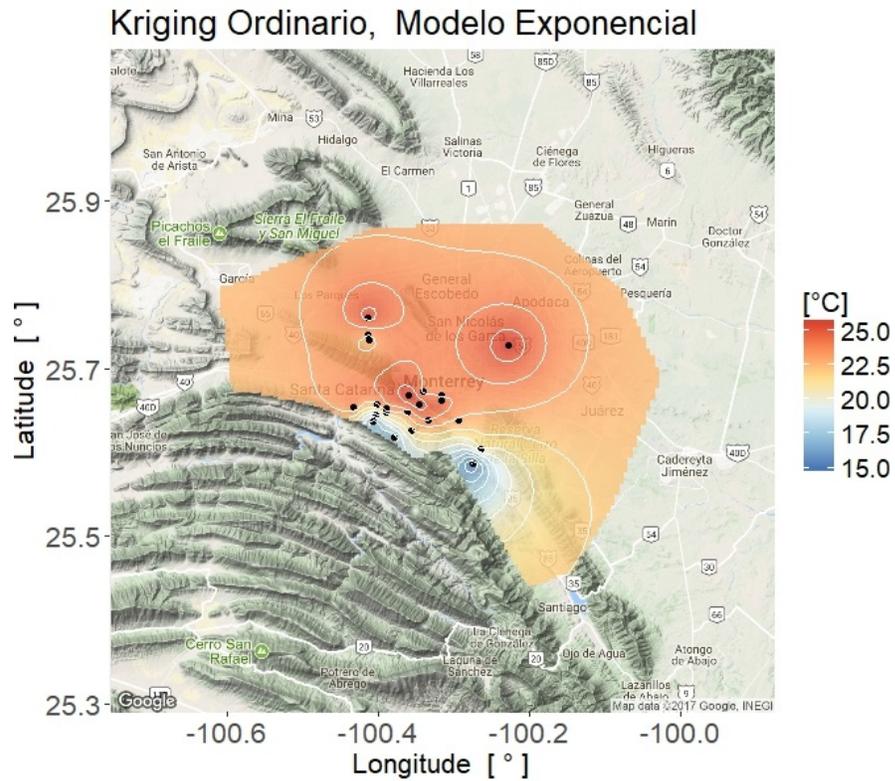


Figura 3.17: Estimación de las temperaturas del día primero de mayo de 2017, de las 00:00 a las 00:15. Modelo exponencial.

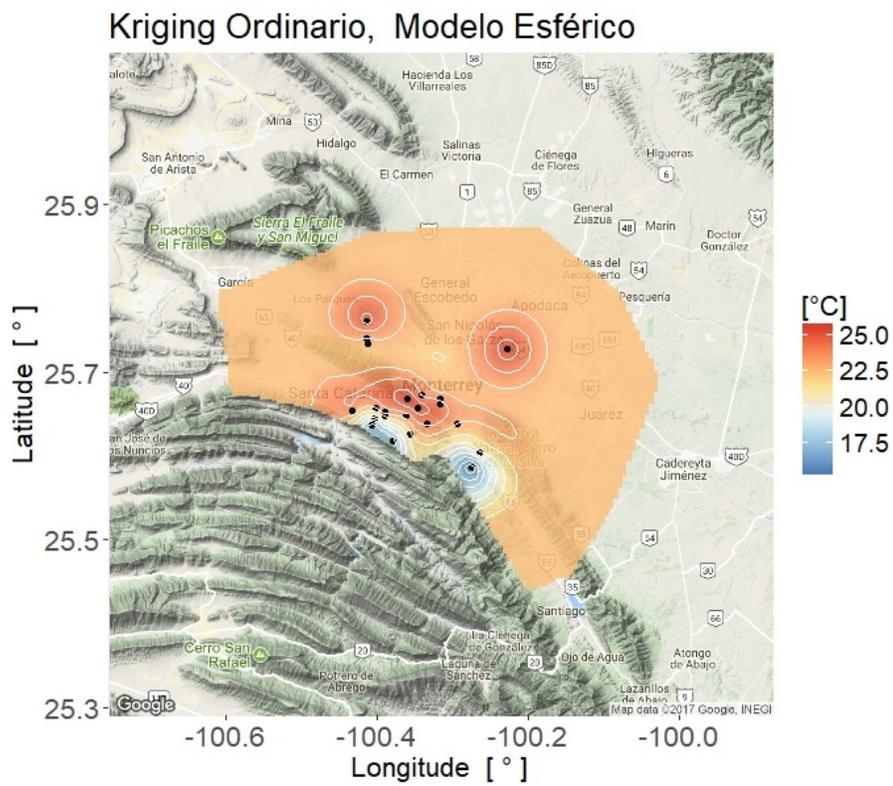


Figura 3.18: Estimación de las temperaturas del día primero de mayo de 2017, de las 00:00 a las 00:15. Modelo esférico.

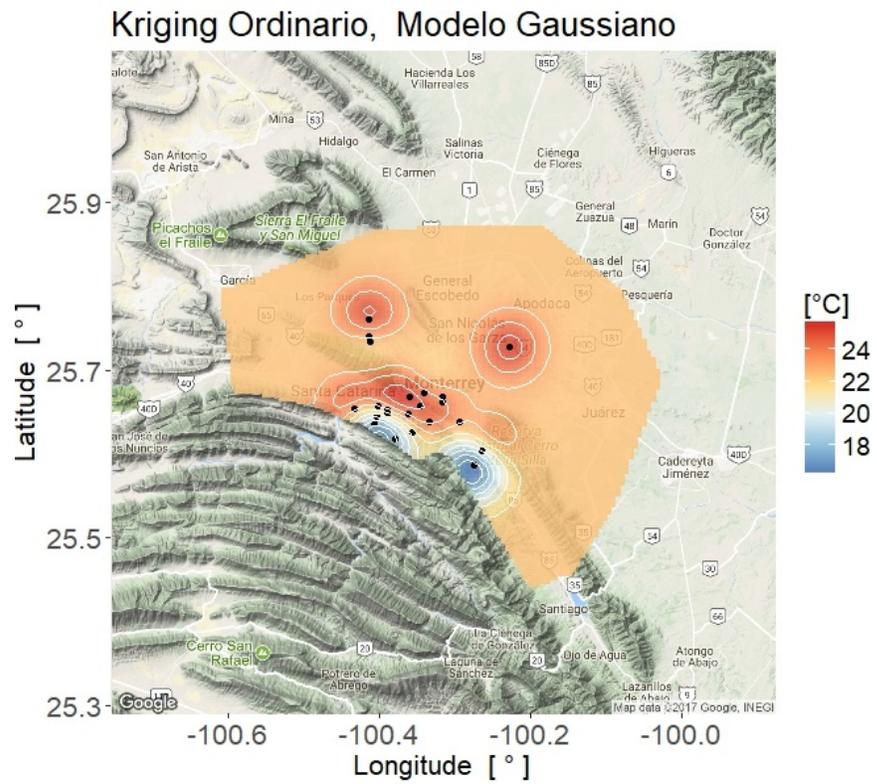


Figura 3.19: Estimación de las temperaturas del día primero de mayo de 2017, de las 00:00 a las 00:15. Modelo gaussiano.

## Capítulo 4

# Aplicación de kriging en datos de humedad relativa

Al igual que en el capítulo anterior, para modelar la correlación espacial se utilizaron las semivarianzas del 1 al 5 de mayo de 2017, en los mismos periodos de quince minutos. Además, no se tomaron en cuenta las mediciones de la estación ubicada en las coordenadas  $(-100.414^\circ, 25.7415^\circ)$  en ninguna de las dos secciones, ya que dificultaba la estimación el variograma, así como también se omitieron las mediciones de la estación ubicada en las coordenadas  $(-100.34^\circ, 25.6726^\circ)$  para el análisis de 00:00 a 00:15. Se muestran en las Figuras 4.1 y 4.2 cómo lucirían los variogramas si se no se omitieran las estaciones mencionadas en las secciones 4.1 y en 4.2, respectivamente.

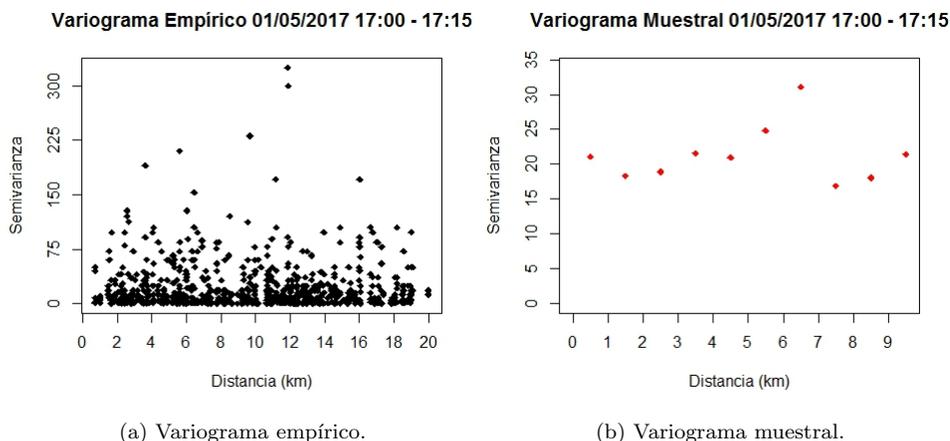


Figura 4.1: Variograma empírico (a) y muestral (b) del 1 al 5 de mayo de 2017, de las 17:00 a las 17:15 horas, incluyendo las estaciones en las coordenadas  $(-100.414^\circ, 25.7415^\circ)$  y  $(-100.34^\circ, 25.6726^\circ)$ .

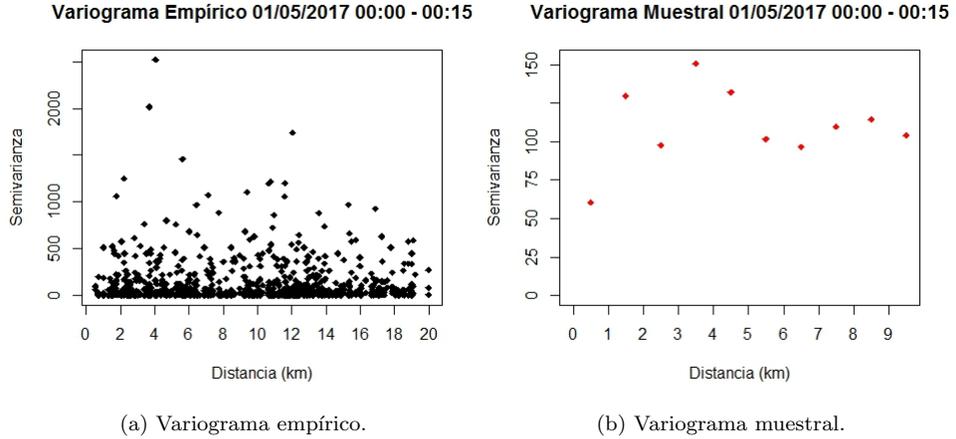


Figura 4.2: Variograma empírico (a) y muestral (b) del 1 al 5 de mayo de 2017, de las 00:00 a las 00:15 horas, incluyendo las estaciones en las coordenadas  $(-100.414^\circ, 25.7415^\circ)$  y  $(-100.34^\circ, 25.6726^\circ)$ .

#### 4.1. Análisis del 1 al 5 mayo de 2017, de 17:00 a 17:15

En esta sección se estimó el porcentaje de humedad en cada punto del AMM en el intervalo de las 17:00 a las 17:15 del día primero de mayo de 2017.

##### Análisis preliminar

Se sometieron al análisis preliminar los datos de los cinco primeros días de mayo, ya que serían estos cinco bloques de datos que se usarían para elaborar el variograma. La primera tarea de este análisis fue descartar los valores atípicos de los datos y como parámetro para detectar este tipo de valores se usó el doble del rango intercuartílico. La Figura 4.3 muestra que en este caso sí hubo varias mediciones atípicas en cada uno de los días. Estas mediciones no serán tomadas en cuenta para elaborar el modelo de variograma.

Una vez omitidas las mediciones mencionadas, se obtiene el resumen estadístico de cada uno de los días mostrados en la Tabla 4.1. El día con mayor humedad fue el 3 de mayo, mientras que el día con menor índice de humedad fue el primero de mayo. Además, es importante notar que los índices de asimetría se mantienen dentro del rango de  $\pm 1$  en todos los casos, por lo que no se hará ninguna otra transformación a los datos antes de estimar los variogramas.

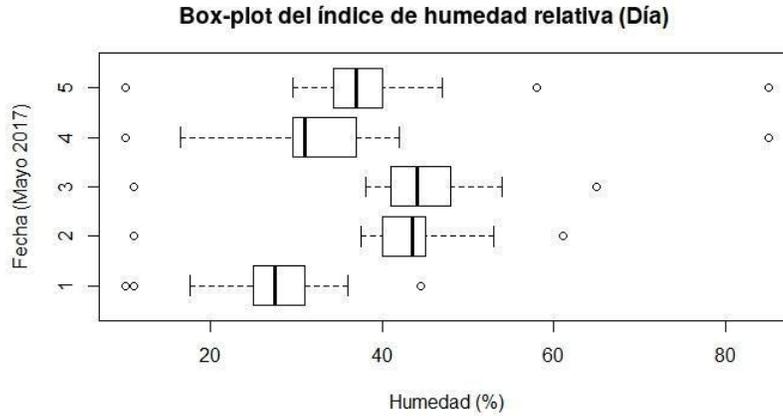


Figura 4.3: Box-plot de los índices de humedad registrados del 1 al 5 de mayo de 2017, de las 17:00 a las 17:15

Tabla 4.1: Resumen estadístico de los índices de humedad registrados del 1 al 5 de mayo de 2017, de las 17:00 a las 17:15 horas.

Mayo 2017 (Día)					
Fecha	1	2	3	4	5
Mínimo	17.50	37.50	38.00	28.00	29.50
Primer cuartil	26.12	40.5	42.00	29.58	34.33
Mediana	28.00	43.00	44.00	31.00	37.00
Media	28.31	43.0	45.03	32.31	36.78
Tercer cuartil	31.00	45.00	48.00	36.50	39.00
Máximo	36.00	53.00	54.00	42.00	47.00
Desviación estándar	4.5772	3.7006	4.4458	6.1412	4.2962
Varianza	20.9506	13.6944	19.7647	37.7138	18.4575
Asimetría	-0.3337	0.7400	0.4712	-0.6889	0.5773
Número de datos	18	19	17	15	17

### Estimación del variograma

La primera etapa de esta parte del proceso es la elaboración del variograma empírico, el cual se muestra en la Figura 4.4. A partir de los datos que arrojó este último, se calculó el variograma muestral, que en este caso toma en cuenta diez intervalos, y que abarca hasta la mitad de la distancia máxima  $d_{\text{máx}} = 20.016$  km, que es  $\frac{d_{\text{máx}}}{2} = 10.008$  km. En la Tabla 4.2 se muestran los intervalos de distancia y semivarianzas que componen el variograma muestral, mientras que su gráfica correspondiente aparece en la Figura 4.5.

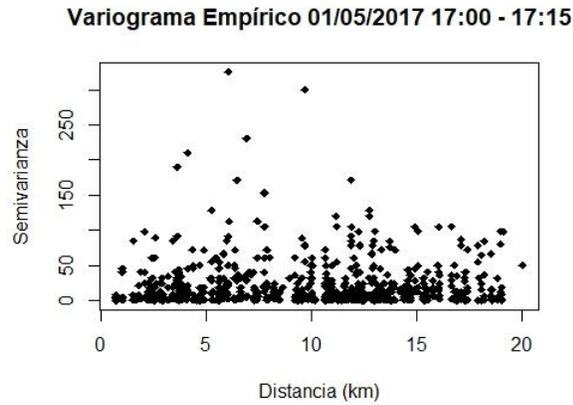


Figura 4.4: Variograma empírico de los índices de humedad registrados del 1 al 5 de mayo de 2017, de las 17:00 a las 17:15 horas.

Es probable que el rango de este variograma sea alrededor de 7 km puesto que es a partir del séptimo intervalo (que corresponde a la distancia  $h = 6.505$ ) que las semivarianzas dejan de ir en aumento. Sin embargo, cabe señalar que a partir de este intervalo las semivarianzas no se estabilizan y, en cambio, oscilan considerablemente.

Tabla 4.2: Intervalos de distancia, semivarianza y conteo de estimaciones de los índices de humedad registrados del 1 al 5 de mayo de 2017, de las 17:00 a las 17:15 horas.

Distancia (km)	Semivarianza	Número de estimaciones
0.500	4.72500	5
1.501	14.19643	21
2.502	14.21965	54
3.503	24.51659	36
4.504	20.19520	37
5.504	21.89410	40
6.505	39.57385	41
7.506	29.86071	35
8.507	13.86111	18
9.508	21.32321	48

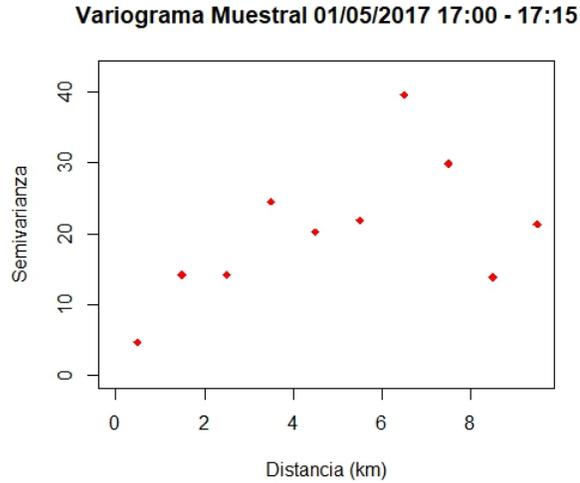


Figura 4.5: Variograma muestral de los datos de humedad relativa registrados del 1 al 5 de mayo de 2017 de las 17:00 a las 17:15 horas.

