Universidad Nacional Autónoma de México

Facultad de Ciencias

Análisis Exploratorio de Datos en la estimación Geoestadística

T E S I S

QUE PARA OBTENER EL TÍTULO DE:
ACTUARIA

P R E S E N T A
AURA ARCHUNDIA AVILA

DIRECTOR DE TESIS
DR. MARTÍN ALBERTO DÍAZ VIERA

2011

Hoja de Datos del Jurado

1.Datos del alumno:
Archundia Avila Aura
55 44 69 93
Universidad Nacional Autónoma de México
Facultad de Ciencias
Actuaría
303049797

2.Datos del tutor:
Dr. Martín Alberto Díaz Viera

3.Datos del sinodal 1:
Dr. Ricardo Casar González

4.Datos del sinodal 2:
Act. Jaime Vázquez Alamilla

5. Datos del sinodal 3:
Dra. María del Pilar Alonso Reyes

6. Datos del sinodal 4:
Mat. Margarita Elvira Chávez Cano

7. Datos del trabajo escrito:
Análisis Exploratorio de Datos en la estimación Geoestadística
260p
2011

Dedicatorias y agradecimientos

Esta tesis se la dedico en especial a la memoria de
Mi abuelita Elenita Osorio y mi abuelito Ricardo Ávila
por todo el amor que siempre me dieron y por ser la luz de mi vida
en todo momento. Siempre estarán en mi corazón.

A mi madre,
Por todo el amor, comprensión y el apoyo incondicional que siempre me has brindado.

A mis hermanos Jorge y Maggy,
Por su comprensión, apoyo y el gran cariño que me han dado a lo largo de mi vida.

A mi padre,
Por tu ayuda, comprensión y todo el cariño que me has dado.

A Jorge, Bere y Yamil,
Por su cariño y apoyo.

A mis amiguis,
Sabrina, Jessica, Aaron, Army, Ruben, Checo, Laura, Marisol, Diego y Mau, por la fortuna de haberlos conocido y haberme brindado su amistad, por aguantarme, aceptarme como soy y darme su cariño incondicional.

A todas aquellas personas que han estado ahí para apoyarme e impulsarme.

A mis maestros, por sus enseñanzas, por transmitirme sus experiencias, por inspirarme, darme aliento y prepararme para mi vida profesional.

A mi Facultad de Ciencias y la UNAM, por el privilegio de ser parte de ellas.

Resumen

Debido a que en muchas ocasiones las herramientas geoestadísticas son utilizadas suponiendo que el geoestadístico tiene conocimientos profundos de estadística o en particular del análisis exploratorio de datos, se producen errores en la aplicación e interpretación de la información. Lo cual conlleva al objetivo principal de este trabajo el cual es revisar el análisis exploratorio de datos con el contexto de analizar la metodología y concentrar los conocimientos. De esta manera se tendrán recomendaciones metodológicas para generar una metodología sistemática que reduzca al mínimo los errores de juicio o de conocimiento. Además, se tienen algunos objetivos particulares como resaltar la dificultad que existe al aplicar la metodología a una base de datos de tamaño insuficiente y que cuenta con alguna característica que no se conoce, lo que resulta en un proceso complejo y poco confiable, mientras que a medida que aumenta el número de observaciones el proceso se vuelve más simple y confiable. Otro objetivo particular es mostrar que a partir de diferentes tipos de muestreo las características implícitas en las bases de datos pueden resultar visibles o no e incluso perturbadas notoriamente debido a la significancia de las muestras elegidas. Esto da lugar a uno más de los objetivos particulares, el cual es destacar la influencia de las características presentadas en las bases de datos. Por último, el presente permite mostrar gráficamente algunos procesos de estimación para mencionar recomendaciones en la manera de identificar las características que pudiera presentar una base de datos, así como un proceso a seguir para tratarlas.

Se utilizan 3 bases de datos (D1, D2 y D3) de las cuales se obtienen 27 escenarios, en los que se tiene asimetría positiva y ninguno cumple normalidad. Los D1 tienen tendencia, sin embargo, éstos representan una base de datos sin tendencia y sin anisotropía debido a que la tendencia no es significativa y mucho menos visible. Los D2 provienen de una estructura anidada, pero la característica observada en los datos fue de tendencia por lo que uno de los resultados relevantes es la demostración del proceso de eliminación de tendencia y la afectación que conlleva en la estimación. Los D3 presentan anisotropía, por lo que se ejemplificó la identificación y el proceso de estimación cuando existe anisotropía y destacó durante el proceso de estimación la complejidad que resulta de identificar las direcciones de los ejes de anisotropía de máximo y mínimo alcance, así como la influencia del tipo de muestreo y el número de observaciones que contenían.

En los resultados, una parte importante dentro del análisis exploratorio de datos es que destacan las transformaciones que se realizan por la asimetría visible y los datos atípicos observados. Para los tres casos de estudio, resultó evidente que si se cuenta con una gran cantidad de información, el tipo de muestreo resulta menos importante y el ajuste del modelo es más fácil, evidente y confiable.

Durante el proceso de aplicación de la metodología de estimación con kriging, la confiabilidad y eficiencia de la estimación geoestadística está basada en el adecuado proceso y análisis de la información, en particular el análisis exploratorio de datos, así como una cantidad suficiente de la misma información, un acertado ajuste del modelo de variograma a utilizar y una adecuada valoración del mismo para llegar a una estimación exitosa.

Índice general

Índice de cuadros

Índice de figuras

Capítulo 1

Introducción

En los años 60, Matheron acuñó el término de Geoestadística. Él formalizó y generalizó matemáticamente un conjunto de técnicas desarrolladas por D.G. Krige (1941) que explotaban la correlación espacial para hacer predicciones en la evaluación de reservas de las minas de oro en Sudáfrica. Matheron es reconocido como el padre de la Geoestadística. Él definió a la Geoestadística como *"la aplicación del formalismo de las funciones aleatorias al reconocimiento y estimación de fenómenos naturales"* (Matheron, 1962).

La geoestadística[1] es una rama de la estadística espacial que se especializa en el análisis y la modelación de la variabilidad espacial en ciencias de la tierra. Su objeto de estudio es el análisis y la predicción de fenómenos en espacio y/o tiempo, tales como: ley de metales, porosidades, concentraciones de un contaminante, pluviosidad, etc. Aunque el prefijo *geo* es usualmente asociado con geología, la geoestadística tiene sus orígenes en la minería. Además, el prefijo viene de raíces griegas que significan Tierra.

Conforme ha pasado el tiempo, la geoestadística se ha formado como una herramienta útil para el análisis y estimación con datos georreferenciados de variables que tengan continuidad espacial.

Dentro de cualquier estudio estadístico, se inicia con el análisis exploratorio de datos para conocer la información con la que se cuenta, y la geoestadística no es la excepción. Sin embargo, en algunas ocasiones las herramientas geoestadísticas son utilizadas suponiendo que el geoestadístico tiene conocimientos profundos de estadística o en particular del análisis exploratorio de datos, lo que ocasiona errores en la aplicación e interpretación de la información.

Uno de los artículos que desataca esta situación es *"Variance of geoestatisticians"*[2]. El artículo hace referencia a que dependiendo del individuo que realiza el estudio, se utilizarán diferentes enfoques al análisis e interpretación de los datos. El artículo cuantifica el efecto de las diferencias de dichos individuos en la calidad de los estimados espaciales geoestadísticos. Éste resultó en que las diferencias en las interpolaciones pueden ser provocadas por decisiones en la

[1] Referencia bibliográfica [2]

[2] *Variance of geoestatisticians, Evan J. Englund 1990.* Ref. bibliográfica [22]

metodología, diferencias en la interpretación de los datos, y en algunos casos a errores en el procedimiento. El proceso de interpolación introduce subjetivamente la decisión por parte del investigador, ya sea en la elección del método de interpolación, la detección de datos atípicos, transformaciones sobre los datos o la interpretación de la estructura de correlación espacial.