### Modelo de ajuste del variograma

Los tres modelos de ajuste fueron aplicados sobre el variograma muestral, con los cuales se obtuvieron los parámetros mostrados en la Tabla 4.3. También se incluyen como indicador de ajuste del modelo la SRC y la MRC de los datos y la estimación del modelo, siendo el modelo esférico el que se ajusta mejor a los datos. Este modelo tiene umbral de 25.62683 y un rango de 6.520 km. Estos parámetros y los que corresponden a los demás modelos aparecen en la Tabla 4.4.

Tabla 4.3: Parámetros de los modelos de ajuste para el variograma del 1 al 5 de mayo de 17:00 a 17:15 horas

Tipo de modelo	Parámetros del modelo			Discrepancia	
	$c_0$	$c$	$a$ (km)	SRC	MRC
Modelo exponencial	0	25.41365	1.951	433.9770	43.39770
Modelo esférico	4.29628	21.33055	6.520	421.8014	42.18014
Modelo gaussiano	5.57094	19.39287	4.561	426.4914	42.64914

Tabla 4.4: Parámetros del variograma para cada modelo de ajuste del 1 al 5 de mayo de 17:00 a 17:15 horas

Tipo de modelo	Varianza nugget	Umbral	Rango efectivo (km)
Modelo exponencial	0	25.41365	5.854
Modelo esférico	4.29628	25.62683	6.520
Modelo gaussiano	5.57094	24.96381	2.344

Las gráficas de cada uno de los modelos de ajuste exponencial, esférico y gaussiano, aplicados al variograma muestral, se muestran en las Figuras 4.6a, 4.6b y 4.6c, respectivamente. En los tres casos los modelos tienen umbral y son acotados, por lo que se interpreta que cumple con la hipótesis de estacionariedad.

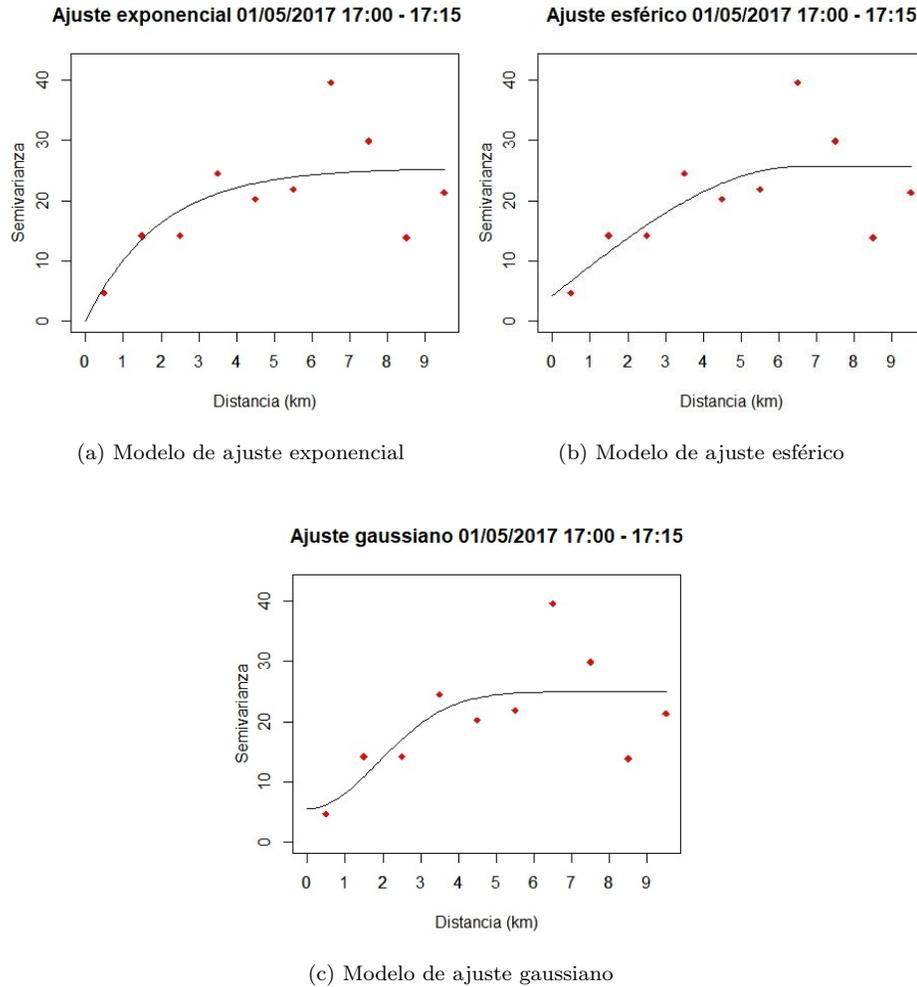


Figura 4.6: Modelo de ajuste exponencial (a), esférico (b) y gaussiano (c) para el variograma del 1 al 5 de mayo de 17:00 a 17:15 horas.

### Interpolación usando kriging

Por último, se realiza la interpolación usando el estimador kriging ordinario para obtener las aproximaciones de la humedad relativa en cada punto del AMM, específicamente en el primer día de mayo de 2017, de las 17:00 a las 17:15.

Las observaciones de este modelo geoestadístico se muestran la Figura 4.7, y es a partir de éstas que se realiza la interpolación. En la Figura 4.8 se muestra

la estimación basada en el modelo exponencial, en la Figura 4.9 la estimación del modelo esférico y, por último, en la Figura 4.10, la construida a partir del gaussiano.

De las estimaciones notamos que la zona norte y central del municipio de Monterrey son regiones donde parece haber menor humedad. En cambio, la zona sur de San Pedro Garza García, cercana al Parque Ecológico Chipinque, registra un mayor índice de humedad.

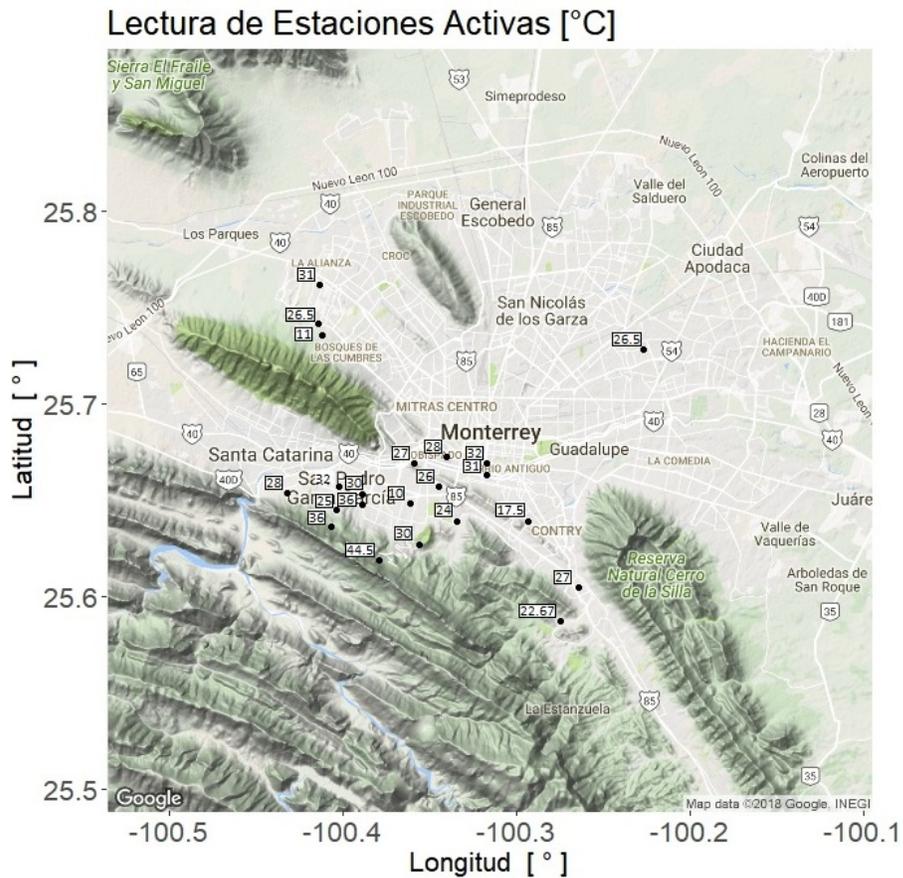


Figura 4.7: Lecturas de las estaciones activas del día primero de mayo de 2017, de las 17:00 a las 17:15

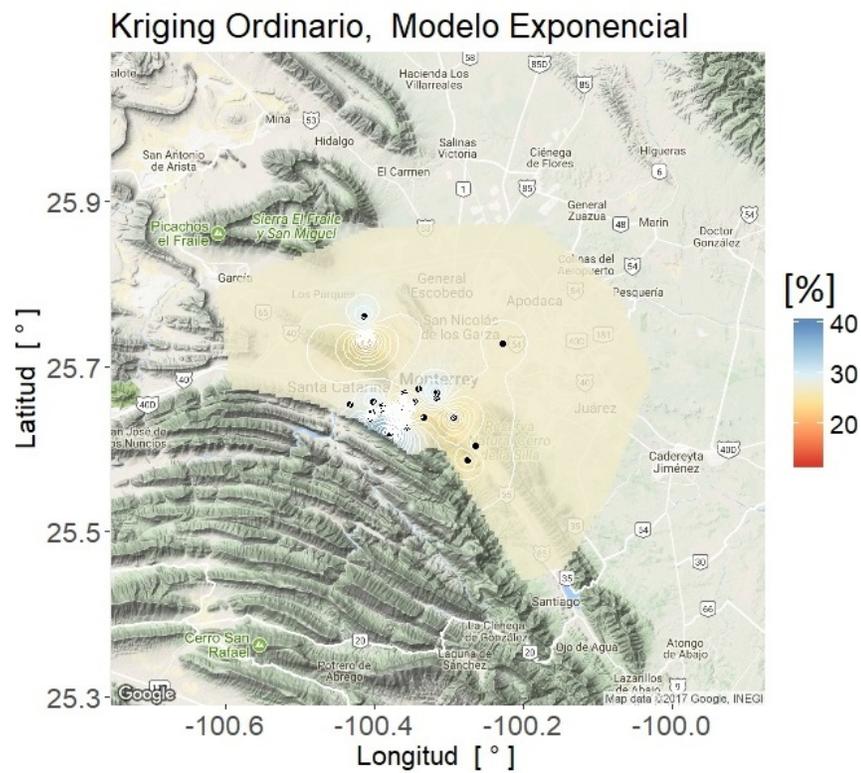


Figura 4.8: Estimación de los índices de humedad relativa del día primero de mayo de 2017, de las 17:00 a las 17:15. Modelo exponencial.

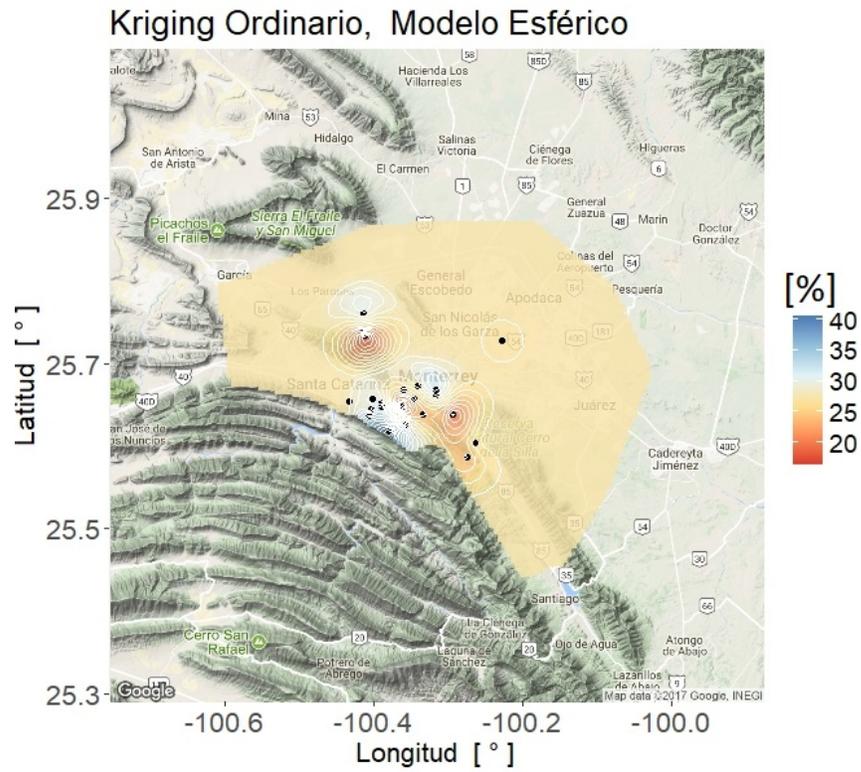


Figura 4.9: Estimación de los índices de humedad relativa del día primero de mayo de 2017, de las 17:00 a las 17:15. Modelo esférico.

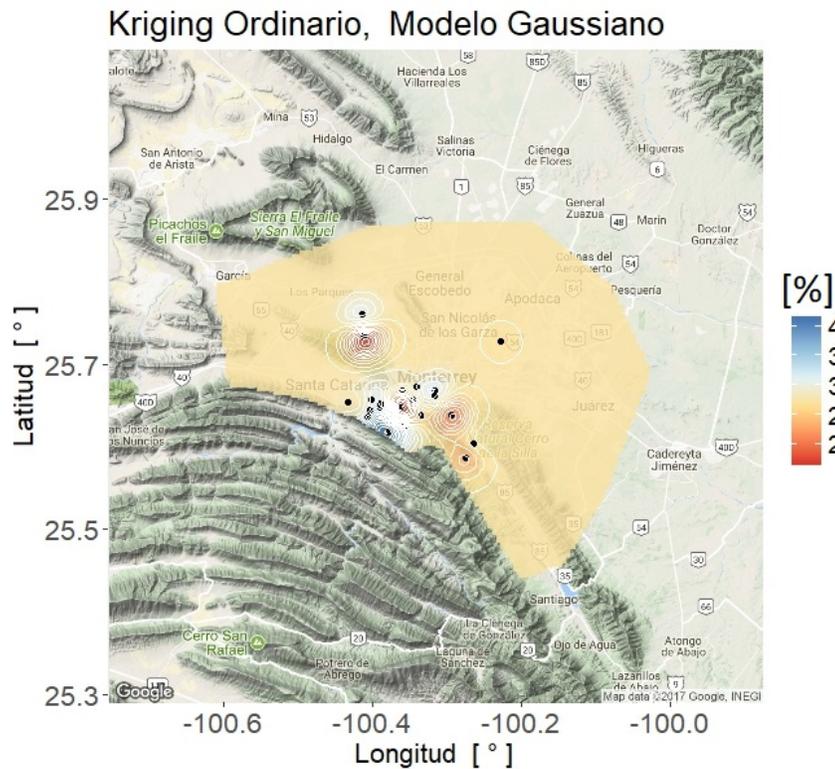


Figura 4.10: Estimación de los índices de humedad relativa del día primero de mayo de 2017, de las 17:00 a las 17:15. Modelo gaussiano.

Puede observarse que el rango de los modelos de variograma fue mayor al estimar la humedad que cuando se estimó la temperatura, y esto se traduce en que el área donde las estimaciones varían por la influencia de las mediciones sobre el estimador es más extensa que en el caso anterior. Fuera de esta área de influencia, al igual que en el capítulo anterior, las estimaciones se vuelven homogéneas.

## 4.2. Análisis del 1 al 5 mayo de 2017, de 00:00 a 00:15

En esta sección se analizan y procesan los datos obtenidos de las mediciones de 1 al 5 de mayo de las 00:00 a las 00:15 para estimar el índice de humedad en cada punto del AMM usando kriging ordinario.

### Análisis preliminar

El análisis preliminar de los datos arroja que en efecto hubo algunas mediciones atípicas en los días 2, 3 y 4 de mayo, como se puede observar en los puntos fuera de los box-plot en la Figura 4.11. En este caso, se utilizó como parámetro para detectar este tipo de valores el doble del rango intercuartílico.

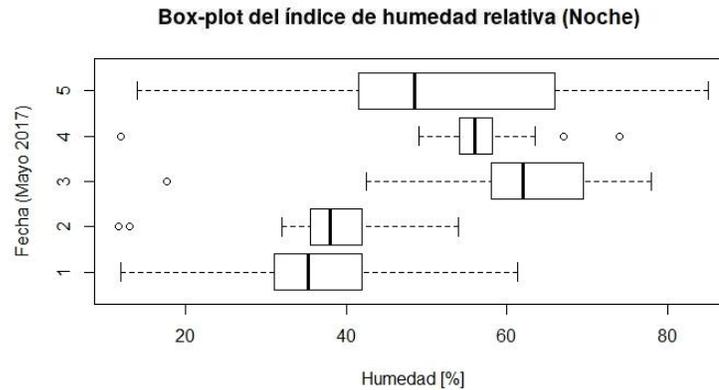


Figura 4.11: Box-plot de los índices de humedad registrados del 1 al 5 de mayo de 2017, de las 00:00 a las 00:15 horas.

Los bloques de datos se mantuvieron dentro del rango de  $\pm 1$  en cuanto al coeficiente de asimetría. Según el resumen estadístico en la Tabla 4.5, la variación en el índice de humedad del día 1 y 5 de mayo fue notoriamente mayor que la del resto de los días.

Tabla 4.5: Resumen estadístico de los datos de humedad relativa registrados del 1 al 5 de mayo de 2017, de las 00:00 a las 00:15 horas.

Mayo 2017 (Día)					
Fecha	1	2	3	4	5
Mínimo	12.00	32.00	42.50	49.00	14.00
Primer cuartil	31.00	36.00	58.50	54.29	42.00
Mediana	34.50	38.00	62.00	56.00	47.00
Media	36.92	38.33	63.18	55.96	51.26
Tercer cuartil	42.00	39.88	69.50	57.25	64.00
Máximo	61.33	46.67	78.00	63.50	85.00
Desviación estándar	12.7916	3.9087	8.3453	3.4338	17.5055
Varianza	163.6239	15.2778	69.6447	11.7907	306.4412
Asimetría	0.30889	0.53353	-0.36258	0.24141	0.15607
Número de datos	21	18	19	16	17

### Estimación del variograma

El variograma empírico se muestra en la Figura 4.12. Para elaborar el variograma muestral, las semivarianzas se agruparon en diez intervalos que abarcan hasta la mitad de la distancia máxima  $\frac{d_{\text{máx}}}{2} = \frac{20.016}{2} = 10.008$  km, y las estimaciones de semivarianza por intervalo se muestran en la Tabla 4.6 y su gráfica se ilustra en la Figura 4.13 .

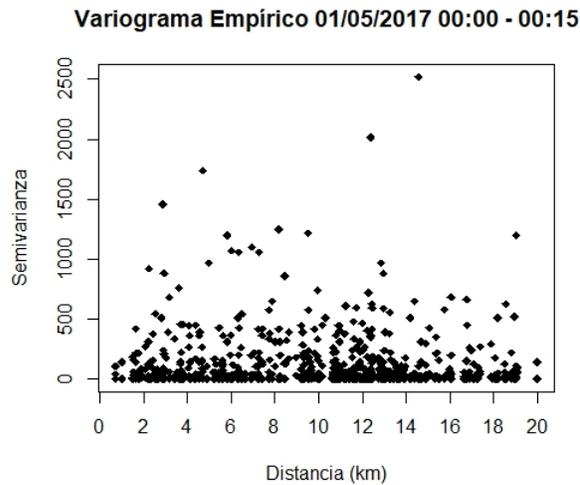


Figura 4.12: Variograma empírico de los datos de humedad relativa registrados del 1 al 5 de mayo de 2017, de las 00:00 a las 00:15 horas.

Tabla 4.6: Intervalos de distancia, semivarianzas y conteo de estimaciones de los datos de humedad relativa registrados del 1 al 5 de mayo de 2017, de las 00:00 a las 00:15 horas.

Distancia (km)	Semivarianza	Número de estimaciones
0.500	53.32500	5
1.501	72.00617	27
2.502	127.59579	58
3.503	110.04380	39
4.504	153.33559	43
5.504	101.78395	45
6.505	145.04663	42
7.506	156.23198	37
8.507	221.38068	22
9.508	143.60289	49

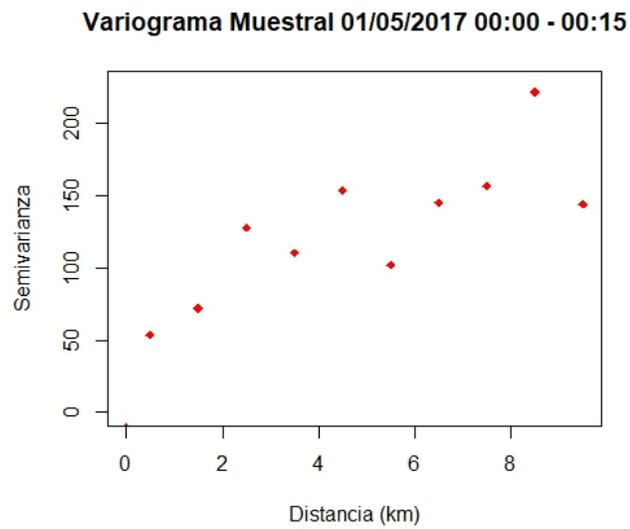


Figura 4.13: Variograma muestral de los datos de humedad relativa registrados del 1 al 5 de mayo de 2017 de las 00:00 a las 00:15 horas.

### Modelo de ajuste del variograma

Se aplican los tres modelos de ajuste para estimar el variograma.

Tabla 4.7: Parámetros de los modelos de ajuste para el variograma del 1 al 5 de mayo de 00:00 a 00:15 horas para los datos de humedad relativa

Tipo de modelo	Parámetros del modelo			Discrepancia	
	$c_0$	$c$	$a$ (km)	SRC	MRC
Modelo exponencial	41.29007	153.36829	5.007	6890.792	689.0792
Modelo esférico	52.02263	118.11461	9.4399	6970.295	697.0295
Modelo gaussiano	68.89017	102.58372	4.751	7629.947	762.9947

Tabla 4.8: Parámetros del variograma para cada modelo de ajuste del 1 al 5 de mayo de 00:00 a 00:15 horas para los datos de humedad relativa

Tipo de modelo	Varianza nugget	Umbral	Rango efectivo (km)
Modelo exponencial	41.29007	194.6584	15.021
Modelo esférico	52.02263	170.1372	9.440
Modelo gaussiano	68.89017	171.4739	8.219

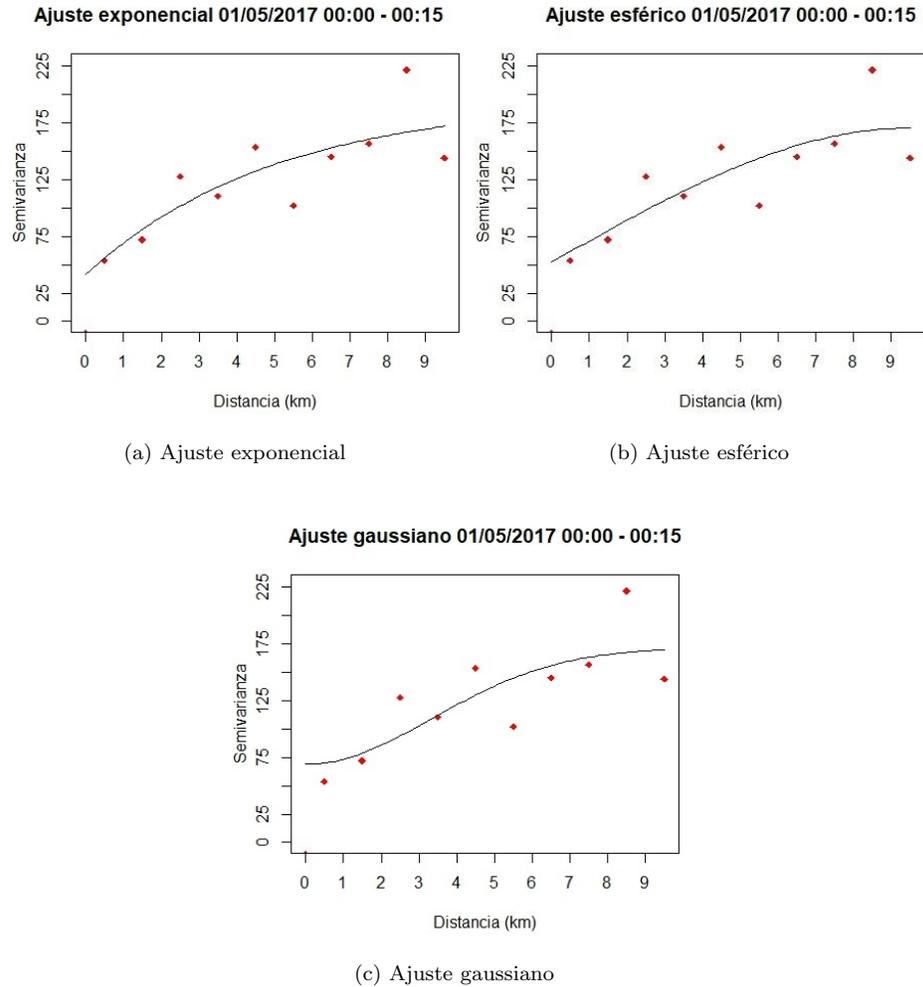


Figura 4.14: Modelo de ajuste exponencial (a), esférico (b) y gaussiano (c) para el variograma del 1 al 5 de mayo de 00:00 a 00:15 horas

### Interpolación usando kriging

El promedio de las mediciones registradas por estación activa de las 00:00 a las 00:15 se muestran en la Figura 4.15. A partir de ellos se realiza el kriging ordinario para interpolar los índices de humedad relativa en cada punto del AMM el día primero de mayo en el bloque de tiempo mencionado.

La interpolación basada en el modelo exponencial se muestra en la Figura 4.16, en la Figura 4.17 está el mapa de estimaciones producido por el modelo

esférico y en la Figura 4.18 aparece la estimación a partir del modelo gaussiano. Notamos que el índice de humedad más bajo se concentra en la zona sur del AMM, cerca de las estaciones ubicadas en el Parque Ecológico Chipinque, mientras que el índice de humedad más alto se concentró en la zona centro y poniente del AMM.

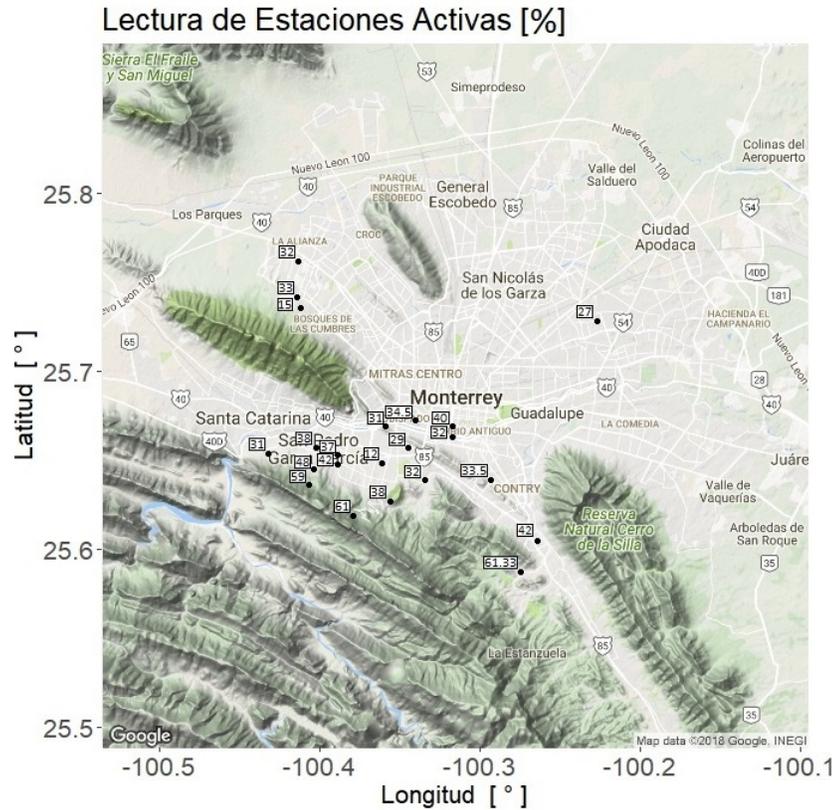


Figura 4.15: Lecturas de las estaciones activas del día primero de mayo de 2017, de las 00:00 a las 00:15 para los datos de humedad relativa

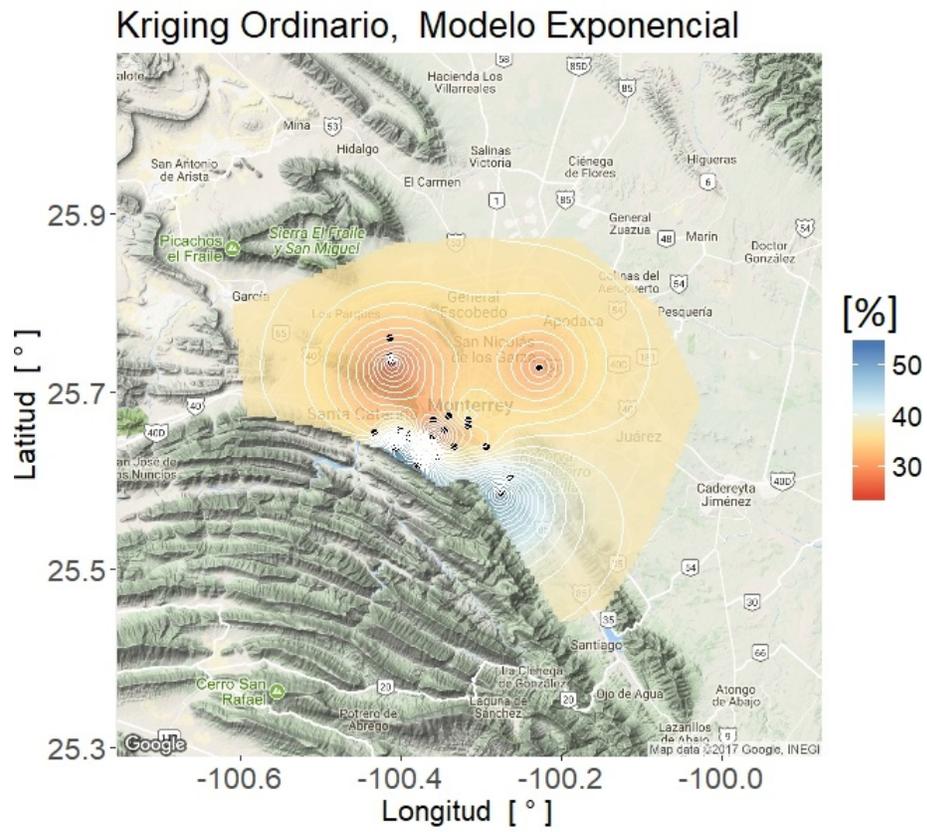


Figura 4.16: Estimación del índice de humedad relativa del día primero de mayo de 2017, de las 00:00 a las 00:15. Modelo exponencial.

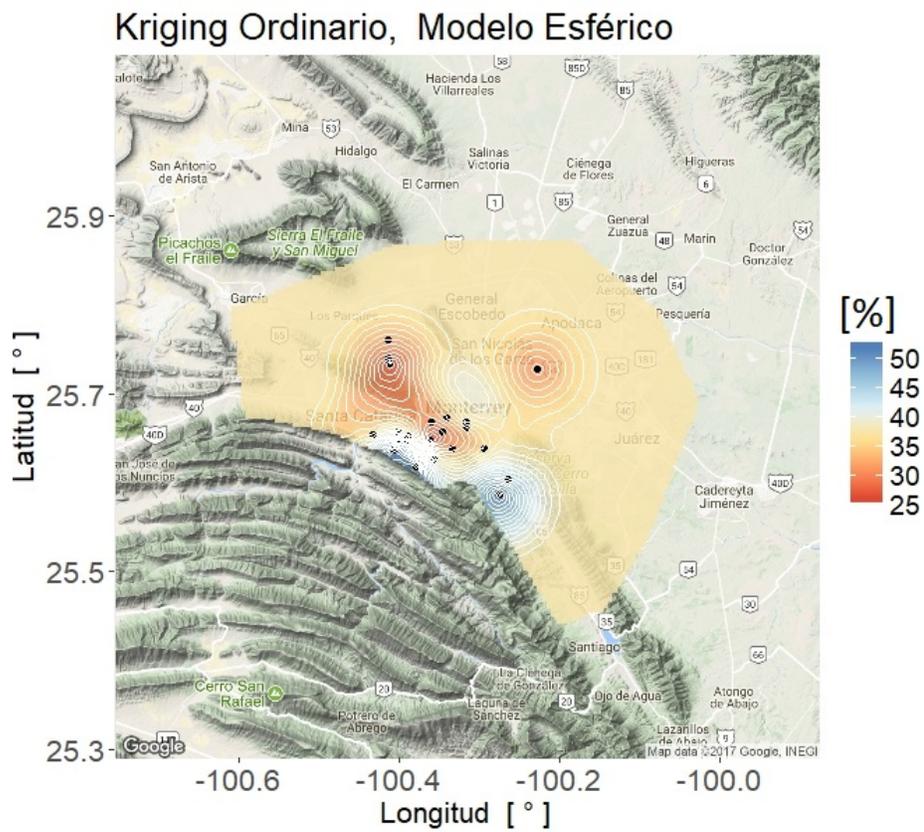


Figura 4.17: Estimación del índice de humedad relativa del día primero de mayo de 2017, de las 00:00 a las 00:15. Modelo esférico.

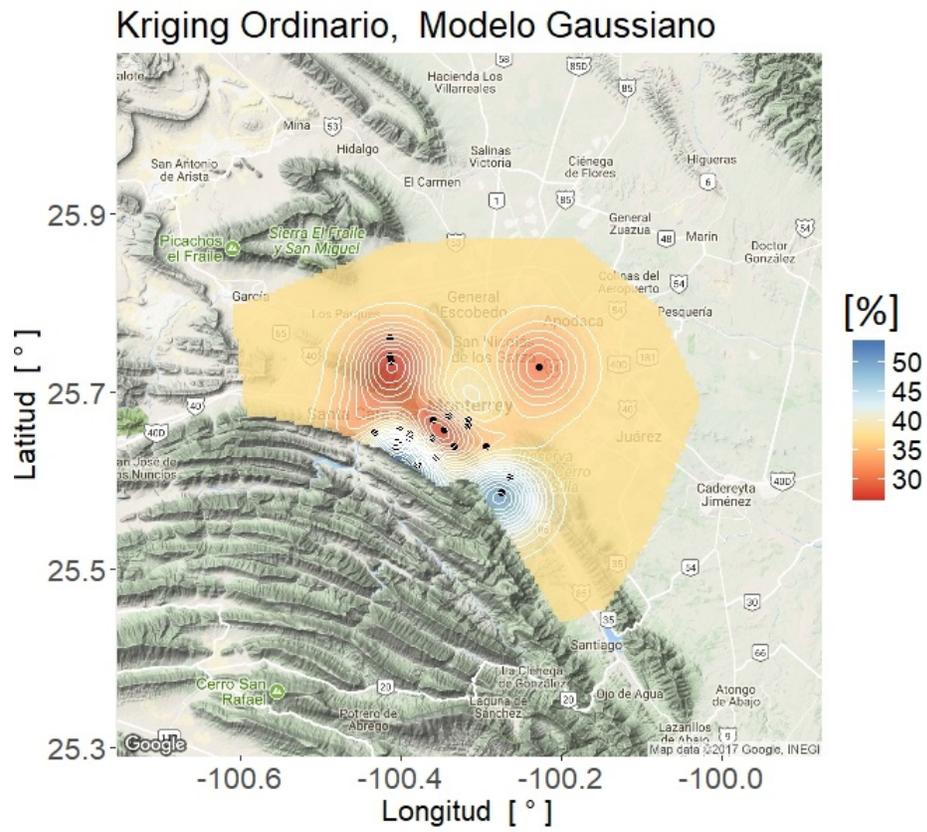


Figura 4.18: Estimación del índice de humedad relativa del día primero de mayo de 2017, de las 00:00 a las 00:15. Modelo gaussiano.