Incluso personas con conocimientos previos pueden errar en la interpretación de algún estudio como lo es el artículo *"Evaluation and Comparison of Spatial interpolators"*[3], en el cual se evalúan quince diferentes estimadores para determinar méritos relativos en la estimación por bloques de contaminantes en sitios. Donde, para medir la calidad de la estimación se utilizaron una función lineal de pérdida y una estadística estandarizada la cual es la media cuadrada del error. Los resultados sorprendentes del estudio fueron que la distancia inversa y el cuadrado de la distancia inversa produjeron mejores resultados que el kriging. Sin embargo, en el estudio no se concluye que necesariamente los métodos de la distancia inversa sean superiores a los estimadores con kriging.

Posteriormente, los mismos autores realizaron un segundo artículo *"Evaluation and Comparison of Spatial interpolators II"*[4], en el cual declaran que en estudios previos se han investigado los efectos de varios parámetros de estimación en la calidad de los estimados espaciales, como Englund (1990), Webber y Englund (1992). En todos los casos, variaciones razonables del experimento pudieron ser imaginadas cuando el kriging se esperaba que tuviera una ventaja distinta sobre el particular algoritmo de la distancia inversa utilizado. La naturaleza de la base de datos pudo favorecer fortuitamente a la distancia inversa. Tanto el método de kriging como el modelado de variograma se realizó de manera muy simple, por lo que cambiar cualquiera de los dos pudo tener un resultado relativamente diferente. Además, la anisotropía marcada y el agrupamiento de las muestras, que es lo que favorece el kriging, no se presentó en los datos. Dentro de los resultados, arrojan que los estimadores de la distancia inversa son sensibles al tipo de base de datos, al número de muestras en la vecindad usadas para estimar y al poder de la distancia usada en los pesos. Y en contraste, el kriging ordinario usando modelos de variograma ajustados es relativamente robusto al tipo de base de datos y al método de estimación del variograma experimental. Los estimados del kriging mejoraron consistentemente cuando se incrementa el número de muestras en la vecindad, sin importar el tipo de base de datos.

Con esta información se encara el problema de que realmente no hay estudios sistemáticos en el contexto del análisis geoestadístico. Esto conlleva al objetivo principal de este trabajo el cual es revisar el análisis exploratorio de datos con el contexto de analizar la metodología y concentrar los conocimientos. De esta manera se tendrán recomendaciones metodológicas lo cual genera una metodología sistemática que reduzca al mínimo los errores de juicio o de conocimiento.

Además, se tienen algunos objetivos particulares como destacar que cuando se tiene poca información, el proceso de estimación es complejo y poco confiable,

[3] *Evaluation and Comparison of Spatial interpolators I, Dennis Weber y Evan Englund, 1992.* Ref. Bibliográfica [23]

[4] *Evaluation and Comparison of Spatial interpolators II, Dennis Weber y Evan Englund, 1993.* Ref. Bibliográfica [24]

mientras que a medida que aumenta el número de observaciones el proceso se vuelve más simple y confiable. Otro objetivo particular es mostrar que a partir de diferentes tipos de muestreo las características implícitas en las bases de datos pueden resultar visibles o no e incluso afectadas notoriamente debido a la significancia de las muestras elegidas. Esto da lugar a uno más de los objetivos particulares, el cual es destacar la influencia de las características presentadas en las bases de datos, ya que cuando se observa adecuadamente la característica implícita de la base de datos en cuestión, el proceso de estimación puede resultar más sencillo y acertado.

La estructura de la tesis se inicia con la teoría de la aplicación de la metodología de estimación con kriging en geoestadística. Después, se da la descripción de la información utilizada, es decir, las bases de datos iniciales y sus características esenciales, los subconjuntos de datos obtenidos de esas 3 bases de datos iniciales y cómo fue realizada la extracción de la información. Luego, se dedica un capítulo a toda la información de los datos del tipo 1. Desde mostrar el proceso de estimación para los datos originales del tipo 1, mostrar un resumen de la información de los modelos de variogramas, así como mencionar las afectaciones más relevantes durante el proceso de estimación en cada uno de los escenarios y luego se muestra la aplicación de todo el proceso de estimación de 3 casos representativos (siempre con 100 observaciones) hasta llegar a su respectivo mapa de estimación y por último se destacan los resultados. Esta estructura de capítulo es aplicada análogamente en los dos siguientes capítulos para los datos del tipo 2 y los datos del tipo 3. Por último, se realiza la discusión sobre procedimientos, visualización y afectación de las características en cada base de datos y se destacan los resultados más relevantes hasta llegar a los resultados generales.

En los resultados, una parte importante dentro del análisis exploratorio de datos es que destacan las transformaciones que se realizan debido a la asimetría visible y los datos atípicos observados.

La base original de datos del tipo 1 tiene tendencia, sin embargo, esta representa una base de datos sin tendencia y sin anisotropía debido a que la tendencia no es significativa y mucho menos visible. Dentro de los resultados destaca que sólo un escenario resultó ajustado como modelo exponencial y fue el escenario de Datos 1 con muestreo aleatorio y 400 observaciones, esto puede ser ocasionado por el muestreo, ya que no es regular y probablemente asignó mayor peso a muestras que no eran tan significativas, lo cual derivó en un modelo diferente al utilizado en los demás escenarios de este tipo de datos los cuales fueron modelados con variogramas esféricos.

La base de datos original del tipo 2 presenta una estructura anidada, sin embargo, para este caso de estudio existió una diferencia de gran trascendencia dentro de los escenarios y dentro de los procesos utilizados para la estimación, debido a que la característica observada en los datos fue de tendencia y no de modelo anidado, posiblemente debido a que el modelo exponencial tenía mayor peso, lo que derivó en variogramas que mostraban una tendencia lineal, así como gráficos utilizados para observar tendencia en los cuales se mostraba de manera contundente. Por lo tanto, uno de los resultados relevantes es la demostración

del proceso de eliminación de tendencia y la afectación que conlleva en la estimación. También resalta que las bases de datos con 36 observaciones son complejas para permitir calcular un variograma adecuadamente y por lo mismo también es difícil ajustar el modelo de variograma, por lo cual la combinación de una estructura anidada que se percibe como tendencia, más un muestreo aleatorio con una cantidad completamente insuficiente de información para definir todas las características resulta en un proceso laborioso y difícilmente confiable en los resultados, aunque sorprendentemente acertado para esta situación en particular.

La base de datos original del tipo 3 presenta anisotropía, por lo que se ejemplificó la identificación y el proceso de estimación cuando existe anisotropía y destacó durante el proceso de estimación la complejidad que resulta de identificar las direcciones de los ejes de anisotropía de máximo y mínimo alcance. La anisotropía es una característica de gran influencia y es posible realizar una estimación de manera más acertada siempre y cuando sea detectada y tratada de manera adecuada. Debido a que la característica de anisotropía depende de la dirección, los mapas de estimación fueron significativamente influenciados por el tipo de muestreo y el número de observaciones que contenían. En dos escenarios, que son el de Datos 3 muestreo aleatorio con 36 observaciones y el de Datos 3 muestreo combinado con 36 observaciones sería adecuado pensar que el número tan reducido de información con la que se cuenta y claramente siendo poco significativa ocasione que la dirección de máximo y mínimo alcance se vea altamente desviada. Pero por otro lado, un resultado sobresale dentro de todos los escenarios de Datos 3 y es el del muestreo combinado con 100 observaciones. Aunque en este escenario se cuenta con una cantidad de observaciones suficiente y un muestreo que se consideraría adecuado, la dirección de anisotropía es casi contraria a la de los datos originales del tipo 3 y la razón posible es que a pesar de que la muestra tiene suficientes observaciones, éstas no son suficientemente significativas y por lo tanto no permite identificar adecuadamente la desviación, con lo cual se afecta considerablemente a la estimación.