## Capítulo 5

# Aplicación de kriging en datos de presión atmosférica

En este capítulo se estima la presión atmosférica en el AMM usando kriging ordinario. La unidad en la que está expresada esta variable es el hectopascal (hPa).

### 5.1. Análisis del 1 al 5 de mayo de 2017, de 17:00 a 17:15

En esta sección se estimó la presión atmosférica en el AMM durante el intervalo de las 17:00 a las 17:15 del primero de mayo de 2017. Para estimar el variograma se usaron de nuevo las semivarianzas del 1 al 5 de mayo de 2017, durante el mismo periodo de quince minutos.

#### Análisis preliminar

Se realizó el análisis preliminar sobre las mediciones registradas con las que se realizará el variograma, para ver si existen valores atípicos y si están dentro del rango aceptable de asimetría. En este caso, para detectar valores atípicos se usa como criterio el doble del rango intercuartílico. La Figura 5.1 muestra que, en efecto, hubo valores atípicos registrados en cada uno de los días y que éstos se alejan considerablemente del resto de los datos. Para fines de la estimación de los modelos de variograma estos valores se omiten de forma provisional.

Una vez que se omiten los valores fuera de rango, se realiza el resumen estadístico que se muestra en la Tabla 5.1, mientras que los box-plot se ven ahora como se muestra en la Figura 5.2. Es importante remarcar que las mediciones de este periodo no son todas simétricas, contrario a lo deseable. En cambio, las mediciones del día 2, 4 y 5 de mayo están fuera del rango de asimetría  $\pm 1$ . A pesar de ello, se tomaron en cuenta los datos originales para el análisis puesto que, al tratar de transformar los datos usando su logaritmo, no ocurrió ningún cambio significativo: después de excluir valores fuera de rango, los coeficientes

de asimetría de los días 2, 4 y 5 pasaron de -1.4305, -1.3505 y -1.3601 a -1.4425, -1.3601 y -1.3725, respectivamente.

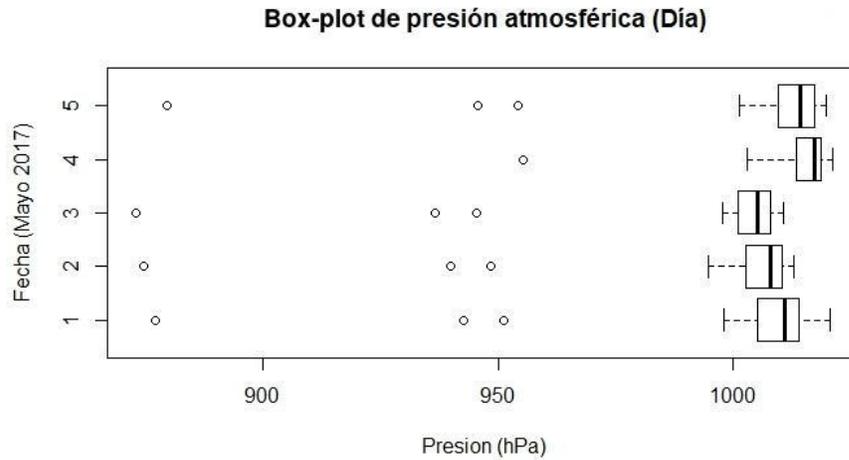


Figura 5.1: Box-plot de la presión atmosférica registrada del 1 al 5 de mayo de 2017, de las 17:00 a las 17:15

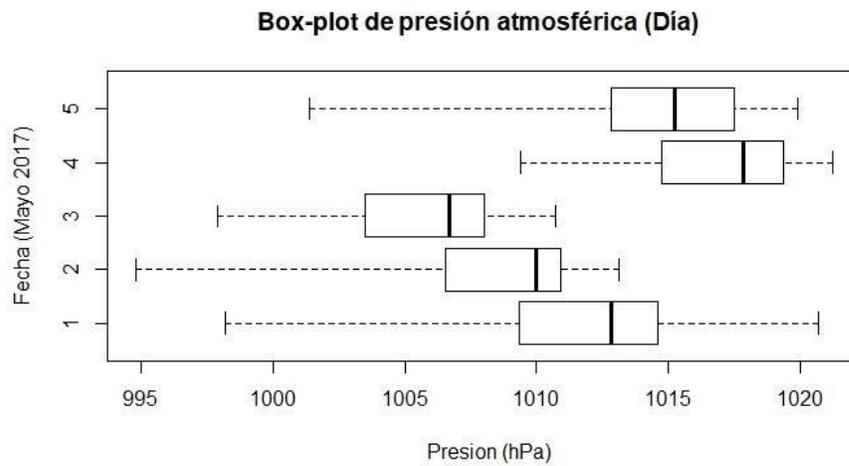


Figura 5.2: Box-plot de la presión atmosférica registrada del 1 al 5 de mayo de 2017, de las 17:00 a las 17:15, sin tomar en cuenta valores atípicos.

Tabla 5.1: Resumen estadístico de la presión atmosférica del 1 al 5 de mayo de 2017, de las 17:00 a las 17:15 horas.

Mayo 2017 (Día)					
Fecha	1	2	3	4	5
Mínimo	998.2	994.8	997.9	1003	1001
Primer cuartil	1009.2	1006.5	1003.3	1014	1013
Mediana	1013.0	1010.0	1006.9	1018	1016
Media	1011.3	1008.1	1005.7	1016	1014
Tercer cuartil	1014.9	1011.1	1008.2	1019	1018
Máximo	1020.7	1013.1	1010.7	1021	1020
Desviación estándar	5.2464	4.6055	3.6478	4.8421	4.6320
Varianza	27.5251	21.2109	13.3066	23.4461	21.4554
Asimetría	-0.7134	-1.4305	-0.6419	-1.3505	-1.3601
Número de datos	18	18	16	16	17

### Estimación del variograma

En la Figura 5.3 se muestra el variograma empírico del modelo geoestadístico de la presión atmosférica, que es el primer paso en el proceso de estimación. Posteriormente se calculó el variograma muestral. A diferencia de los anteriores, en éste se utilizaron únicamente ocho intervalos, y éstos sólo abarcan una proporción de 0.4 de la distancia máxima, es decir  $\frac{2d_{m\acute{a}x}}{5} = \frac{2 \cdot 19.036}{5} = 7.614$  km. Dicha proporción entra en el rango aceptable de  $[\frac{1}{3}, \frac{1}{2}]$

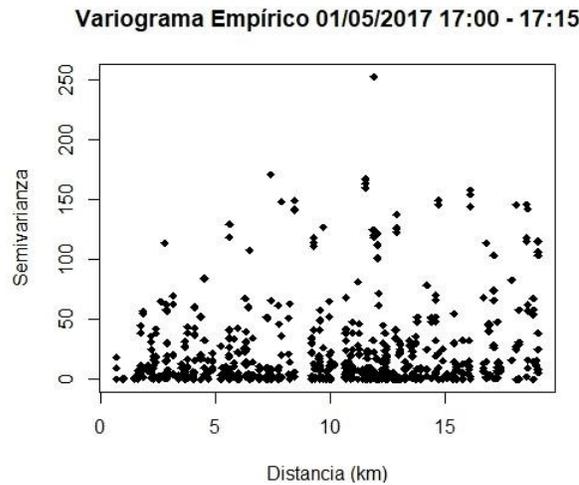


Figura 5.3: Variograma empírico de la presión atmosférica registrada del 1 al 5 de mayo de 2017, de las 17:00 a las 17:15 horas.

Se tomaron parámetros distintos para este variograma muestral puesto que

al estimar las semivarianzas para distancias mayores a 7.614 km, los valores oscilan fuertemente, y se presentan outliers. No se usaron diez intervalos puesto que las estimaciones de las semivarianzas se vuelven difícil de modelar debido a la poca cantidad de datos para estimar cada una.

Tabla 5.2: Intervalos de distancia, semivarianza y conteo de estimaciones de la presión atmosférica registrada del 1 al 5 de mayo de 2017, de las 17:00 a las 17:15 horas.

Distancia (km)	Semivarianza	Número de estimaciones
0.476	5.691889	5
1.428	11.780355	22
2.380	14.855405	40
3.331	16.157409	48
4.283	15.330935	32
5.235	17.583346	34
6.187	15.975430	43
7.139	16.447118	25

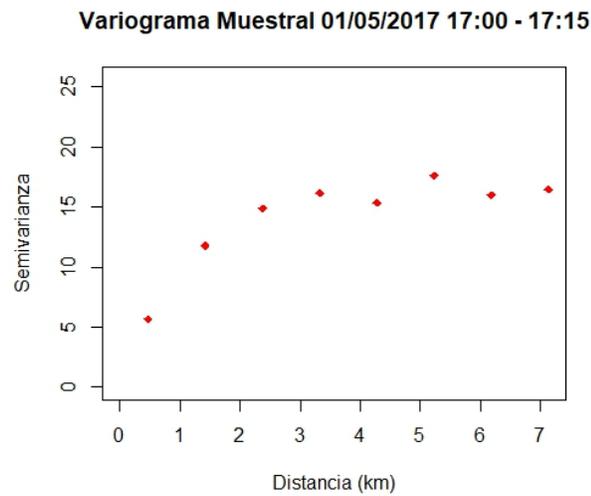


Figura 5.4: Variograma muestral de la presión atmosférica registrada del 1 al 5 de mayo de 2017 de las 17:00 a las 17:15 horas.

Se observa que a partir de la distancia  $h = 3.331$  km la semivarianza deja

de incrementarse y parecen estabilizarse, por lo que es probable que el rango de este variograma sea alrededor de dicha distancia.

**Modelo de ajuste del variograma**

En la Tabla 5.3 se muestran los parámetros obtenidos al aplicar los tres tipos de modelos de ajuste sobre el variograma muestral. En las últimas dos columnas se incluyen la SRC y la MRC como indicador de ajuste a los datos del variograma muestral de cada modelo.

De acuerdo a dichos indicadores, es el modelo exponencial el que tiene mejor ajuste, con un umbral de 16.59939 y un rango efectivo de 3.358 km y sin varianza nugget, parámetros que aparecen en la Tabla 5.4 junto con los que corresponden a los demás modelos. El umbral de los tres modelos es, en general, alrededor de 16.3, mientras que el rango efectivo varía de 2.6 km a 3.4 km.

Las gráficas de los tres modelos de ajuste, el exponencial, el esférico y el gaussiano, se muestran en las Figuras 5.5a, 5.5b y 5.5c, respectivamente. En los tres casos se puede suponer cierta la hipótesis de estacionariedad puesto que los tres tienen umbral y son acotados.

Tabla 5.3: Parámetros de los modelos de ajuste para el variograma del 1 al 5 de mayo de 17:00 a 17:15 horas

Tipo de modelo	Parámetros del modelo			Discrepancia	
	$c_0$	$c$	$a$ (km)	SRC	MRC
Modelo exponencial	0	16.59939	1.119	2.7015	0.33769
Modelo esférico	2.76674	13.50646	3.118	3.0432	0.38040
Modelo gaussiano	4.68212	11.58636	1.513	2.9879	0.37349

Tabla 5.4: Parámetros del variograma para cada modelo de ajuste del 1 al 5 de mayo de 17:00 a 17:15 horas

Tipo de modelo	Varianza nugget	Umbral	Rango efectivo (km)
Modelo exponencial	0	16.59939	3.358
Modelo esférico	2.76674	16.27320	3.118
Modelo gaussiano	4.68212	16.26848	2.618

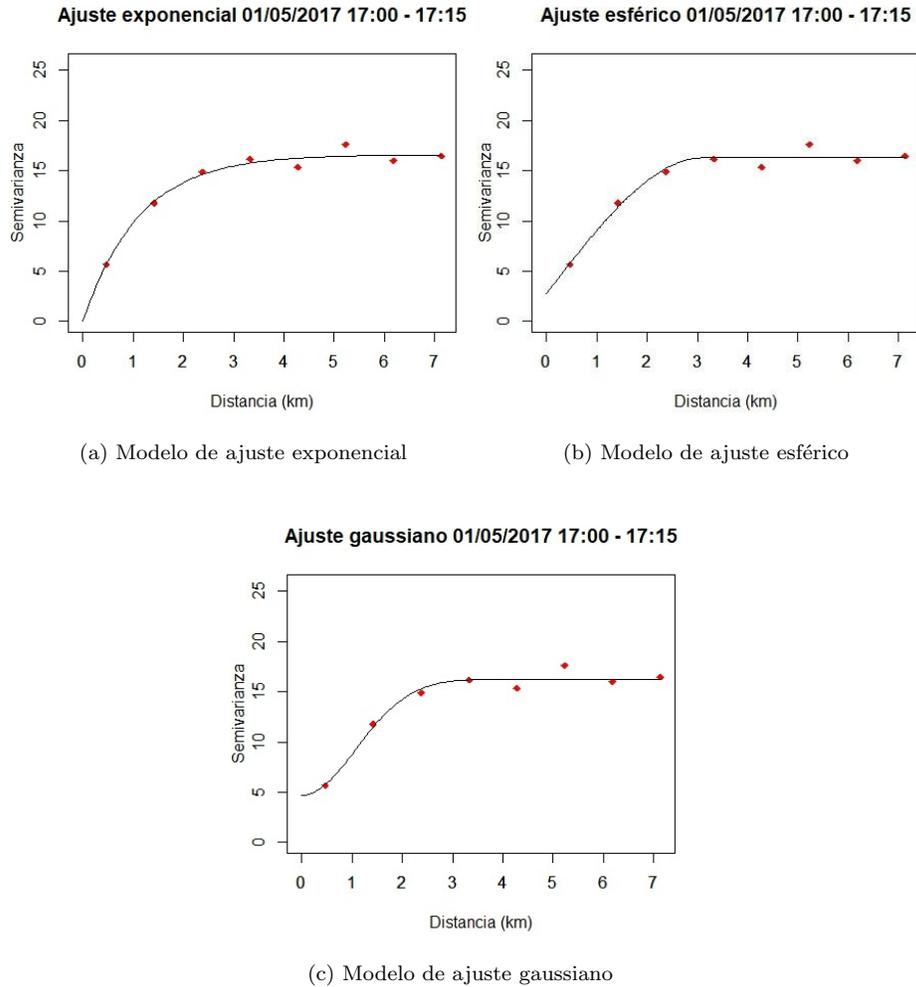


Figura 5.5: Modelo de ajuste exponencial (a), esférico (b) y gaussiano (c) para el variograma del 1 al 5 de mayo de 17:00 a 17:15

### Interpolación usando kriging

Finalmente, se calculó la estimación de la presión atmosférica para cada punto del AMM utilizando kriging ordinario sobre los promedios de lecturas por cada estación del 1 de mayo de 2017, en el periodo de 15 minutos a partir de las 17:00. Estas observaciones son mostradas en la Figura 5.6. La aplicación de kriging ordinario basado en el modelo exponencial de variograma se muestra en la Figura 5.7, el basado en el modelo esférico aparece en la Figura 5.8, y por último la Figura 5.9 es la estimación hecha a partir del modelo gaussiano.

Aunque, de acuerdo a estas imágenes, las tres estimaciones lucen muy similares, la estimación basada en el modelo exponencial parece ser el mejor modelo dado el ajuste que tuvo su variograma.

Antes de continuar, notamos que existen dos áreas que registran presión atmosférica menor al resto, y están ubicadas en las faldas del Parque Ecológico Chipinque, al sur de AMM y, más al norte de ésta, en la zona del cerro del Obispado. Esto puede explicarse parcialmente a través de la fórmula barométrica, que establece que, dada una temperatura constante, la presión atmosférica depende de la altura:

$$P = P_0 \cdot \left[ \frac{T}{T + L \cdot h} \right]^{\frac{g \cdot M}{R \cdot L}},$$

donde

- $P_0$  es la presión atmosférica promedio a nivel del mar: 1013.25 hPa,
- $T$  representa la temperatura (K). Puede considerarse como 298.15° K, que en Celcius equivale a 25° C,
- $L$  es la tasa de caída de temperatura:  $-0.0065^\circ \text{ K} \cdot \text{m}^{-1}$ ,
- $h$  es la altura (m) sobre el nivel del mar,
- $R$  es la constante universal de los gases ideales:  $8.3144598 \text{ J} \cdot \text{mol}^{-1} \cdot \text{K}^{-1}$ ,
- $g$  es la aceleración gravitacional:  $9.80665 \text{ m} \cdot \text{s}^{-2}$  y
- $M$  es la masa molar del aire terrestre:  $0.0289644 \text{ kg} \cdot \text{mol}^{-1}$ .

Utilizando dicha fórmula con una altura  $h = 1,260$  m sobre el nivel del mar, que según el Instituto Nacional de Estadística y Geografía (2014) es la altura aproximada de la estación en las coordenadas  $(-100.379^\circ, 25.6185^\circ)$ , en la zona del Parque Ecológico Chipinque en San Pedro Garza García, la presión atmosférica sería  $P = 875.2635$  hPa. Dicho valor teórico es similar al valor real de 877 hPa.

Al hacer el mismo procedimiento con la estación ubicada en las coordenadas  $(-100.341^\circ, 25.673^\circ)$ , en el Cerro del Obispado, a una altura aproximada de  $h = 700$  m de acuerdo con el INEGI, se obtiene  $P = 933.8057$  hPa. como valor teórico de presión atmosférica, contra un valor real de 942.7 hPa.

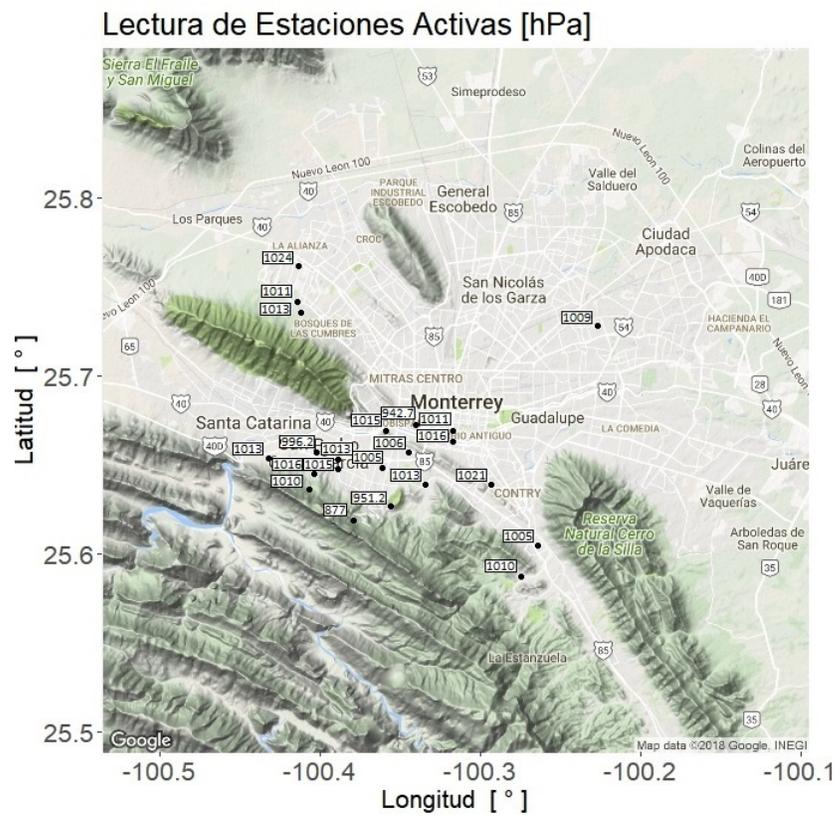


Figura 5.6: Lecturas de las estaciones activas del día primero de mayo de 2017, de las 17:00 a las 17:15

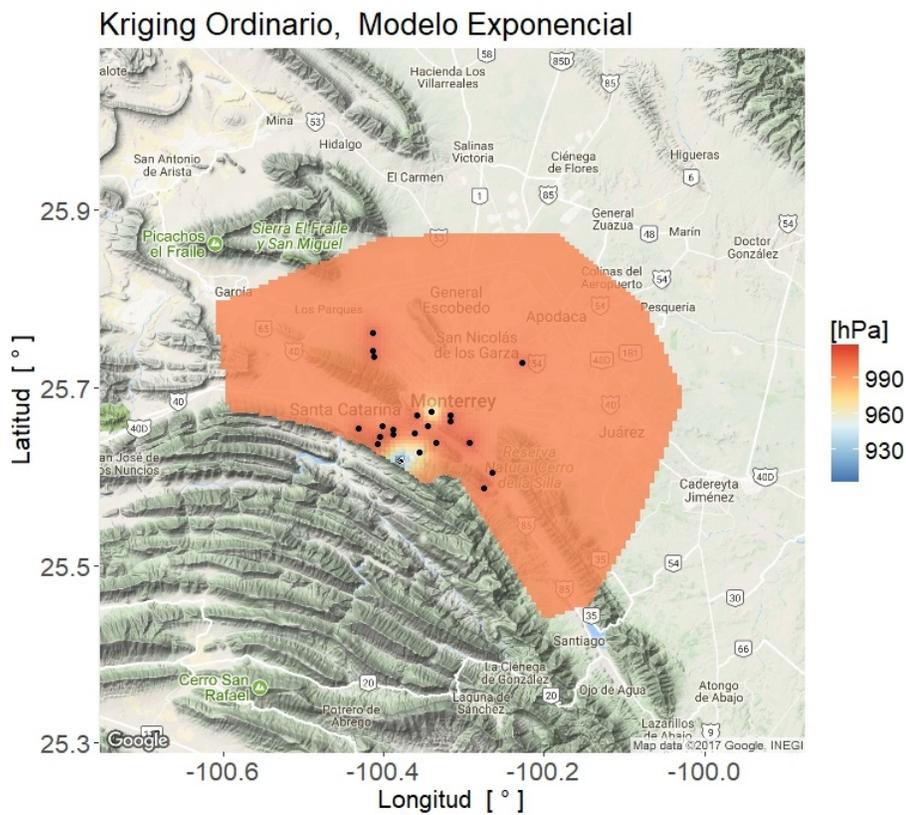


Figura 5.7: Estimación de la presión atmosférica del día primero de mayo de 2017, de las 17:00 a las 17:15. Modelo exponencial.

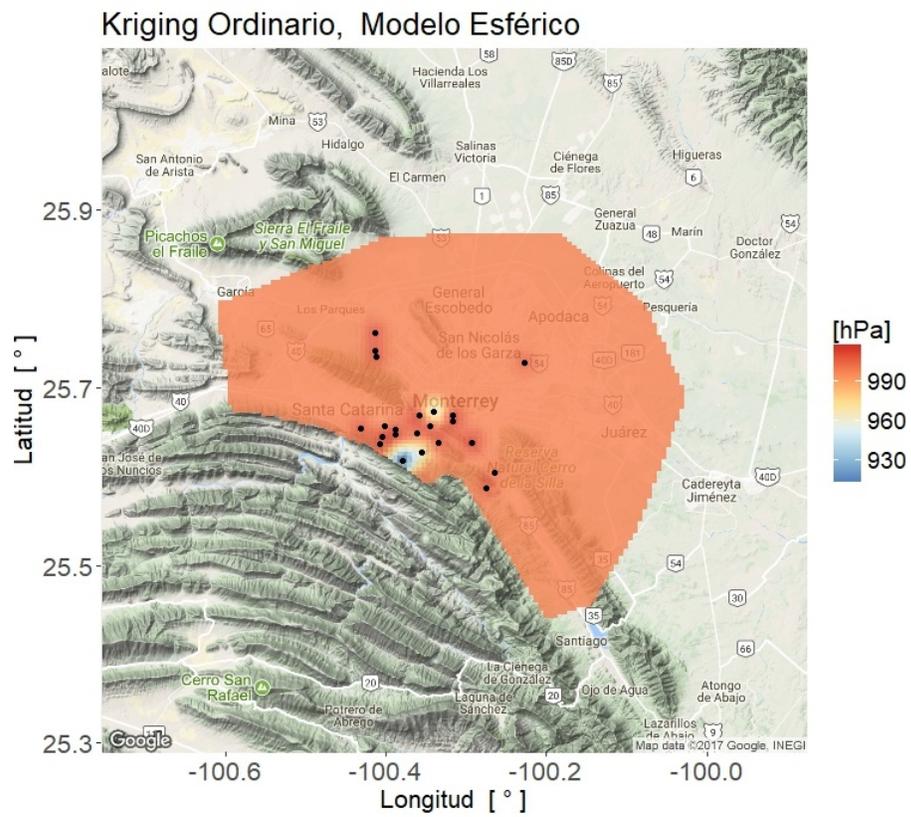


Figura 5.8: Estimación de la presión atmosférica del día primero de mayo de 2017, de las 17:00 a las 17:15. Modelo esférico.

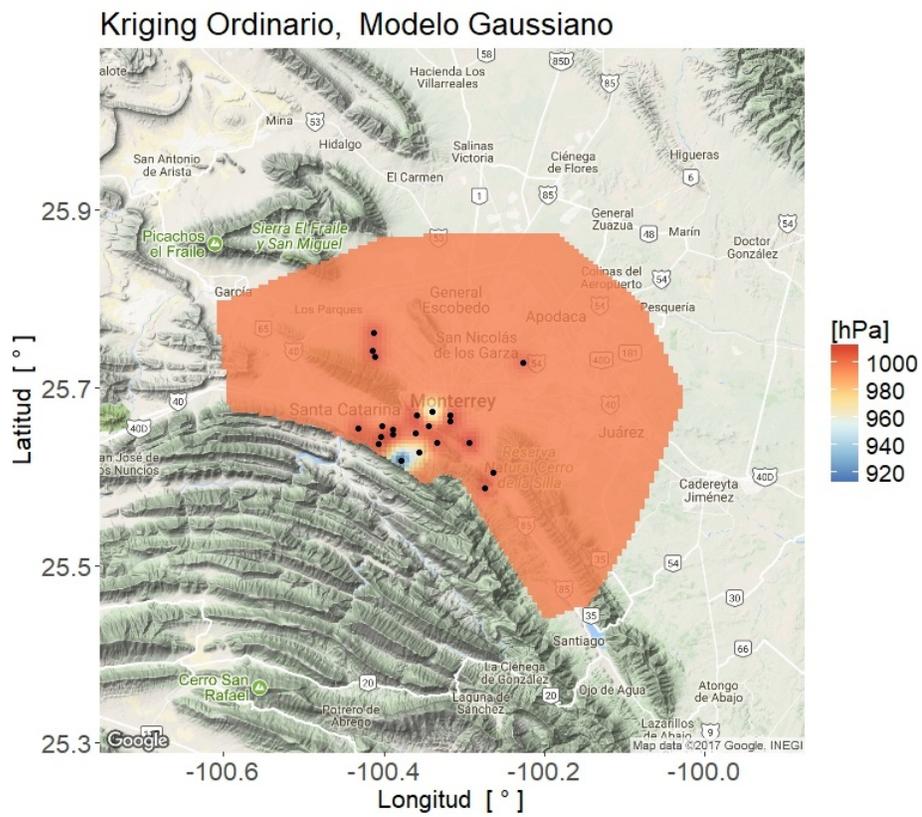


Figura 5.9: Estimación de la presión atmosférica del día primero de mayo de 2017, de las 17:00 a las 17:15. Modelo gaussiano.

## 5.2. Análisis del 1 al 5 de mayo de 2017, de 00:00 a 00:15

La última colección de datos que fue analizada y utilizada para hacer kriging fueron las mediciones de presión atmosférica registradas del 1 al 5 de mayo, de las 00:00 a las 00:15 horas. Al igual que en la sección precedente, dichas mediciones fueron registradas en hectopascales.

### Análisis preliminar

Como primer tarea en este procedimiento, se considera un análisis preliminar para descartar posibles valores atípicos en los datos. Como criterio para descartar este tipo de valores se tomó como parámetro el doble del rango intercuartílico. En la Figura 5.10 se muestra el box-plot de la presión atmosférica registrada del 1 al 5 de mayo de 2017, de las 00:00 a las 00:15 horas. La incidencia de valores fuera de rango es similar al caso anterior, y una vez omitidos el box-plot de los datos actualizados se convierte en la mostrada en la Figura 5.11.

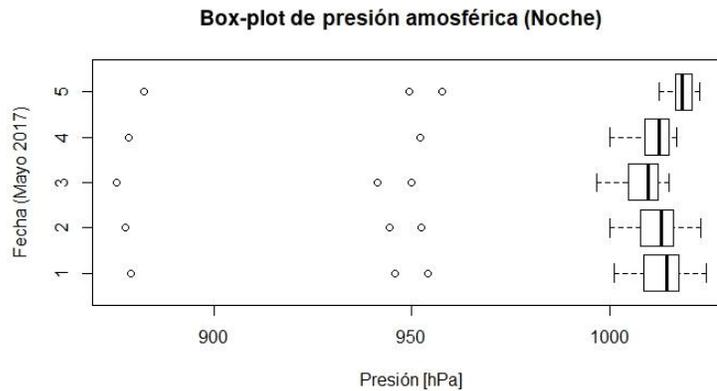


Figura 5.10: Box-plot de la presión atmosférica registrada del 1 al 5 de mayo de 2017, de las 00:00 a las 00:15 horas.

En promedio hubo 18 estaciones disponibles por día, y la noche con menos estaciones fue la de la madrugada del 5 de mayo, con sólo 15 estaciones. Se observa de la Figura 5.11 y puede corroborarse en la Tabla 5.5 que los días 3 y 4 son particularmente asimétricos, al tomar en cuenta que el índice de asimetría sale del rango de  $\pm 1$ . Se siguió trabajando con los datos originales puesto que, al considerar los logaritmos de estos datos, después de haber descartado valores atípicos, los coeficientes de asimetría de los días mencionados pasa de -1.3406 y -1.2675 a -1.3521 y -1.2772, respectivamente.

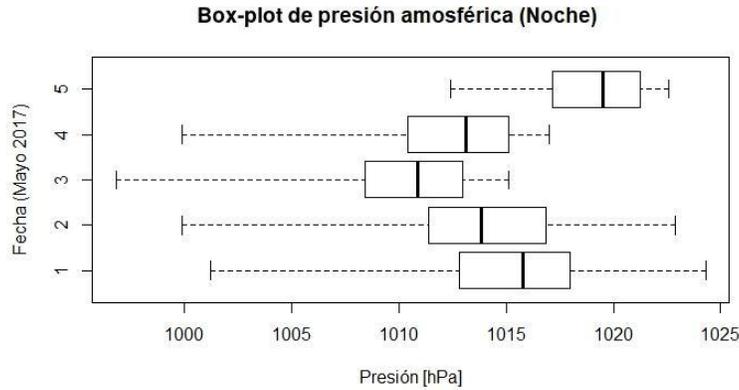


Figura 5.11: Box-plot de la presión atmosférica registrada del 1 al 5 de mayo de 2017, de las 00:00 a las 00:15 horas. Valores atípicos excluidos.

Tabla 5.5: Resumen estadístico de los datos de presión atmosférica registrados del 1 al 5 de mayo de 2017, de las 00:00 a las 00:15 horas.

Mayo 2017 (Día)					
Fecha	1	2	3	4	5
Mínimo	1001	999.9	996.8	999.9	1012
Primer cuartil	1013	1011.4	1008.4	1010.4	1017
Mediana	1016	1014.4	1011.7	1013.8	1020
Media	1015	1013.2	1009.7	1012.1	1019
Tercer cuartil	1018	1016.9	1013.0	1015.1	1021
Máximo	1024	1022.9	1015.1	1017.0	1023
Desviación estándar	5.3115	5.1633	4.6756	4.5658	2.9419
Varianza	28.2121	26.6600	21.8611	20.8466	8.6548
Asimetría	-0.7250	-0.7133	-1.3406	-1.2675	-0.6939
Número de datos	18	19	17	17	14

### Estimación del variograma

Las semivarianzas en el variograma empírico, mostrado en la Figura 5.12, se toman en cuenta hasta la distancia  $0.6 \cdot d_{\text{máx}} = 0.6 \cdot 20.016 = 12.010$  y se agrupan en diez intervalos para calcular el variograma muestral, que puede verse en la Figura 5.13. En la Tabla 5.6 se incluyen las estimaciones de semivarianza por cada intervalo.

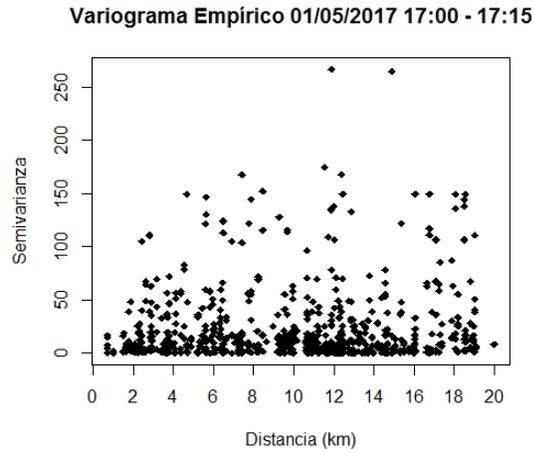


Figura 5.12: Variograma empírico de los datos de presión atmosférica registrados del 1 al 5 de mayo de 2017, de las 00:00 a las 00:15 horas.

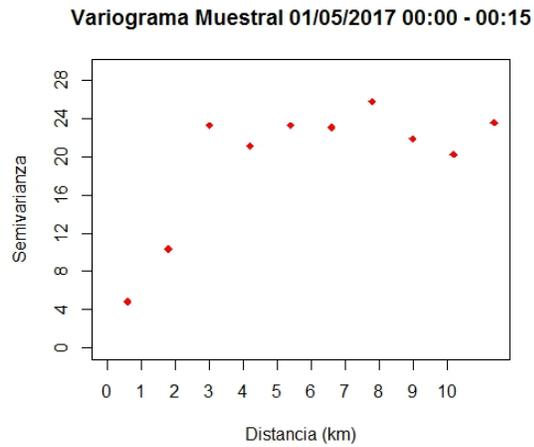


Figura 5.13: Variograma muestral de los datos de presión atmosférica registrados del 1 al 5 de mayo de 2017 de las 00:00 a las 00:15 horas.

Tabla 5.6: Intervalos de distancia, semivarianzas y conteo de estimaciones de los datos de presión atmosférica registrados del 1 al 5 de mayo de 2017, de las 00:00 a las 00:15 horas.

<b>Distancia (km)</b>	<b>Semivarianza</b>	<b>Número de estimaciones</b>
0.600	4.848028	10
1.801	10.350324	33
3.002	23.220810	47
4.203	21.078787	53
5.404	23.237322	44
6.605	23.033748	43
7.806	25.739302	44
9.007	21.847935	34
10.208	20.191612	54
11.409	23.539521	68

En esta ocasión, el variograma muestral abarca más allá de la mitad de la distancia máxima, contrario a lo que se había recomendado en la sección 2.6. Esto se debe a que, de considerar una distancia menor, el variograma no tiene una forma que aparente convergencia asintótica, y por lo tanto, no se distingue claramente el rango efectivo del modelo. Este fenómeno se puede apreciar en la Figura 5.14.