Para los tres casos de estudio, resultó evidente que cuando se cuenta con una mínima cantidad de información la estimación no es muy confiable aún sin tener características influyentes sobre el modelo y hacia la estimación. Si se cuenta con suficiente información significativa es posible, más no forzoso, que se puede llegar a una estimación aceptable si se adapta un modelo adecuado. Pero si se cuenta con una gran cantidad de información, el tipo de muestreo resulta menos importante y el ajuste del modelo es más fácil, evidente y confiable.

Los 3 casos de estudio presentan por separado 3 características influyentes dentro del proceso de estimación, sin embargo, en la realidad es posible que se presente más de una característica en la misma base de datos u otras características que no se han mencionado, por lo que se recomienda utilizar las herramientas, indicadores, procesos, gráficos y toda la información presentada en este trabajo como apoyo visual y analítico para identificar y tratar las diferentes afectaciones que puedan presentarse al realizar algún otro estudio.

Con toda esta información destaca que durante el proceso de aplicación de la metodología de estimación con kriging, la confiabilidad y eficiencia de la es-

timación geoestadística está basada en el adecuado proceso de análisis de la información, en particular el análisis exploratorio de datos, así como una cantidad suficiente información, un acertado ajuste del modelo de variograma a utilizar y una adecuada valoración del mismo para llegar a una estimación exitosa.

Capítulo 2

Metodología de Geoestadística

Dado que el estudio de la metodología no es el objetivo principal de este trabajo, únicamente se describe en lo general los procesos metodológicos de la geoestadística[1].

2.1. Conceptos básicos

2.1.1. Variable aleatoria

Definición: Sea (Ω, F, P) un espacio de probabilidad y $Z : \Omega \to R$. Se dice que z es una variable aleatoria si: $\{ x \in \Omega : Z(x) \leq z\} \in F, \forall z \in R$.

Se le llama *variable regionalizada* a la variable distribuida en el espacio de tal manera que presenta una estructura espacial de correlación, es decir, una variable regionalizada es una variable aleatoria z definida en un punto del espacio $\underline{x}$. Donde en el caso más general $\underline{x}$ es un punto en el espacio tridimensional, es decir, $\underline{x} = (x_1, x_2, x_3)$.

2.1.2. Función Aleatoria

En geoestadística, una función aleatoria es un conjunto de variables aleatorias que tienen una cierta ubicación espacial y cuya dependencia la una de la otra es especificada por un mecanismo probabilístico.

Matemáticamente, si a cada punto $\underline{x}$ que pertenece a un dominio en el espacio se le hace corresponder una variable aleatoria $z(x)$, que en sentido general pueden ser dependientes, entonces el conjunto de variables aleatorias espacialmente distribuidas $\{z(x), x \in \Omega\}$ será una función aleatoria $Z(x)$.

[1]Para más información sobre la metodología se recomienda revisar la bibliografía y en particular las referencias bibliográficas [1] [2] [12]

Una realización de una función aleatoria $Z(x)$ es una variable regionalizada Z'.

2.1.3. Función de distribución de probabilidad

La *función de distribución de la variable aleatoria* Z es $F_Z : R \to [0,1]$ y está dada por:

$$F_Z(z) = Pr\left[Z \leq z\right] = Pr\left[x \in \Omega | Z(x) \leq z\right] \tag{2.1}$$

Si la distribucion de probabilidad es conocida, se pueden calcular muchos parámetros que describen características interesantes de la variable aleatoria. Hay un máximo y un mínimo para toda variable aleatoria, así como media, varianza y desviación estándar. Si el conjunto de resultados es suficientemente grande tendrá mediana y cuartiles.

Cabe mencionar que la distribución de la variable aleatoria no puede ser determinada por conocer sólo algunos de los parámetros, así como también los parámetros no pueden ser calculados sólo observando los resultados de la variable aleatoria.

Ahora se define la *función de distribución de una función aleatoria*:

Sea $Z(x)$ una función aleatoria definida en R^3 y la función $F_{Z(x_1),Z(x_2),Z(x_3),\ldots,Z(x_n)} : R^n \to R$, entonces el vector aleatorio $\{Z(x_1), Z(x_2), Z(x_3), \ldots, Z(x_n)\}$ se caracteriza por su función de distribución de probabilidad n-variada:

$$\begin{aligned} F_{Z(x_1),Z(x_2),Z(x_3),\ldots,Z(x_n)}\left(z_1, z_2, z_3, \ldots, z_n\right) = \\ Pr\left[Z(\underline{x_1}) \leq z_1, Z(\underline{x_2}) \leq z_2, Z(\underline{x_3}) \leq z_3, \ldots, Z(\underline{x_n}) \leq z_n\right] \end{aligned} \tag{2.2}$$

El conjunto de todas las distribuciones para todo valor de n y para cualquier selección de puntos en R^3 constituye el espacio de probabilidad de la función aleatoria $Z(x)$.

2.1.4. Modelo de Función aleatoria

Dentro de la rama de ciencias de la Tierra, muy pocos casos se conocen y entienden suficientemente para permitir la aplicación de modelos determinísticos. Las variables de interés en las bases de datos de ciencias de la Tierra, por lo general, son el resultado de un amplio número de procesos cuyas interacciones son muy complejas, las cuales todavía no se pueden describir cuantitativamente. Debido a esta complejidad en las ciencias de la Tierra y en sus procesos, es obligado pensar en la incertidumbre que hay acerca de la forma en la que se comporta el fenómeno en las ubicaciones conocidas.

Los *modelos de función aleatoria* que se utilizan, reconocen de manera importante esta incertidumbre y se convierten en una herramienta para estimar valores en ubicaciones desconocidas una vez que se hayan formulado suposiciones sobre las características estadísticas del fenómeno.

Con cualquier procedimiento de estimación, ya sea determinístico o probabilístico, es inevitable saber qué tan acertados son los valores estimados.

Dentro de la conceptualización del fenómeno que permite predecir qué está pasando en las ubicaciones donde no existe muestreo, los modelos no son ni buenos ni malos, sin información adicional no hay prueba de que los modelos son válidos. Sin embargo, los modelos se pueden clasificar en apropiados e inapropiados, tomando en cuenta las metas del estudio y toda la información cualitativa posible.

El modelo planteado debe tener un constante recordatorio de qué parte es modelada y qué parte es real, esto permite ver más claramente la naturaleza de las suposiciones. Con una muestra de datos que tiene una visión limitada del perfil completo, se vuelve muy tentador el reemplazar la realidad de un problema de estimación con la conveniencia matemática de un modelo, y si se hace eso, se pierden de vista las suposiciones sobre las cuales el procedimiento de estimación está basado. Un típico indicio de esto es la dependencia de las pruebas de hipótesis estadísticas para probar los parámetros del modelo. Ya que mientras estas pruebas demuestran que el modelo es consistente consigo mismo, no prueban que el modelo es apropiado.

En los modelos probabilísticos la muestra es vista como el resultado de un proceso aleatorio. Esta conceptualización resulta bastante útil para el problema de estimación, ya que el resultado de un proceso aleatorio sí ayuda con el problema de predecir valores, no sólo aporta en cuanto a los procedimientos y métodos de estimación, sino que también evalúa la exactitud de los estimados y asigna intervalos de confianza a ellos.

La aplicación de los métodos geoestadísticos de estimación más utilizados no requiere de una definición completa de un proceso aleatorio, es suficiente con especificar sólo algunos parámetros del proceso aleatorio, por ejemplo, la media y varianza de una combinación lineal de variables aleatorias.

2.1.5. Estadísticos (media, varianza, covarianza y semivariograma)

El *momento de primer orden* de $Z(x)$ es la esperanza matemática o media definida como:

$$m(x) = E\left[Z(x)\right] \tag{2.3}$$

Los momentos de segundo orden considerados en geoestadística son:

1. La varianza de $Z(x)$

$$\sigma^2(x) = Var\left[Z(x)\right] = E\left[\{Z(x) - m(x)\}^2\right] \tag{2.4}$$

2. La covarianza de dos variables aleatorias $Z(x_i)$ y $Z(x_j)$ o también llamada función de autocovarianza es definida como:

$$C(x_i, x_j) = E\left[\{Z(x_i) - m(x_i)\}\{Z(x_j) - m(x_j)\}\right] \tag{2.5}$$

3. El semivariograma $\gamma(x_i, x_j)$ o también llamado función de semivarianza se define como:

$$\begin{aligned} 2\gamma(x_i, x_j) &= Var\left[Z(x_i) - Z(x_j)\right] \\ \gamma(x_i, x_j) &= \frac{1}{2}E\left[\{Z(x_i) - Z(x_j)\}^2\right] \end{aligned} \tag{2.6}$$

Con frecuencia se usa el término del variograma $2\gamma(x_i, x_j)$ indistintamente para designar también a $\gamma(x_i, x_j)$.

Cabe destacar que tanto la varianza como el variograma son siempre positivos, mientras que la covarianza puede tomar valores negativos.

2.1.6. Estacionariedad de funciones aleatorias

Se dice que una función aleatoria es estrictamente estacionaria[2] si su función de distribución es invariante a cualquier traslación respecto a un vector $\underline{h}$ o lo que es equivalente, la función de distribución del vector aleatorio $Z(x_1), Z(x_2), Z(x_3), \ldots, Z(x_n)$ es idéntica a la del vector $Z(x_1 + \underline{h}), Z(x_2 + \underline{h}), Z(x_3 + \underline{h}), \ldots, Z(x_n + \underline{h})$ para cualquier $\underline{h}$.

En términos prácticos, la hipótesis de estacionariedad se limita a los primeros momentos.

Se dice que una función aleatoria es estacionaria de segundo orden si se cumple que:

1. Su valor esperado existe y no depende de $\underline{x}$.

$$E\left[Z(x)\right] = m; \forall \underline{x} \tag{2.7}$$

2. Para cualquier par de variables aleatorias $Z(x)$ y $Z(x + h)$, su covarianza existe y sólo depende del vector de separación $\underline{\mathrm{h}}$.

$$C(h) \equiv C(x + h, x) = E\left[Z(x + h)Z(x)\right] - m^2 \tag{2.8}$$

La estacionariedad de la varianza implica que la varianza existe, es finita y no depende de x, es decir,

$$\sigma^2 = C(0) = Var\left[Z(x)\right] \tag{2.9}$$

Así mismo, bajo esta hipótesis el semivariograma también es estacionario y se cumple que:

$$\gamma(h) \equiv \gamma(x + h, x) = \frac{1}{2}E\left[\{Z(x + h) - Z(x)\}^2\right] \tag{2.10}$$

Además, existe una relación discreta entre el semivariograma y la función de covarianza

$$\gamma(h) = C(0) - C(h) \tag{2.11}$$

[2]Información complementaria y ref. bibliográficas [2][17]

En este caso resulta suficiente usar una de las dos funciones para caracterizar la dependencia espacial.

Para explicar mejor esta relación se puede ver por medio de la distribución de probabilidad univariada, si ésta no depende de la ubicación de x en todas las regiones muestreadas, entonces sin importar su ubicación, todas las parejas de variables aleatorias separadas por una distancia en particular h tienen la misma distribución de probabilidad conjunta.

Esta independencia de la distribución de probabilidad univariada y bivariada con respecto a la ubicación de $\underline{x}$ es referida como *estacionariedad*.

Si la función aleatoria es estacionaria, entonces los parámetros univariados como el valor esperado y la varianza pueden ser utilizados para resumir el comportamiento univariado del conjunto de variables aleatorias.

Adicionalmente, existen funciones aleatorias $Z(x)$ que representan fenómenos físicos que muestran una capacidad casi ilimitada de variación, por lo que para estas funciones no están definidas la varianza ni la covarianza. Sin embargo, existen casos en que sus incrementos o diferencias $Z(x+h) - Z(x)$ tienen varianza finita. En otras palabras, esto quiere decir que las diferencias son estacionarias de segundo orden.

Por lo tanto, las funciones aleatorias intrínsecas son aquellas que cumplen las siguientes condiciones:

1. El valor esperado de las diferencias es: $E\left[Z(x+h) - Z(x)\right] = 0$
2. La varianza de las diferencias es: $Var\left[Z(x+h) - Z(x)\right] = 2\gamma(h)$

Estas condiciones son conocidas como *Hipótesis Intrínseca*.

Cabe mencionar que una función aleatoria estacionaria de segundo orden es siempre intrínseca, haciendo notar que el regreso no se cumple. A las funciones que cumplen con la hipótesis intrínseca se les considera como *débilmente estacionarias*.

Funciones aleatorias no estacionarias

Las funciones aleatorias no estacionarias son aquellas cuya esperanza matemática depende de $\underline{x}$:

$$E[Z(x)] = m(x) \tag{2.12}$$

A $m(x)$ se le conoce como *función de deriva o tendencia*.

En los casos en los que existe tendencia, se descompone a la función aleatoria $Z(x)$ en la suma de una componente determinística $m(x)$ y un residuo $R(x)$ estacionario con media nula, es decir:

$$Z(x) = m(x) + R(x) \tag{2.13}$$

Para este caso, el semivariograma de $Z(x)$ depende de $\underline{x}$:

$$\gamma(x+h, x) = \gamma_R(h) + \frac{1}{2}\{m(x+h) - m(x)\}^2 \tag{2.14}$$

En el caso en que la deriva sea lineal $m(x) = m_0 + m_1 x$ el semivariograma no depende de $\underline{x}$.

$$\gamma(h) = \gamma_R(h) + \frac{1}{2}(m_1 h)^2 \tag{2.15}$$

Además, existe un enfoque que considera a las funciones aleatorias no estacionarias como intrínsecas de orden k. Esto significa que si se tomas las diferencias de un orden k apropiado, éstas resultan ser estacionarias.

2.1.7. Consideraciones para el cómputo del semivariograma

- La *anisotropía* (opuesta de isotropía) es la propiedad general según la cual determinadas propiedades varían según la dirección en que son examinadas, es decir, algo anisótropo podrá presentar diferentes características según la dirección. En un sentido más general, se habla de anisotropía cuando cualquier cambio de escala de una figura o un cuerpo, como en un gráfico $x - y$, se produce con factores distintos (o en dependencia de una función) en cada coordenada.

- *Lag*: Es un intervalo, es decir, se refiere a la distancia entre muestras o dispersiones sucesivas en los diagramas de dispersión.