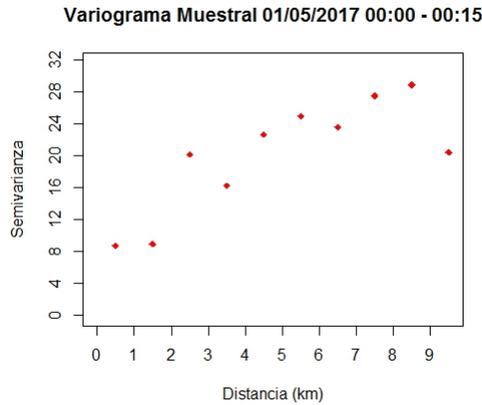


Figura 5.14: Variograma muestral, calculado hasta una distancia de 10.008 km, de los datos de presión atmosférica registrados del 1 al 5 de mayo de 2017 de las 00:00 a las 00:15 horas.

### Modelo de ajuste del variograma

Se aplicaron sobre el variograma muestral los tres tipos de modelo de variograma con los que se ha trabajado. Los parámetros obtenidos por cada modelo se incluyen en la Tabla 5.7. La SRC y MRC indican que el modelo con mejor ajuste es el gaussiano, cuyo rango efectivo es  $1.73 \cdot 2.163 = 3.742$  km, que tiene un umbral de 22.97531 y cuya varianza nugget es de 2.50788.

Tabla 5.7: Parámetros de los modelos de ajuste para el variograma del 1 al 5 de mayo de 00:00 a 00:15 horas

Tipo de modelo	Parámetros del modelo			Discrepancia	
	$c_0$	$c$	$a$ (km)	SRC	MRC
Modelo exponencial	0	23.58123	1.928	59.96158	5.996158
Modelo esférico	0	22.80824	4.298	40.29131	4.029131
Modelo gaussiano	2.50788	20.46743	2.163	35.93564	3.593564

Tabla 5.8: Parámetros del variograma para cada modelo de ajuste del 1 al 5 de mayo de 00:00 a 00:15 horas

Tipo de modelo	Varianza nugget	Umbral	Rango efectivo (km)
Modelo exponencial	0	23.58123	5.784
Modelo esférico	0	22.80824	4.298
Modelo gaussiano	2.50788	22.97531	3.742

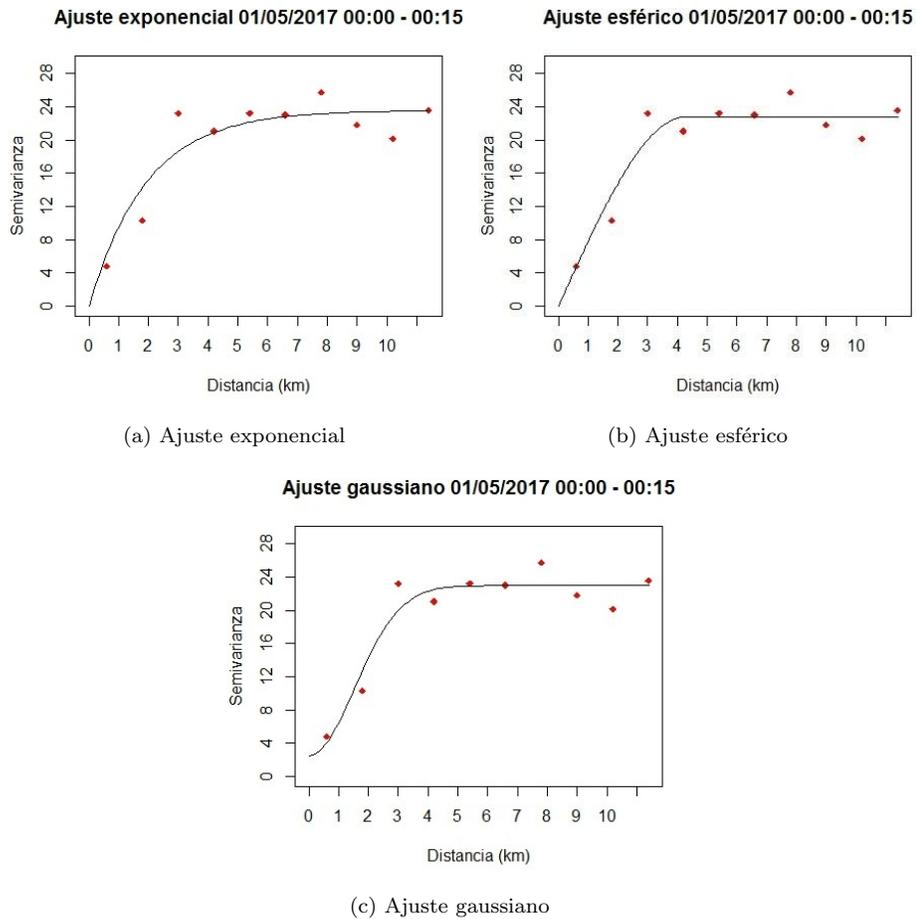


Figura 5.15: Modelo de ajuste exponencial (a), esférico (b) y gaussiano (c) para el variograma del 1 al 5 de mayo de 00:00 a 00:15 horas

La gráfica del modelo exponencial puede observarse en la Figura 5.15a, al igual que las gráficas de los modelos esférico y gaussiano se incluyen en las Figuras 5.15b y 5.15c, respectivamente.

### Interpolación usando kriging

Por último, se muestran los mapas que incluyen las estimaciones por cada punto del AMM. En la Figura 5.17 se encuentra la estimación basada en el modelo exponencial, en la Figura 5.18 se encuentra la estimación basada en el modelo esférico, y en la Figura 5.19, la elaborada a partir del modelo gaussiano. Los datos muestrales sobre los que se hacen las estimaciones se pueden apreciar en la Figura 5.16.

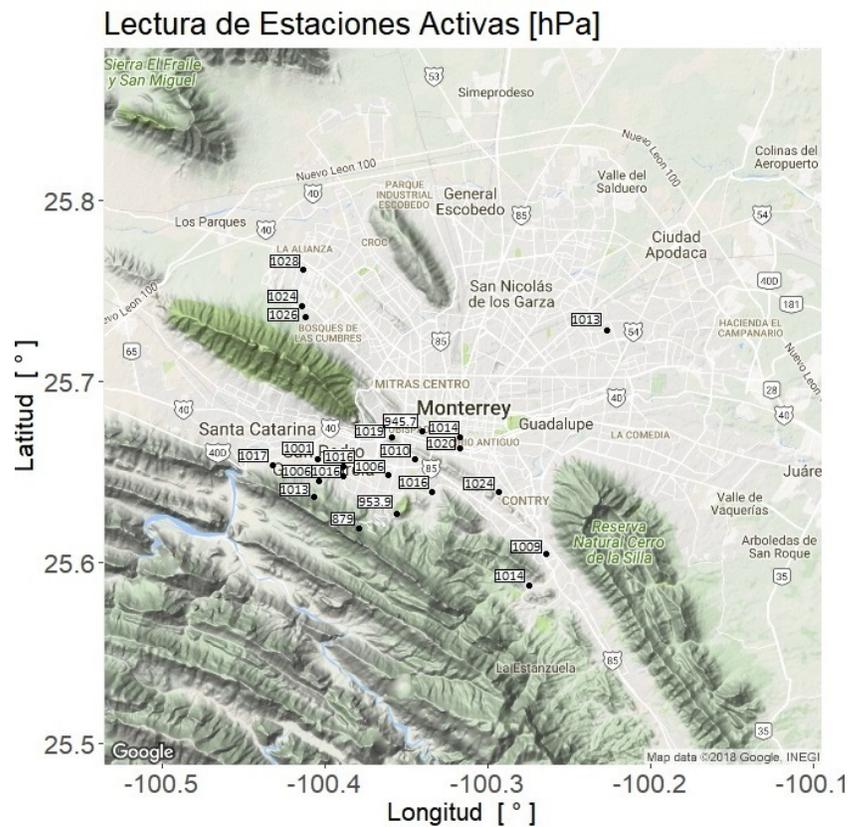


Figura 5.16: Lecturas de presión atmosférica de las estaciones activas del día primero de mayo de 2017, de las 00:00 a las 00:15

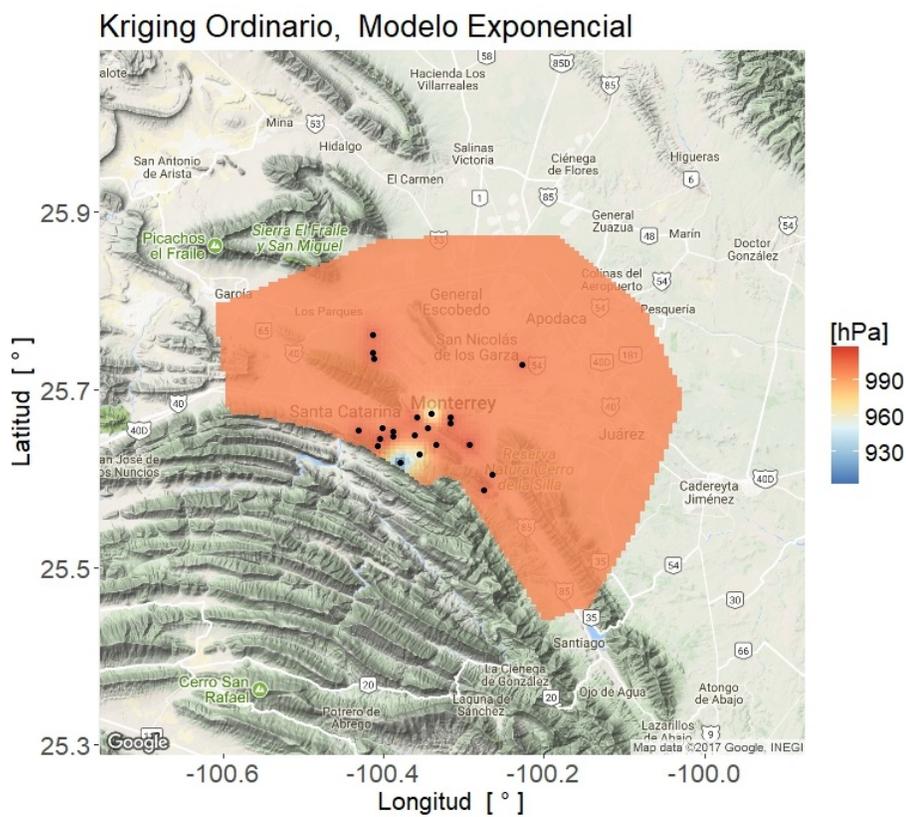


Figura 5.17: Estimación de la presión atmosférica del día primero de mayo de 2017, de las 00:00 a las 00:15. Modelo exponencial.

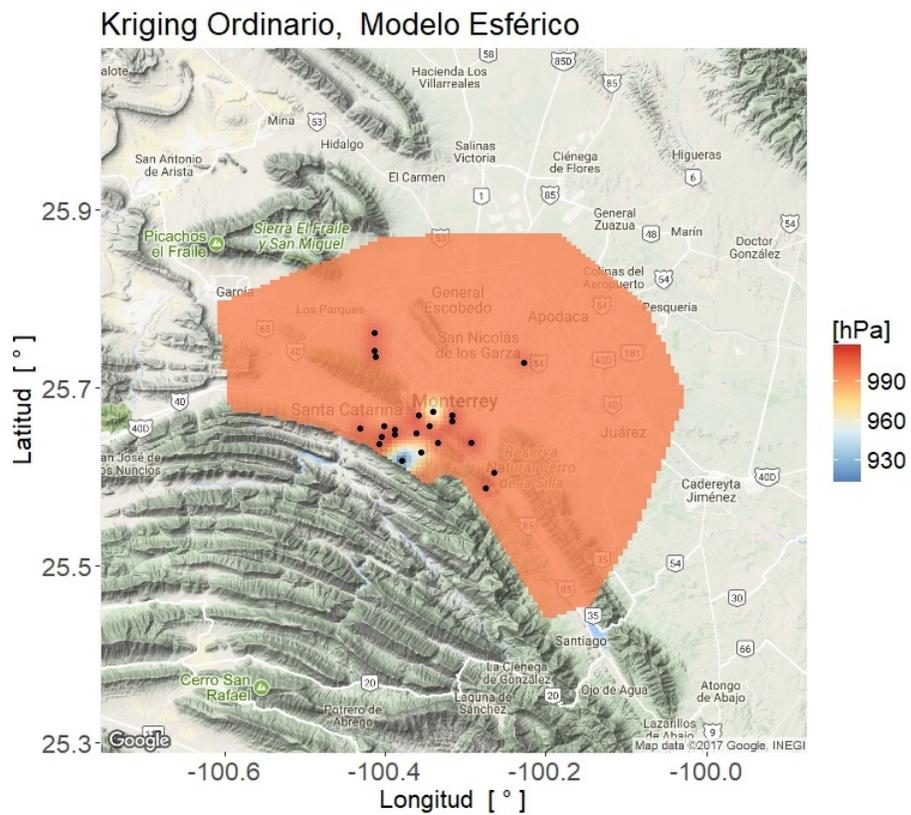


Figura 5.18: Estimación de la presión atmosférica del día primero de mayo de 2017, de las 00:00 a las 00:15. Modelo esférico.

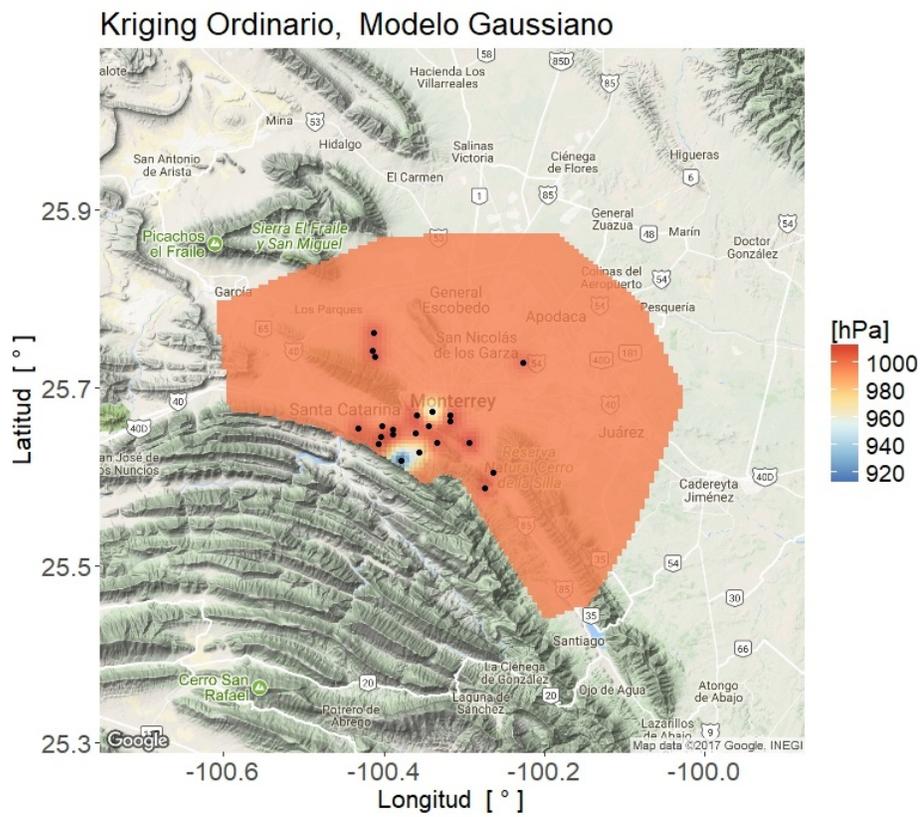


Figura 5.19: Estimación de la presión atmosférica del día primero de mayo de 2017, de las 00:00 a las 00:15. Modelo gaussiano.



## Capítulo 6

# Conclusión

Como fue mencionado en la sección 1.1, dos de los objetivos de esta tesis eran el estudio del método de kriging como estimador de variables climatológicas para el área metropolitana de Monterrey, así como identificar diferencias entre la aplicación de este método en los contextos que aquí se abordan. Esta metodología, junto con el análisis geoespacial propio de ella, se llevó a cabo en los capítulos 3, 4 y 5.

Durante el proceso de implementación se observaron algunas importantes características del modelo que deben ser tomadas en cuenta. En el proceso de modelar la correlación espacial, se observó que fue bueno el ajuste de los semivarianzas de la temperatura, y de la presión atmosférica durante el día, presentados en el capítulo 3 y en la sección 5.1, respectivamente. Por el contrario, en el capítulo 4 se muestra que no hubo buen ajuste del variograma muestral del índice de humedad en ninguno de los dos tiempos del día.

De los tres tipos de modelos estudiados, el gaussiano fue el que más veces se ajustó mejor a los datos con tres casos: con datos de presión atmosférica a las 00:00 horas (Tabla 5.7) y con datos de temperatura a las 00:00 y 17:00 horas (Tablas 3.7 y 3.3); seguido del modelo exponencial con dos casos: con datos de presión atmosférica a las 17:00 horas (Tabla 5.3) y con datos de humedad relativa a las 00:00 horas (Tabla 4.7); mientras que el último fue el esférico, que sólo resultó tener mejor ajuste con las lecturas del índice de humedad obtenidas durante el día (Tabla 4.3). Al analizar la temperatura, el modelo gaussiano tuvo el mejor ajuste los dos momentos en que se hizo la estimación. En cuanto a las demás variables, no hubo un modelo que ajustara mejor en ambos momentos del día.

Un aspecto a tomar en cuenta al utilizar este método de estimación para distintos momentos del día son las diferencias entre los modelos de variograma estimados para el día y para la noche. Al estudiar las tres variables climatológicas, tanto la varianza nugget como el umbral y el rango de los modelos de

variograma aumentaron considerablemente al cambiar de las mediciones obtenidas en el día a las de la noche. Esta observación fue particularmente notoria al estimar la temperatura, donde el rango del variograma exponencial aumentó de 2.535 km a 19.378 km (ver Tabla 3.8). Cabe mencionar también que, al estimar la temperatura a las 17:00 horas, el rango de correlación espacial fue muy pequeño y, como resultado, las estimaciones resultantes fueron homogéneas para la mayor parte del AMM.