- *Tolerancia*: Es el parámetro que se define para la distancia entre dispersiones, es decir, el parámetro de intervalo de los valores admisibles. Se debe definir una tolerancia direccional suficientemente grande que permita que haya suficientes[3] pares de datos, y sin embargo que tenga suficientemente pocos pares para que los variogramas de distintas direcciones no sean tan borrosos que no se puedan reconocer.

Los parámetros de distancia y dirección permiten producir una estructura más clara. Un incremento apropiado entre lags sucesivos y una tolerancia para la distancia pueden usualmente ser escogidos después de pocas pruebas.

- *Rango o alcance (Range)*: Es la distancia a la cual el variograma llega a su meseta. Conforme la separación entre parejas de datos crece, el valor correspondiente del variograma generalmente también crecerá. Sin embargo, eventualmente el incremento en la distancia de separación no causa un crecimiento correspondiente al incremento en el cuadrado del promedio de la diferencia entre pares de valores y el variograma que alcanza una meseta (comportamiento parecido a constante).

- *Umbral o meseta (Sill)*: La meseta que alcanza el variograma con la distancia del rango es llamada *Sill* y es el umbral al cual el variograma alcanza un comportamiento de tipo constante o de estabilización.

- *Efecto Pepita (Nugget Effect)*: Aunque el valor para $h = 0$ es estrictamente 0, algunos factores como el error al muestrear o una escala de variabilidad corta pueden causar que los valores de la muestra separados por una distancia muy pequeña sean muy distintos. Esto causa una discontinuidad en

[3] El significado de suficientes o pocos pares de datos se expresa mejor en el punto 4 de la sección de estimadores de variograma del presente capítulo

el origen del variograma. El salto vertical del valor 0 en el origen al valor del variograma en una distancia de separación extremadamente pequeña es llamado el efecto pepita (*Nugget Effect*). La relación del *Nugget Effect* hacia el umbral o meseta es frecuentemente llamado *Relative Nugget Effect* o efecto pepita relativo y es generalmente expresado en porcentajes.

2.1.8. Semivariograma

Se define al semivariograma[4] o comúnmente llamado también variograma.

Dada una variable regionalizada $Z(x)$ que cumpla la Hipótesis Intrínseca, entonces existe la función de semivarianza y se define:

$$\gamma(h) = \frac{1}{2} Var\left[Z(x) - Z(x+h)\right] = \frac{1}{2} E\left[\{Z(x) - Z(x+h)\}^2\right] \tag{2.16}$$

El semivariograma es una función que relaciona la semivarianza con el vector h conocido como *lag*, el cual denota la separación en distancia y dirección de cualquier par de valores $Z(x)$ y $Z(x+h)$.

2.1.9. Estimadores de semivariograma

El estimador más común del semivariograma es:

$$\widetilde{\gamma}(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} \left[Z(x_i + h) - Z(x_i)\right]^2 \tag{2.17}$$

Donde $N(h)$ es el número de pares $Z(x_i)$ y $Z(x_i + h)$ separados a una distancia $h = |\underline{h}|$.

$\widetilde{\gamma}(h)$ es un estimador no paramétrico, pero esencialmente es una media muestral, por lo que una de sus desventajas es la no robustez. Es óptimo cuando la distribución de la variable es normal, por lo que el sesgo es el mínimo posible.

Cabe destacar que por definición $\gamma(0) = 0$, pero en la práctica cuando $|\underline{h}|$ tiende a cero $\gamma^*(h)$ no necesariamente se anula. (*nugget efect*)

El estimador del variograma puede ser errático debido a algunas desviaciones como en los siguientes casos:

- Debido a que la distribución es sesgada y con grandes colas
- Si no hay heterocedasticidad
- Debido a sesgo en el muestreo
- Debido a valores atípicos

Para el cálculo del semivariograma se muestran algunas reglas prácticas, independientemente del estimador utilizado:

[4]Notas del Dr. Martín Díaz Viera, p.9, Ref. Bibliográfica [2]

1. Para el semivariograma suavizado o regularizado, los pares de observaciones se agrupan según la distancia dentro de un intervalo $h = |\underline{h}|$ con una tolerancia $\pm\Delta h/2$ y dentro de una dirección θ con una tolerancia $\pm\Delta\theta/2$.

2. El semivariograma muestral debe ser considerado solamente para pequeñas distancias por lo que generalmente se estima para valores de $|\underline{h}|$ menores que la mitad de la distancia máxima ($|\underline{h}| < d_{max}/2$).

3. A pesar de que la elección del número de intervalos es arbitraria, en el caso práctico se considera un *mínimo de 10 intervalos y un máximo de 25 intervalos* para determinar con precisión el rango y meseta del semivariograma.

4. La longitud de los intervalos debe ser suficiente para que el número de pares se encuentre entre 30 y 50 pares como mínimo para que el estimado del semivariograma sea relativamente estable.

Existen dos tipos principales de semivariogramas:

- El tipo de semivariograma transitivo que es el que se incrementa con el incremento del valor absoluto del intervalo $|\underline{h}|$ hasta alcanzar un valor máximo a partir del cual se mantiene relativamente constante y oscila alrededor del mismo. Donde las variables que tienen este tipo de semivariograma cumplen no sólo la hipótesis intrínseca, sino que también son estacionarias de segundo orden.

- El segundo tipo de semivariograma es el no acotado, el cual aparenta un incremento sin límites, por lo que no presentan varianza *a priori* finita.

Los elementos más importantes para ajustar el semivariograma son:

1. Un intercepto con la ordenada (efecto nugget).

2. Una sección monótonamente creciente (estructura del variograma).

3. Una meseta (estabilización de la función o rango de correlación).

2.1.10. Relación entre semivariograma, covarianza y correlograma

Los parámetros que son comúnmente utilizados para resumir el comportamiento bivariado de una función aleatoria estacionaria son la función de covarianza, $C(h)$, su correlograma o función de correlación, $\rho(h)$, y el variograma $\gamma(h)$. Para funciones aleatorias estacionarias, estos tres parámetros están relacionados por unas pocas expresiones simples[5].

Es importante remarcar que las *estadísticas descriptivas* calculadas de los datos de la muestra, *no son lo mismo que los parámetros del modelo conceptual.* En el caso de las expresiones simples de la función de covarianza, el correlograma, y el variograma que describen la relación entre los parámetros de una

[5]Información adicional en Isaaks 1989, p.55, Ref. Bibliográfica [1]

función aleatoria no son válidas para las estadísticas descriptivas correspondientes.

Para la mayoría de las funciones aleatorias utilizadas en la geoestadística práctica, las parejas de funciones aleatorias ampliamente separadas son independientes una de otra. Por lo tanto, la función de covarianza y el correlograma eventualmente llegan a cero, mientras que el variograma eventualmente llega a su máximo valor, usualmente llamado umbral o meseta (*sill*). Este valor del variograma es también la varianza de la función aleatoria, lo que permite que se exprese lo siguiente:

$$C_Z(h) = \gamma_Z(\infty) - \gamma_Z(h) \tag{2.18}$$

Para los modelos de función aleatoria estacionaria, los cuales son los más comunes en la práctica de geoestadística, la función de covarianza, el correlograma y el variograma proveen de exactamente la misma información de una forma ligeramente diferente. El correlograma y la covarianza tienen la misma forma, al correlograma se le hace una escala tal que su valor máximo es 1. El variograma también tiene la misma forma que la función de covarianza, excepto que está invertida; mientras la covarianza empieza de un máximo de $\widetilde{\sigma}^2$ en $h = 0$, y decrece hasta cero, el variograma empieza en cero y se incrementa hasta un máximo de $\widetilde{\sigma}^2$.

Al resumir la función de distribución conjunta entre pares de variables, como una función de distancia, el variograma (o la covarianza o el correlograma) proporciona una *medida de continuidad espacial* de la función aleatoria.

En la geoestadística, es preferible utilizar el variograma ya que la media es desconocida y no se requiere para el cálculo del variograma, mientras que si se requiere para la covarianza. Además, si la función no tiene varianza estacionaria (infinita), la covarianza no está definida en cero, mientras que el variograma es nulo.

2.1.11. Diagramas de dispersión

Cuando se realiza una gráfica de dispersión de los valores verdaderos contra los valores estimados se observa evidencia adicional sobre qué tan bueno ha sido el comportamiento del método. Un método de estimación perfecto sería el que siempre proporcionara valores que coinciden con el valor verdadero, para tales estimadores ideales, el diagrama de dispersión de los valores reales contra los estimados sería una línea recta de 45°. Los métodos de estimación se pueden juzgar en base a qué tan cercanos son los *diagramas de dispersión* a lo ideal.

En la práctica, siempre se tiene cierto error en los estimados y en la gráfica de dispersión de los valores verdaderos contra los estimados se obtiene una nube de puntos. También se utiliza el coeficiente de correlación, ya que es un buen índice para saber qué tan cerca están los datos de caer sobre la línea de 45°.

Si la media de los errores es 0 para cualquier rango de valores estimados, entonces la curva de esperanza condicional de los valores verdaderos dados los

estimados será una línea de 45°. Aún cuando rara vez se espera obtener un estimador completamente condicionalmente insesgado, comparar la curva de esperanza condicional con la línea de 45° ayuda a valorar el método de estimación y llega a hacer evidentes las causas del sesgo.

Por lo tanto, es necesario tener un resumen de la información que contiene un diagrama de dispersión. Una de las características esenciales de los diagramas de dispersión de h es la densidad de la nube de puntos. Un buen indicador de la densidad es el coeficiente de correlación. Conforme la nube de puntos se hace más densa, se espera que el coeficiente de correlación decrezca.

La relación entre el coeficiente de correlación de un diagrama de dispersión y h es llamada *función de correlación o correlograma*. El coeficiente de correlación depende de h, el cual, al ser un vector, tiene magnitud y dirección. Para mostrar gráficamente la función de correlación se puede utilizar un mapa de contorno que muestra la función de correlación de un diagrama de dispersión de h como función de la magnitud y dirección. Aunque esto proporciona un gráfico completo y efectivo de $\rho(h)$, no es el formato tradicional. En vez de eso, se grafica la función de correlación contra la magnitud de h por separado para varias direcciones, por lo que el coeficiente de correlación varía en función tanto de la distancia de separación como de la dirección.

Un índice alternativo para continuidad espacial es la covarianza $C(h)$. La relación entre la covarianza de un diagrama de dispersión y h es llamada *función de covarianza*. Esta función decrece constantemente de manera similar al coeficiente de correlación.

Otro buen índice para medir la densidad de la nube de puntos es el *momento de inercia* sobre la línea $x = y$, el cual puede ser calculado con lo siguiente:

$$momento\ de\ inercia = \frac{1}{2n}\sum_{i=1}^{n}(x_i - y_i)^2 \tag{2.19}$$

Es la mitad del promedio cuadrado de la diferencia entre las coordenadas x y y de cada par de puntos en el diagrama de dispersión, el factor $1/2$ es consecuencia del hecho de que el interés está en la distancia perpendicular de los puntos de la línea de 45°. En un diagrama de dispersión tiene mucha relevancia ya que se están emparejando valores de la misma variable el uno con el otro. Todos los puntos del diagrama de dispersión para $h = (0, 0)$ caen exactamente en la línea $x = y$ ya que cada valor está emparejado consigo mismo. Conforme $|h|$ se incrementa, los puntos se alejarán de la línea y el momento de inercia alrededor de la línea de 45° es por lo tanto una medida de la densidad de la nube.

A diferencia de los otros dos índices de continuidad espacial, el momento de inercia se incrementa conforme la nube se hace más densa. Por lo que conforme el coeficiente de correlación y la covarianza decrecen, el momento de inercia aumenta. La relación entre el momento de inercia de un diagrama de dispersión de h es tradicionalmente llamado *semivariograma* o simplemente *variograma* $\gamma(h)$.

Las tres estadísticas propuestas para la densidad de la nube de puntos son

sensibles a puntos discrepantes. Es importante evaluar el impacto significativo que puede tener un punto errático en cualquiera de las tres funciones. En la práctica, frecuentemente la función de correlación, la función de covarianza y el variograma no describen correctamente la continuidad espacial debido a unos cuantos puntos erráticos. Si la forma de cualquiera de estas funciones no está bien definida, vale la pena examinar los diagramas de dispersión para determinar si unos cuantos puntos tienen un efecto indebido.

En los diagramas de dispersión se pueden observar los datos atípicos, corroborando con el diagrama de caja y brazos o el histograma. Dentro de la estimación de geoestadística la influencia de estos outliers puede ser de gran importancia, por ello se deben de observar y tratar cuidadosamente estos posibles valores erráticos. Para reducir la influencia de los valores extremos, se puede proceder de las siguientes formas:

- Transformar los datos para reducir el sesgo o asimetría de sus histogramas.
- Emplear otros estadísticos para describir los gráficos de dispersión-h que sean sensibles a los valores extremos.

2.1.12. Análisis de continuidad espacial para el modelo del variograma

Existen tres funciones que resumen la continuidad espacial[6], la función de correlación $\rho(h)$, la función de covarianza $C(h)$, y el variograma $\gamma(h)$. Estas funciones utilizan estadísticas descriptivas de los diagramas de dispersión para describir cómo cambia la continuidad espacial como función de la distancia y dirección. Cualquiera de estas funciones es adecuada solamente para propósitos de descripción. Sin embargo, el variograma es la decisión más común. Aún cuando la covarianza y la correlación son funciones igualmente útiles, y probablemente en otros enfoques sean más comunes, para este enfoque se utilizará el variograma y sólo se recurrirá a otras estadísticas descriptivas en caso de que no se pueda encontrar otra forma de mejorar el variograma de muestra.

Normalmente se comienza el análisis de continuidad espacial con un variograma adireccional para el cual la tolerancia es suficientemente grande tal que la dirección para cualquier vector de separación h_{ij}, se vuelve importante. Con todas las direcciones posibles combinadas en un solo variograma, sólo la magnitud de h_{ij} es importante. Un variograma adireccional puede ser pensado en términos generales como un promedio de varios variogramas direccionales. No es estrictamente un promedio, ya que la ubicación de las muestras puede causar que algunas direcciones estén sobrerrepresentadas.

Los cálculos de un variograma adireccional no implican que la continuidad espacial es la misma en todas las direcciones, apenas sirve como un punto de inicio para establecer algunos de los parámetros que se requieren para los cálculos del variograma de muestra.

[6]Información adicional en Isaaks 1989, p.141, Ref. Bibliográfica [1]

Como la dirección no juega un papel en los cálculos del variograma adireccional, se concentra en encontrar los parámetros de distancia que producen la estructura más clara. Un incremento apropiado entre lags sucesivos y una tolerancia para la distancia pueden usualmente ser escogidos después de pocas pruebas.