El análisis preliminar de los datos contempló, en lo general, excluir algunas mediciones que dificultaban el estudio de la correlación espacial en la que se elaboran los distintos modelos de variograma. En particular, se detectó que, en los seis procesos de estimación, la estación ubicada en las coordenadas ( $-100.414^\circ$ ,  $25.7415^\circ$ ), al poniente del AMM, dificultaba la estimación el variograma.

El relieve desigual de la ciudad es otra limitante del método de estimación tal como fue planteado, que supone que las estaciones tienen condiciones similares y, por lo tanto, omiten la altitud del terreno como variable. Según el modelo geostatístico, dos estaciones a distancia  $h_0$ , ubicadas en altitudes similares, tienen la misma correlación espacial que dos estaciones a la misma distancia, ubicadas en altitudes completamente distintas, aunque esto no sea necesariamente cierto.

Una posible solución para tratar de reducir el impacto de este parámetro en la que se hallan las estaciones es agregar algún factor que dependa linealmente de la altitud. Dicho factor puede sumarse a las distintas mediciones para hacer kriging sobre los datos modificados y, una vez realizada la interpolación, se elimina dicho factor para estimar en cada punto.

Por último, cabe mencionar que se cumplió también con el último de los objetivos planteados en la sección 1.1, que era hacer la implementación de kriging en el paquete estadístico R que pudiera adaptarse a distintos parámetros. Gracias a esta implementación, este código podría utilizarse en el futuro como base para la estimación geostatística de otra variable como la concentración de contaminantes en el AMM. Asimismo, la información recabada de la estimación de variables meteorológicas y de contaminantes podría utilizarse en trabajos posteriores para analizar la correlación entre ellos en el contexto de la metrópoli regiomontana.

## Apéndice A

# Código de R de las funciones definidas para todo el proceso

Las funciones definidas en este código se utilizan tanto en el análisis preliminar (Apéndice B), como en el proceso de kriging (Apéndice C).

```
library("lubridate")
library("sp")
library("nnls")
library("calibrate")
library("ggplot2")
library("ggmap")
library("RColorBrewer")
library("moments")
rm(list=ls())

### Establecer el directorio correspondiente ####
setwd("/home/ferjosafath/Documents/Tesis/Codigo/")
#####

#####
###      get_obs      ### =
#####
#   *** NOTA ***
#   En esta funcion se modifica el tipo de dato a estudiar
#
#   wdata.....contiene los registros en intervalo de
#           tiempo
#   coord_stations..son las coordenadas de las estaciones
#   blocked.....estaciones con valores atipicos
#####
#####
# Funcion addalpha() para incluir un nivel alfa a los
# colores para obtener transparencia
# Usar como: colorwithalpha = addalpha(color, alphavalue)
# donde color puede ser, por ejemplo "#0007F" o un
```

```

# alfa-valor esta en el rango (0,1]
#####
addalpha <- function(colors, alpha=1.0) {
  r = col2rgb(colors, alpha=T)
  # Apply alpha
  r[4,] = alpha*255
  r = r/255.0
  return(rgb(r[1,], r[2,], r[3,], r[4,]))
}

#####
# la funcion colorRampPaletteAlpha() crea una paleta de
# colores incluyendo transparencia
# Usar como: colors = colorRampPaletteAlpha(basicColors,
#                                           resolution)
# donde basicColors es un arreglo de colores definiendo la
# paleta de colores, y resolution es el numero de colores
# diferentes a ser construidos dentro de la escla definida
# por los colores en basicColors
# Por ejemplo, para definir una escala linear de colores
# con 80 colores entre azul y rojo, uno tiene que definir
# > res = 80
# > mybasics = c("blue","red")
# > colors = colorRampPaletteAlpha(mybasics, res)
#####

colorRampPaletteAlpha <- function(colors, n=32,
                                  interpolate='linear') {
  # Create the color ramp normally
  cr = colorRampPalette(colors, interpolate=interpolate)(n)
  # Find the alpha channel
  a = col2rgb(colors, alpha=T)[4,]
  # Interpolate
  if (interpolate=='linear') {
    l = approx(a, n=n)
  } else {
    l = spline(a, n=n)
  }
  l$y[l$y > 255] = 255 # Clamp if spline is > 255
  cr = addalpha(cr, l$y/255.0)
  return(cr)
}

#####
# Construccion del Jet color palette usando transparencias
#####
JetPalletTransparency <- function(resolution, alpha){
  col9 = addalpha("#00007F", alpha)
  col8 = addalpha("blue", alpha)
  col7 = addalpha("#007FFF", alpha)
}

```

```

col6 = addalpha("cyan"    , alpha)
col5 = addalpha("#7FFF7F", alpha)
col4 = addalpha("yellow"  , alpha)
col3 = addalpha("#FF7F00", alpha)
col2 = addalpha("red"     , alpha)
col1 = addalpha("#7F0000", alpha)
basicColors = c(col1,col2,col3,col4,col5,col6,col7,col8,
                col9)

colors = colorRampPaletteAlpha(basicColors, resolution)
return(colors)
}

#####
# Construccion de la palteta de color azul usando
# transparencias
#####
BluePalletTransparency <- function(resolution, alpha){
  col1 = addalpha("#0000FF", alpha)
  col2 = addalpha("white", alpha)
  basicColors = c(col1,col2)
  colors = colorRampPaletteAlpha(basicColors, resolution)
  return(colors)
}

#####
# Construccion de la palteta de color rojo usando
# transparencias
#####
RedPalletTransparency <- function(resolution, alpha){
  col1 = addalpha("#EE0000", alpha)
  col2 = addalpha("white", alpha)
  basicColors = c(col1,col2)
  colors = colorRampPaletteAlpha(basicColors, resolution)
  return(colors)
}

#####
# Plot de contorno llenado con transparencia
#####
# Funcion para plotear los contornos llenados usando mapa
# de color predefinido
# Usar como: ContourLonLat(colormap,lon,lat,z,
#                          contourlevels,alpha,
#                          extrapoints,titletext,keytext,
#                          xlabel,ylabel)
# donde el plot esta hecho usando los argumentos como
# sigue
# colormap:      Cadena con el colormap ("Jet","Blue",
#                          "Red",etc.)
# lon, lat:      Vectores de puntos de cuadrícula en

```

```

#                               direcciones de latitud y longitud
# z:                             Arreglo con los valores numericos a
#                               graficar
# contourlevels: Numero de contornos a considerar.
#                               Opcional, el valor por default es 8
# alpha:                         Parametro de transparencia en (0,1].
#                               Opcional, valor por default 0.8
# extrapoints:                   Puntos adicionales a poner en la grafica
#                               como puntos negros
# titletext:                     Cadena con el titulo de la grafica
# keytext:                       Cadena a mostrar sobre la barra de color
#                               e.g. unidades. Opcional
# xlabel:                        Etiqueta del eje x. Opcional, el valor
#                               por default es "Longitude"
# ylabel:                        Etiqueta del eje y. Opcional, el valor
#                               por default es "Latitude"
#####

ContourLonLat = function(colormap="Jet",lon,lat,z,
                        contourlevels=8,alpha=0.8,
                        linewidth=0,extrapoints,
                        titletext,keytext="",
                        xlabel ="Longitude",
                        ylabel ="Latitude"){

  minz = min(z)
  maxz = max(z)
  switch(casefold(colormap),
    "jet" = {pallet_transparency <-
      JetPalletTransparency(contourlevels,alpha)},
    "blue"= {pallet_transparency <-
      BluePalletTransparency(contourlevels,alpha)},
    "red" = {pallet_transparency <-
      RedPalletTransparency(contourlevels,alpha)},
    print( paste("Colormap is not known. Known",
      "maps are Jet, Blue, Red" ) )
    # Default statement
  )

  if (exists("pallet_transparency")){
    ctlevels = seq(minz,maxz,
                  by=(maxz-minz)/(contourlevels-1))
    z2 = z
   ctlns = contourLines(lon,lat,z2, levels=ctlevels)
    filled.contour(lon,lat,z,levels=ctlevels,
                  col = pallet_transparency,
                  plot.axes = {
                    for (j in 1:nrow(extrapoints)){
                      points(extrapoints[j,1],
                            extrapoints[j,2],
                            type="p",pch = 19)
                    }
                  })
  }
}

```

```

    }
    axis(1)
    axis(2)
    sapply(1:contourlevels,
           function(x) lines(ctlns[[x]][[2]],
                             ctlns[[x]][[3]],
                             lwd=linewidth))
  },
  xlab = xlabel,
  ylab = ylabel,
  main = titletext,
  key.title = title(main = keytext)
)
}
else{
  print("Plotting simple contour plot")
}
}

get_obs <- function(wdata, coord_stations, blocked = NULL){
  #Obtener las temperaturas y el ID de las estaciones
  #####
  #### TIPO DE DATO A OBTENER ####
  #####
  tmpdata = wdata$PRESSURE
  #####

  n_stations = length(coord_stations[,1])
  id = wdata$ID_LOCATIONS_W

  # Numero de lecturas durante el intervalo
  n_lectures = length(tmpdata)

  # La temperatura promedio de esa estacion durante el
  # intervalo
  temp_upd = rep(0, n_stations)

  # EL numero de lecturas por estacion
  count_station = rep(0, n_stations)

  # Organizar los datos anteriores
  for (i in 1:n_lectures) {
    if (as.character(tmpdata[i]) != "NULL") {
      # Convertir a promedio
      x = as.numeric(as.character(tmpdata[i]))
      temp_upd[id[i]] = temp_upd[id[i]]*
        count_station[id[i]]
      temp_upd[id[i]] = temp_upd[id[i]]+x
      count_station[id[i]] = count_station[id[i]] + 1
      temp_upd[id[i]] =

```

```

        temp_upd[id[i]]/count_station[id[i]]
        # Este proceso va actualizando el promedio de
        # cada una de las estaciones
    }
}
# Aqui ya tiene el promedio de las mediciones de cada
# estacion
for(i in 1:n_stations) {
  if(count_station[i] == 0) {
    temp_upd[i] = NA
    # Marcar con NA las estaciones de las que no se
    # obtuvo lectura
  }
}

coord_stations = cbind(coord_stations[,1],
                       coord_stations[,2],1:32)
# Borrar las lecturas de las estaciones que se
# bloquearon por default
temp_upd[blocked] = NA

# Tomar en cuenta solo las estaciones no bloqueadas
x = which(!is.na(temp_upd))
temp_upd = temp_upd[x]
coord_stations = coord_stations[x,]

lectures = data.frame(temp_upd,coord_stations[,1],
                      coord_stations[,2],
                      coord_stations[,3])

names(lectures) = c("obs","long","lat")

return(lectures)
}

#####
###   estimar_krig   ###
#####
#   x0.....es el punto donde se hace la estimacion
#   stations....son las coordenadas de las estaciones
#   obsrv.....son las observaciones
#   var_model...es el modelo admisible de variograma
#####
estimar_krig <- function(A, x0,stations,obsrv,var_model) {

  n = length(obsrv);

  # Inicializar b
  b = rep(0,n+1);
  b[n+1] = 1

```

```

for (i in 1:n) {
  h = spDistsN1(t(stations[i,]),x0,longlat=TRUE);
  b[i] = var_model(h);
}
lambda = solve(A, b)

# inicializar el estimador
estm = 0

#calcular la estimador
for (i in 1:n) {
  estm = estm + obsrv[i]*lambda[i]
}
return(estm)
}

#####
##      registro      ##
#####
# wdata.....contiene el registro bruto de los datos
# initialdate...son las coordenadas de las estaciones
# timeln.....el tiempo en el que se consideran las
#               mediciones
#####
#El proposito de este programa es hallar las mediciones
#ocurridas en un intervalo de tiempo.

#En wdata estan guardadas las mediciones y el tiempo en
# que fue tomada dicha medicion esta guardada en
# time_stamp
#Notese que wdata debe estar ordenado por fecha

#Para que funcione se debe llamar a la libreria
#library("lubridate")

registro <- function(wdata,initialdate,
                    timeln = time_length(minutes(15))) {
  time_stamp = wdata$TIMESTAMP_W

  #cuantas mediciones hay en total
  n_tmst = length(time_stamp)

  #Buscar las fechas que se requieren por busqueda binaria
  #stt es el indice de inicio
  stt = 1
  #end es el indice del final
  end = n_tmst

  #aqui ocurre lo importante (Parte 1)

```

```

while (stt < end) {
  med = floor((stt+end)/2)
  #timelength para saber si el intervalo es positivo
  # o negativo (antes/despues)
  if(time_length(interval(initialdate,
                           dmy_hm(time_stamp[
                               floor((stt+end)/2)])))<0)
  {
    stt = med + 1
  }else {end = med}
}

#aqui ocurre lo importante (Parte 2)
inicio = stt
end = n_tmst

while (stt < end) {
  med = floor((stt+end)/2)
  if(time_length(interval(initialdate,
                           dmy_hm(time_stamp[med]))) <=
      timeln){
    stt = med + 1
  } else {end = med}
}

# final dice el primer elemento despues de los 15
# minutos, por lo que nos interesa una antes
final = stt-1

# wdata ahora tiene el intervalo de 15 minutos que me
# interesa
wdata = wdata[inicio:final,]

return(wdata)
}

#####
###      exp_mod      ###
#####
#   variogram.....contiene el variograma muestral
#   r.....rango del variograma
#   showit.....mostrar la curva
#####
exp_mod <- function(variogram,r,showit = FALSE) {

  x = variogram$h
  y = variogram$v
  n = length(x)

  A = matrix(data = 1,nrow = n,ncol = 2)

```

```

for (i in 1:n) {
  A[i,2] = 1-exp(-1*x[i]/r)
}

coeff = nnls(A, y)$x
eq <- function(x){coeff[1]+ coeff[2]*(1-exp(-x/r))}

if (showit) {
  curve(eq, from=0, to = x[n],add = TRUE)
}
return(eq)
}

#####
###      gss_mod      ###
#####
#   variogram.....contiene el variograma muestral
#   r.....rango del variograma
#   showit.....mostrar la curva
#####
gss_mod <- function(variogram,r,showit = FALSE) {

  x = variogram$h
  y = variogram$v
  n = length(x)

  A = matrix(data = 1,nrow = n,ncol = 2)

  for (i in 1:n) {
    A[i,2] = 1-exp(-1*(x[i]/r)^2)
  }

  coeff = nnls(A, y)$x

  eq <- function(x){coeff[1]+coeff[2]*(1-exp(-1*(x/r)^2))}
  if (showit) {
    curve(eq, from=0, to = x[n],add = TRUE)
  }
  return(eq)
}

#####
###      sph_mod      ###
#####
#   variogram.....contiene el variograma muestral
#   r.....rango del variograma
#   showit.....mostrar la curva
#####
sph_mod <- function(variogram , r, showit = FALSE) {

```

```

x = variogram$h
y = variogram$v
n = length(x)

A = matrix(data = 1,nrow = n,ncol = 2)

for (i in 1:n) {
  if (x[i] <= r) {
    A[i,2] = 3*x[i]/(2*r)-1/2*(x[i]/r)^3
  }
}

#hallar la mejor aproximacion
coeff = nnls(A, y)$x

eq <- function(x){coeff[1]+
  coeff[2]*(3*x/(2*r)-1/2*(x/r)^3)}
#trazar la primera parte de la curva
if (showit) {
  curve(eq, from=0, to = r,add = TRUE)
}
#trazar la segunda parte de la curva
if (showit && r < x[n]) {
  curve((0*x + coeff[1]+ coeff[2]),add = TRUE, from=r,
    to = x[n])
}

eq <- function(x) {
  if (x < r) {
    return(coeff[1]+ coeff[2]*(3*x/(2*r)-1/2*(x/r)^3))
  }
  else {
    return(coeff[1]+ coeff[2])
  }
}
return(eq)
}

#####
###      adm_model      ###
#####
#  variogram.....contiene el variograma muestral
#  r.....rango del variograma
#  mod_type.....tipo de modelo de variograma
#####
adm_model <- function(variogram, r, mod_type = "exp",
  showit = FALSE){
  modelf=function(x) {1}
  switch(casefold(mod_type),

```

```

        "exp" = modelf = exp_mod(variogram , r,showit),
        "sph" = modelf = sph_mod(variogram , r ,showit),
        "gss" = modelf = gss_mod(variogram , r,showit),
        print(paste("Modelo no conocido.",
                    "Utilice exp, sph, gss"))
    )
    return(modelf)
}

#####
###      emp_var_o      ###
#####
#   wdata.....contiene los registros en bruto
#   coord_stations..son las coordenadas de las estaciones
#   days.....dias en los que se toman semivarianzas
#   timeln.....minutos que seran incluidos
#   initialdate.....dias inicial del variograma
#   blocked.....estaciones con valores atipicos
#   outl.....indica si se van a excluir o no los
#                   valores fuera del rango
#   range.....factor con el que se decide si un
#                   valor queda fuera o no
#####
emp_var_o <- function(wdata,coord_stations,initialdate,
                    timeln = time_length(minutes(15)),
                    dayz = 1,blocked = NULL,
                    outl = FALSE,range = 1.5) {

    wdat = wdata
    lags = NULL
    semivar = NULL
    for ( d in 1:dayz ) {
        # Se obtienen los registros de ese dia
        wdata = registro(wdat,initialdate,timeln)
        wdata = get_obs(wdata,coord_stations,blocked)

        n_stat = length(wdata$obs)
        samples = wdata$obs

        #quitar outliers
        if(outl == TRUE) {
            qt = IQR(samples,na.rm = TRUE)
            qt1 = quantile(samples,probs = 0.25, names = FALSE,
                            na.rm = TRUE)
            qt3 = quantile(samples,probs = 0.75, names = FALSE,
                            na.rm = TRUE)
            for (f in 1:n_stat) {
                #revisa si es outlier
                if(samples[f]-qt3 > range*qt||
                    qt1-samples[f] > range*qt)
            {