Otra razón para empezar con cálculos adireccionales es que pueden servir como una advertencia para variogramas direccionales erráticos. El variograma adireccional contiene más pares de datos que cualquier variograma direccional y por lo tanto es más probable mostrar una estructura clara para interpretar. Si el variograma adireccional no produce una estructura clara, no se debe esperar mucho éxito con variogramas direccionales.

Una forma de buscar lo errático es examinar los diagramas de dispersión, ya que pueden revelar que un simple valor está teniendo una gran influencia en los cálculos. Un mapeo de las ubicaciones de pares de datos erráticos también puede revelar problemas que no se habían visto. Si la razón de lo errático se puede identificar, entonces se puede adaptar el cálculo del variograma para enfrentar el problema. En ocasiones esto implica que se deben remover por completo pares de datos de la muestra. En caso de que los intentos por mejorar el variograma de la muestra no funcionen, entonces se debe considerar utilizar una forma diferente de continuidad espacial.

Una vez que el variograma adireccional se comporta de manera correcta, se procede a explorar el patrón de anisotropía con los variogramas direccionales. En estudios prácticos, se conoce información con anticipación, es decir, dependiendo del fenómeno en cuestión se puede asignar una dirección. Por ejemplo, la información hidrológica de la contaminación de los mantos acuíferos podría ser muy útil al decidir la dirección del cálculo de los variogramas, en el caso de que ése sea el fenómeno en cuestión.

Si no se tiene la información, un mapa de contorno de la muestra puede ofrecer una idea de las direcciones del mínimo y máximo de continuidad. Sin embargo, no se debe utilizar sólo un mapa de contorno para decidir, a pesar de que la mayoría de las veces es acertado.

En adición, se necesitan decidir dos parámetros, uno es la distancia entre dispersiones sucesivas en los diagramas de dispersión, usualmente conocido como *lag* o incremento en el lag; el otro parámetro es la tolerancia que se permitirá para la distancia entre dispersiones.

El patrón de la muestra en ocasiones sugiere un incremento en el *lag*. Si las muestras están ubicadas dentro de una cuadrícula semiregular, se puede utilizar el espacio de la cuadrícula como lag. Si la muestra es muy aleatoria se puede utilizar un lag inicial y estimar el promedio de espacio entre muestras vecinas.

Si el patrón de la muestra es notablemente anisotrópico, con un espacio entre muestras mucho más pequeño en algunas direcciones que en otras, los parámetros de distancia dependerán de la dirección. En estos casos, no es recomendable utilizar el variograma adireccional para establecer los parámetros

de distancia, se deben agrupar las muestras que tienen un espaciamiento similar.

Una forma muy común es definir el *lag* de tolerancia como la mitad del *lag* de espacio, es decir, si las muestras están colocadas en una cuadrícula regular o semiregular, se puede escoger el *lag* de tolerancia como menos de la mitad del *lag* de espaciamiento. Aunque esto pueda resultar en que algunos pares de datos no sean usados en los cálculos del variograma podría hacer la estructura más clara.

La existencia de muestras ubicadas muy cerca una de la otra puede afectar la decisión de los parámetros de distancia. Se podría pensar en incluir un *lag* adicional para distancias de separación pequeñas y usar una tolerancia también pequeña para el primer *lag* de separación para que cualquier muestra duplicada o muestras gemelas se agrupen, proporcionando el punto del variograma de muestra más cercano al origen.

Una vez que se tienen las direcciones de máxima y mínima continuidad, lo que falta es la tolerancia angular o direccional. Al calcular los variogramas direccionales sería ideal escoger una tolerancia angular tan pequeña como sea posible para limitar lo borroso de la anisotropía que resulta de combinar pares de diferentes direcciones.

Desafortunadamente, la tolerancia direccional muy pequeña proporciona tan pocas parejas, que el variograma direccional se vuelve demasiado errático. Lo mejor es probar con distintas tolerancias y utilizar la más pequeña que todavía cumpla con buenos resultados.

Para cualquier *lag* en particular, el número de pares que contribuyen a los cálculos del variograma aumenta conforme la tolerancia direccional aumenta.

2.1.13. Modelos de variograma

La decisión de algún modelo de variograma[7] o covarianza es un paso muy importante en los procedimientos geoestadísticos de estimación. La elección de un modelo en particular de variograma implica directamente una creencia sobre algún tipo de continuidad espacial. Dentro de las muchas características de los datos referentes a ciencias de la tierra, el patrón de continuidad espacial es uno de los problemas más importantes de la estimación. Si el fenómeno es muy continuo, entonces los estimados basados sólo en las muestras más cercanas son los más confiables. Por otro lado, si el fenómeno es muy errático, entonces los estimados basados sólo en las muestras más cercanas pueden no ser muy confiables, para tal fenómeno los buenos estimados requieren del uso de mucha más información de la muestra más allá de sólo los datos cercanos.

Las funciones aleatorias para las que los espacios entre valores cercanos son muy diferentes tendrán variogramas que se elevan muy rápidamente del origen;

[7]Información adicional en Isaaks 1989, p.369 y notas del Dr. Martín Díaz Viera, p.17, Ref. Bibliográficas [1][2]

funciones aleatorias para las que los espacios entre valores cercanos son muy similares tendrán variogramas que se elevan mucho más lento. De igual manera entre más continua sea la estacionariedad, es decir, entre mayor sea la probabilidad de mantenerse en el mismo punto, el variograma alcanza el punto más alto de manera más lenta; mientras que en las que son menos continuas, es decir, que tienen una función de probabilidad en la que es menos probable que se quede en el mismo punto, o sea, menos estacionariedad, el variograma alcanza su punto más alto casi inmediatamente.

En la práctica de geoestadística usualmente se adopta una función aleatoria estacionaria como modelo y se especifica sólo su covarianza o variograma. Se pueden tener muchas versiones que sean convincentes al perfil real, es decir, que pasan por todos los puntos muestreados. Debido a eso, *la clave para una estimación exitosa, es escoger el variograma o la covarianza que capture el patrón de continuidad espacial que se considera que el perfil debe tener.* Se elige el variograma o la covarianza dependiendo de los conocimientos que se tengan sobre el fenómeno, es decir, que el mecanismo con el cual se generan los valores observados sea más o menos continuo, por lo que se debe escoger una función aleatoria que sea consistente con la información.

Para la construcción de los variogramas, se examinan las restricciones que un modelo debe respetar, por lo que los modelos autorizados de variograma deben cumplir que:

1. La función de covarianza $C(h)$ si existe debe ser positiva semidefinida.

2. Sea $Y = \sum_{i=1}^{n} \lambda_i Z(x_i)$, donde $\lambda_i = 1, \ldots, n$ son pesos arbitrarios y entonces $Var(Y) = \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j C(x_i, x_j)$, la varianza de Y puede ser positiva o cero, pero no negativa.

3. Para la restricción anterior la función de covarianza $C(x_i, x_j)$ debe asegurar la condición de la varianza por lo que, la matriz de covarianzas debe ser positiva definida.

4. Por último, el variograma debe tener un crecimiento inferior a h^2, es decir, $lim_{h \to 0} \frac{\gamma(h)}{h^2} = 0$ por lo que si no se cumple, puede ser indicador de no estacionariedad.

Cabe destacar que cualquier combinación lineal de modelos autorizados es un modelo autorizado.

Los modelos que se introducirán son los modelos básicos, son modelos simples, isotrópicos e independientes de dirección. Los modelos de variograma básicos pueden ser divididos en dos tipos: los que llegan a una meseta y los que no. Los modelos de variograma del primer tipo son mejor conocidos como modelos de transición. La meseta a la que llegan es llamada *sill* y la distancia a la cual ellos llegan a esta meseta es llamada rango.