```

```

        samples[f] = NA
    }
}
#quitar los NA
wdata = wdata[which(!is.na(samples)),]
samples = samples[which(!is.na(samples))]
n_stat = length(samples)
}

h = rep(0,n_stat*(n_stat-1)/2)
v = rep(0,n_stat*(n_stat-1)/2)
coord = cbind(wdata$long, wdata$lat)

#k es el contador
k = 0

#obtener lags y semivarianzas
for (i in 1:(n_stat-1)) {
  for (j in (i+1):n_stat) {
    k = k + 1
    h[k] = spDistsN1(t(coord_stations[i,]),
                     coord_stations[j,],longlat=TRUE)
    v[k] = ((samples[i]-samples[j])^2)/2;
  }
}

#agregar nuevos datos
lags = c(lags,h)
semivar = c(semivar,v)

initialdate = initialdate + days(1);
}
#variogram es el producto final
variogram = data.frame(lags,semivar)
#se ordena el variograma con respecto al lag
variogram = variogram[order(variogram[,1]),]
return(variogram)
}

#####
###      smp_var      ###
#####
# variogram.....contiene el variograma empirico
# intervals.....el numero de intervalos del variograma
# indx.....porcion de dist_max a aparecer
#####

smp_var <- function(variogram,n_int = 10,indx = 0.5) {
  # los lags y las semivarianzas
  lags = variogram$lags;

```

```

semivar = variogram$semivar;

#cuanto mide el variograma empirico
n_vgm = length(lags);

#distancia maxima absoluta
dist_max = lags[n_vgm];

#distancia maxima del variograma muestral
dist_max = dist_max*indx;

#Los intervalos distribuidos equidistantes
intervals = seq(0,dist_max,length = n_int+1)

#inicializar las nuevas componentes del variograma
h = rep(0,n_int)
v = rep(0,n_int)
n = rep(0,n_int)

#inicializar el contador
k=1
for(i in 2:(n_int+1)) {
  h[i-1] = (intervals[i-1] + intervals[i])/2

  # promediar las semivarianzas que caigan en el
  # intervalo

  v[i-1] = mean(semivar[which((lags > intervals[i-1])
                              &(lags <= intervals[i]))])
  n[i-1] = length(semivar[which((lags > intervals[i-1])
                              &(lags <= intervals[i]))])
}

sample_var = data.frame(h,v,n)
sample_var = sample_var[which(!is.na(sample_var[,2])),]
return(sample_var)
}

#####
###      findrange      ###
#####
#   variogram.....contiene el variograma muestral
#   mod_type.....tipo de modelo de variograma
#####
findrange <- function(variogram,mod_type,stt = 0.1,
                      fin = -1) {
  minm = 10000

  if (fin == -1) {
    fin = max(variogram$h,na.rm = TRUE)

```

```
}

n_smp = length(variogram$h)

for (i in 1:4) {
  vec = seq(stt,fin,length = 11)
  for(j in 1:11) {

    var_model = adm_model(variogram,vec[j],mod_type,
                          showit = FALSE)

    sq = 0

    for(i in 1:n_smp) {
      sq = sq+(variogram$v[i]-
               var_model(variogram$h[i]))^2
    }

    if (minm > sq) {
      k = j
      minm = sq
    }
  }

  minm = 10000

  if(k == 1) {
    stt = vec[1]
  } else {
    stt = vec[k-1]
  }
  if(k == 11) {
    fin = vec[11]
  } else {
    fin = vec[k+1]
  }
}

return((stt+fin)/2)
}
```

## Apéndice B

# Código de R para el análisis preliminar de los datos

```
#Import the location of the weather stations
weather = read.csv("locations_weather.csv")

#stations coordinates
coord_stations = cbind(weather$LONG,weather$LAT)

#Fecha a partir de la cual se toman las mediciones
wdat = read.csv("weather_May01_to_Aug14_2017.csv")
initialdate = dmy_hm("01/05/2017_00:00")
#initialdate = dmy_hm("08/05/2017 18:00")

timeln = time_length(minutes(15))

##### RANGO PARA VALORES ATIPICOS #####
rang = 1.5
#####

dayz = 5

datos = matrix(data = NA,nrow = 32,ncol = dayz)
colnames(datos) = c("1","2","3","4","5");

for (d in 1:dayz) {
  #observaciones
  wdata = registro(wdat,initialdate,timeln)
  x = get_obs(wdata,coord_stations,blocked = c(7))
  n = length(x$obs);
  datos[1:n,d] = x$obs;
  initialdate = initialdate + days(1);
}

boxplot.matrix(datos,ylab = "Fecha_(Mayo_2017)",
```

```

range = rang,xlab = "Humedad[%"",
main = paste("Box-plot",
            "del índice de humedad",
            "(Noche)",horizontal = TRUE);

datos2 = datos

for (g in 1:5) {
  qt = IQR(datos2[,g],na.rm = TRUE)
  qt1 = quantile(datos2[,g], probs = .25, na.rm = TRUE,
                names = FALSE)
  qt3 = quantile(datos2[,g], probs = .75, na.rm = TRUE,
                names = FALSE)
  for (f in 1:32) {
    #revisa si es outlier
    if (!is.na(datos2[f,g])
        && (datos2[f,g] - qt3 > rang*qt
            || qt1 - datos2[f,g] > rang*qt)) {
      datos2[f,g] = NA
    }
  }
}

print(c(sd(datos2[,1],na.rm = TRUE),
        sd(datos2[,2],na.rm = TRUE),
        sd(datos2[,3],na.rm = TRUE),
        sd(datos2[,4],na.rm = TRUE),
        sd(datos2[,5],na.rm = TRUE)))

print(c(var(datos2[,1],na.rm= TRUE),
        var(datos2[,2],na.rm = TRUE),
        var(datos2[,3],na.rm= TRUE),
        var(datos2[,4],na.rm = TRUE),
        var(datos2[,5],na.rm = TRUE)))

print(c(skewness(datos2[,1],na.rm = TRUE),
        skewness(datos2[,2],na.rm = TRUE),
        skewness(datos2[,3],na.rm = TRUE),
        skewness(datos2[,4],na.rm = TRUE),
        skewness(datos2[,5],na.rm = TRUE)))

```

## Apéndice C

# Código de R para el proceso de kriging

```
### Ubicaciones de las estaciones #####
weather = read.csv("locations_weather(2).csv")
#####

### Numero de estaciones y sus ubicaciones #####
n_stations = length(weather$LONG)
coord_stations = cbind(weather$LONG,weather$LAT)

#Fecha a partir de la cual se toman las mediciones
wdata = read.csv("weather_May01_to_Aug14_2017.csv")

### fecha a partir de la que se toman muestras ###
initialdate = dmy_hm("01/05/2017_17:00")

### intervalo de tiempo para tomar muestras #####
timeln = time_length(minutes(15))

### Rango para determinar valores atipicos ###
rang = 2

wdat = registro(wdata,initialdate,timeln)
lectures = get_obs(wdat, coord_stations, blocked = NULL)

### Generar grafica con las mediciones #####
plot(lectures$long,lectures$lat,pch=18,
      main = paste("Estaciones_Activas[hPa]_01/05/2017",
                  "17:00_-17:15"),
      ylab="Latitud" ,ylim = c(min(lectures$lat)-0.008,
                              max(lectures$lat)+0.008),
      xlab="Longitud",
      xlim = c(min(lectures$long)-0.008,
              max(lectures$long)+0.008))
```

```

textxy(lectures$long,lectures$lat,
       signif(lectures$obs,digits = 4),offset = 1)
#####

### Importar de nuevo las ubicaciones ####
weather = read.csv("locations_weather.csv")

n_stations = length(weather$LONG)

### variograma empirico #####
variogram = emp_var_o(wdata,coord_stations,
                     initialdate,timeln,dayz = 5,
                     blocked = c(7),outl = TRUE,
                     range = rang)
#####

plot(variogram$lags, variogram$semivar,pch=18,
     main="Variograma Empirico 01/05/2017 17:00-17:15",
     ylab=expression('Semivarianza',hat(gamma) ('u')),
     xlab="Distancia(km)",xaxt = 'n')
axis(1, at=seq(0,30,2), tick = TRUE)

#### sample variogram #####
sampl = smp_var(variogram,n_int = 8 ,indx = 0.4) #
#####

sq = c(0,0,0)
r = c(0,0,0)

plot(c(0,sampl$h), c(-10,sampl$v),pch=18, col="red",
     main="Variograma Muestral 01/05/2017 17:00-17:15",
     ylab=expression('Semivarianza',hat(gamma) ('u')),
     xlab="Distancia(km)",
     ylim = c(0,max(sampl$v,na.rm = TRUE)+2.7),
     xaxt = 'n')
axis(1, at=seq(0,10,1), tick = TRUE)

polyg = read.csv("df.csv");
totalp = length(polyg$x)

# mycolorbar = scale_fill_distiller(name = "[hPa]",
#   values = rev(brewer.pal(11,"RdYlBu")))

mycolorbar = scale_fill_distiller(name = "[hPa]",
                                  palette = "RdYlBu",
                                  direction = -1)

#####
Map = get_map(location = "Monterrey, Mexico",

```

```

        color = "color",
        source = "google",
        maptype = "terrain",
        zoom = 11)
save(Map, file="./MtyMap.Rdata")
#####

# Si ya se cargo el mapa en el archivo MtyMap.Rdata
# se pueden omitir las instrucciones anteriores y
# cambiarse por:
# load("./MtyMap.Rdata")

### Crear el mapa para el plot ###
Mtymap = ggmap(Map,
               extent = "panel",
               ylab = "Latitud",
               xlab = "Longitud",
               legend = "right"
)

mytheme = theme(axis.text=element_text(size=16),
                axis.title.x=element_text(size=16),
                axis.title.y=element_text(size=16),
                title=element_text(size=16),
                legend.text=element_text(size=14))

for (i in 1:3) {
  switch(as.character(i),
        "1" = mod_type = "exp",
        "2" = mod_type = "sph",
        "3" = mod_type = "gss"
  )
  switch(as.character(i),
        "1" = lbl = paste("Ajuste exponencial",
                          "01/05/2017 17:00-17:15"),
        "2" = lbl = paste("Ajuste esferico 01/05/2017",
                          "17:00-17:15"),
        "3" = lbl = paste("Ajuste gaussiano 01/05/2017",
                          "17:00-17:15")
  )
  plot(c(0, sampl$h), c(-10, sampl$v), pch=18, col="red",
        main=lbl, ylab=expression('Semivarianza',
                                  hat(gamma) ('u')),
        xlab="Distancia (km)",
        ylim = c(0,
                 max(sampl$v,
                      na.rm=TRUE)+
                 2.7),
        xaxt = 'n')
  axis(1, at=0:15, tick = TRUE)
}

```

```

r[i] = findrange(sampl,mod_type)

var_model = adm_model(sampl,r[i],mod_type,showit=TRUE)

n_smp = length(sampl$h)

sq[i] = 0

for(j in 1:n_smp) {
  sq[i] = sq[i]+(sampl$v[j]-var_model(sampl$h[j]))^2
}

wdat = registro(wdata,initialdate,timeln)

lectures = get_obs(wdat, coord_stations, blocked = c())
stations = cbind(lectures$long,lectures$lat)
stations2 = data.frame(stations)
names(stations2) = c("lon","lat")

# Parametros de la cuadrícula para el area de latitud y
# longitud
nlon=80
nlat=60

temp = matrix(0,nlon,nlat)

lon_limits = c(min(stations[,1]-0.05),
               max(stations[,1]+0.05))
lat_limits =c (min(stations[,2]-0.05),
               max(stations[,2]+0.05))

lon_v = seq(lon_limits[1],lon_limits[2],length = nlon)
lat_v = seq(lat_limits[1],lat_limits[2],length = nlat)

obsrv = lectures$obs
n = length(obsrv);

# Inicializa la matriz de las semivarianzas A
A = matrix(1,n+1,n+1);
A[n+1,n+1] = 0;

# Escribir las semivarianzas en la matriz A
for (i1 in 1:n) {
  for (j1 in 1:n) {
    h = spDistsN1(t(stations[i1,]),
                  as.matrix(stations[j1,]),
                  longlat=TRUE);
    A[i1,j1] = var_model(h);
  }
}

```

```

    A[i1,i1] = 0;
  }
  estim = 10

  for(i1 in 1:totalp){
    x0 = c(polyg[i1,1],polyg[i1,2])
    polyg[i1,3] = estimar_krig(A,x0,stations,
                              lectures$obs,var_model)
  }

  contourlevels = 20
  contourlinewidth = 0
  alpha = .9

  switch(as.character(i),
    "1" = Mtymap +
      ggtitle("Kriging_Ordinario_Modelo_Exponencial")
      + geom_tile(data=polyg, alpha=.8,
                  aes(x, y, fill=predict)) +
      geom_point(data = stations2, color = "black") +
      geom_contour(data = polyg, aes(x,y, z=predict),
                  binwidth=1,color="white") +
      xlab("Longitud_[]_[]") +
      ylab("Latitud_[]_[]") +
      mycolorbar +
      mytheme,
    "2" = Mtymap +
      ggtitle("Kriging_Ordinario_Modelo_Esferico")+
      geom_tile(data=polyg, alpha=.8, aes(x, y,
                                          fill=predict)) +
      geom_point(data = stations2, color = "black") +
      geom_contour(data = polyg, aes(x,y, z=predict),
                  binwidth=1,color="white") +
      xlab("Longitud_[]_[]") +
      ylab("Latitud_[]_[]") +
      mycolorbar +
      mytheme,
    "3" = Mtymap +
      ggtitle("Kriging_Ordinario_Modelo_Gaussiano")+
      geom_tile(data=polyg, alpha=.8, aes(x, y,
                                          fill=predict))+
      geom_point(data = stations2, color = "black") +
      geom_contour(data = polyg, aes(x,y, z=predict),
                  binwidth=1,color="white") +
      xlab("Longitud_[]_[]") +
      ylab("Latitud_[]_[]") +
      mycolorbar +
      mytheme
  )
}

```



# Bibliografía

- Armstrong, M. (1998). *Basic Linear Geostatistics*. Springer.
- Díaz-Viera, M. A., Herrera-Zamarrón, G. S., y Valdés-Manzanilla, A. (2009). Un modelo de correogionalización lineal para la estimación espacial de la precipitación en el valle de la Ciudad de México, combinando datos de pluviógrafos con imágenes de radar meteorológico. *Ingeniería Hidráulica en México*, 24, 63-90. Descargado de <http://www.revistatyca.org.mx/ojs/index.php/tyca/article/view/99/95>
- Diggle, P. J., y Ribeiro, P. J., Jr. (2007). *Model-based Geostatistics*. Springer.
- Graffelman, J. (2013). calibrate: Calibration of scatterplot and biplot axes [Manual de software informático]. Descargado de <https://CRAN.R-project.org/package=calibrate> (Paquete de R versión 1.7.2)
- Instituto Nacional de Estadística y Geografía. (2014). *Carta Topográfica G14C25 (San Pedro Garza García). Escala 1:50000*. Descargado de <https://www.beta.inegi.org.mx/app/biblioteca/ficha.html?upc=702825208011>
- Kahle, D., y Wickham, H. (2016). ggmap: Spatial visualization with ggplot2 [Manual de software informático]. Descargado de <https://CRAN.R-project.org/package=ggmap> (Paquete de R versión 2.6.1)
- Montalvo-Urquizo, J., Marroquín, M. G. V., Hernández-Castillo, J. J., y Hernández-González, H. E. (2017). MWTP: monterrey weather, traffic and pollution database for geospatial analysis. *CoRR*, abs/1703.04526. Descargado de <http://arxiv.org/abs/1703.04526>
- Mullen, K. M., y van Stokkum, I. H. M. (2012). npls: The lawson-hanson algorithm for non-negative least squares (npls) [Manual de software informático]. Descargado de <https://CRAN.R-project.org/package=npls> (Paquete de R versión 1.4)
- Myers, D. E. (1989). To Be or Not to Be... Stationary? That Is the Question. *Mathematical Geology*, 21(3), 347-362. Descargado de <https://link.springer.com/article/10.1007/BF00893695>

- National Imagery and Mapping Agency. (2000). *Department of Defense World Geodetic System 1984, Its Definition and Relationships with Local Geodetic Systems* (Inf. Téc. n.º TR8350.2). Autor.
- Oliver, M. A., y Webster, R. (2015). *Basic Steps in Geostatistics: The Variogram and Kriging*. Springer.
- Pebesma, E., y Bivand, R. (2017). sp: Classes and methods for spatial data [Manual de software informático]. Descargado de <https://CRAN.R-project.org/package=sp> (Paquete de R versión 1.2-5)
- Rathbun, S. (1998). Spatial Modelling in Irregularly Shaped Regions: Kriging Stuaries. *Environmetrics*, 9, 109-129.
- Soto Rodríguez, A. N. (2015). *Solución numérica de la ecuación de advección-difusión para las islas de calor urbano empleando diferencias finitas generalizadas* (Tesis de Licenciatura). Facultad de Ciencias Físico-Matemáticas, Universidad Michoacana de San Nicolás Hidalgo.
- Spinu, V., Grolemond, G., y Wickham, H. (2017). lubridate: Make dealing with dates a little easier [Manual de software informático]. Descargado de <https://CRAN.R-project.org/package=lubridate> (Paquete de R versión 1.7.1)
- Wickham, H., y Chang, W. (2016). ggplot2: Create elegant data visualisations using the grammar of graphics [Manual de software informático]. Descargado de <https://CRAN.R-project.org/package=ggplot2> (Paquete de R versión 2.2.1)
- Zavaleta Sánchez, J. (2014). *Un sistema para la reconstrucción de superficies mediante técnicas de interpolación de datos dispersos* (Tesis de Máster). Facultad de Ciencias, Universidad Nacional Autónoma de México.