Algunos de los modelos de transición llegan a su meseta asintóticamente. Para éstos modelos, el rango es definido arbitrariamente para ser la distancia a la cual el 95 % de la meseta es alcanzada. En esta parte, la meseta de todos los

modelos de transición se estandarizó a uno.

Los modelos de variograma del segundo tipo no alcanzan su meseta, pero continúan incrementándose mientras la magnitud de h se incremente. Estos modelos son frecuentemente necesarios cuando hay tendencia o desviación en los datos.

El grupo de modelos *transitivos o acotados* se deriva a partir de la noción de autocorrelación entre valores promedios dentro de los bloques. La idea es que la función aleatoria, de la cual la propiedad medida es una realización, depende del grado de solapamiento de los dos bloques, es decir, una zona de transición.

Modelo Esférico: Probablemente es el modelo más utilizado comúnmente. Se puede derivar el modelo considerando el traslape de los volúmenes de dos esferas de diámetro a y nombrando h a la distancia que separa sus centros. El volumen de intersección es:

$$V = \frac{\pi}{4}\frac{2a^3}{3} - a^2h + \frac{h^3}{3};\ para\ h \leq a \tag{2.20}$$

Dividiendo por el volumen de la esfera ($\pi a^3/6$) se obtiene la función de autocorrelación:

$$\rho(h) = 1 - \frac{3h}{2a} + \frac{1}{2}(\frac{h}{a})^3 \tag{2.21}$$

Y el semivariograma:

$$\gamma(h) = \left\{ \begin{array}{ll} \frac{S}{2}\left[3(\frac{h}{a}) - (\frac{h}{a})^3\right] & \text{para } 0 \leq h \leq a \\ S & \text{para } h > a \end{array} \right\} \tag{2.22}$$

Gradiente= $3S/(2a)$
Por lo cual su ecuación estadarizada es:

$$\gamma(h) = \left\{ \begin{array}{ll} 1{,}5\frac{h}{a} - 0{,}5(\frac{h}{a})^3 & \text{si } h \leq a \\ 1 & \text{en otro caso} \end{array} \right\} \tag{2.23}$$

Donde a es el rango.

Tiene un comportamiento lineal en distancias de separación cercanas al origen pero se aplana a grandes distancias, y alcanza la meseta en a. Cuando se ajusta este modelo a un variograma de muestra es frecuentemente útil recordar que la tangente al origen alcanza la meseta en aproximadamente dos tercios del rango.

El Modelo exponencial: Otro modelo de transición común es el modelo exponencial. Si el traslape de los bloques varía su tamaño de forma aleatoria, entonces el semivariograma resulta exponencial. En el caso isotrópico es:

$$\gamma(h) = S(1 - \exp(-\frac{h}{a}));\ h \geq 0 \tag{2.24}$$

Gradiente= S/a
Se considera como rango efectivo $r = 3a$.

Su ecuación estandarizada es:

$$\gamma(h) = 1 - \exp(-\frac{3h}{a}) \tag{2.25}$$

Este modelo alcanza su meseta asintóticamente con el rango definido como la distancia a la cual el valor del variograma es 95 % de la meseta. Como el modelo esférico, el modelo exponencial es lineal en las distancias cortas cercanas al origen, sin embargo, se eleva más abruptamente y luego se aplana más gradualmente. Cuando se ajusta el modelo al variograma de muestra es útil recordar que la tangente al origen alcanza la meseta en aproximadamente un quinto del rango.

El Modelo Gaussiano: El modelo Gaussiano es un modelo de transición que es frecuentemente usado para modelar fenómenos extremadamente continuos. Está dado por:

$$\gamma(h) = S(1 - \exp -(\frac{h}{a})^2);\ \text{para}\ h \geq 0 \tag{2.26}$$

Donde a es un parámetro no lineal que determina la escala espacial de la variación, como en el caso exponencial. El rango efectivo se considera $r = \sqrt{3a}$, que corresponde al valor $0{,}95S$ del variograma.

Su ecuación estandarizada es:

$$\gamma(h) = 1 - \exp(-\frac{3h^2}{a^2}) \tag{2.27}$$

Al igual que el modelo exponencial, el modelo Gaussiano alcanza su meseta asintóticamente, y el parámetro a está definido cómo el rango práctico o la distancia a la cual el valor del variograma es 95 % de la meseta. La característica diferente del modelo Gaussiano es el comportamiento parabólico cerca del origen. Es el único modelo de transición en el cual su forma tiene un punto de inflexión.

El modelo de efecto agujero o efecto pepita (nugget effect): Muchos variogramas de muestra tienen una discontinuidad en el origen. Mientras el valor del variograma para $h = 0$ sea estrictamente 0, el valor del variograma en cada distancia de separación puede ser significativamente más grande que 0 dando lugar a una discontinuidad.

El efecto agujero es indicativo de fenómenos con componentes periódicas. Las expresiones más comunes de modelos de semivariogramas son:

$$\gamma(h) = S(1 - \frac{\sin h}{h});\ \text{para}\ h > 0 \tag{2.28}$$

Este puede ser usado para representar procesos regularmente continuos y que muestran un comportamiento periódico, el cual es frecuentemente encontrado, donde existe una sucesión de zonas ricas y pobres. Es un modelo negativo definido en tres dimensiones.

Otra alternativa es:

$$\gamma(h) = S(1 - \cos h); \text{para}\ h \geq 0 \tag{2.29}$$

Si el efecto agujero es observado muy acentuado en cierta dirección, por ejemplo la vertical, cuando el fenómeno es una sucesión pseudoperiódica de estratificación horizontal, entonces este modelo es negativo definido en una dimensión.

El modelo Lineal: El modelo lineal no es un modelo de transición, ya que no alcanza una meseta, pero se incrementa linealmente con h. En su forma estandarizada se escribe simplemente como:

$$\gamma(h) = |h| \tag{2.30}$$

Ahora ya se tienen algunos modelos básicos para ajustar un variograma de muestra direccional. Actualmente, en el caso isotrópico, el variograma muestral depende únicamente de la distancia de separación y no de la dirección, así que todos los variogramas de muestra direccionales serán los mismos. En estos casos, se puede modelar el variograma de muestra adireccional como si fuera un variograma de muestra direccional. De hecho, el variograma de muestra adireccional se prefiere debido a que usualmente tiene mejor comportamiento y por lo tanto es más fácil de modelar.

Aunque en ocasiones se puede modelar un variograma de muestra adireccional satisfactoriamente usando un modelo básico, la combinación de modelos básicos se requiere y es más frecuente para obtener un ajuste satisfactorio. Esto lleva a una propiedad importante de los modelos de variograma positivos definidos; cualquier combinación de modelos de variograma positivos definidos con coeficientes positivos es también un modelo positivo definido.

Para ajustar una combinación de modelos de variograma básicos a un variograma de muestra direccional en particular, se debe decidir cuál de los modelos básicos describe mejor la forma total. Si el variograma de muestra tiene una meseta, uno de los modelos de transición será el más adecuado, sino entonces tal vez el modelo lineal sea más apropiado.

Ahora se mencionan algunos modelos *no acotados*:

Modelo potencia: Existen casos en que la varianza aparenta incrementarse indefinidamente. Así también si se toma cada vez un menor intervalo de muestreo, siempre existe alguna variación que queda sin resolver. Un punto de partida para entender esta variación es el movimiento Browniano en una dimensión, en el cual:

$$Z(x) - Z(x+h) = \epsilon \tag{2.31}$$

Es una variable aleatoria gaussiana, espacialmente independiente.

Su semivariograma es:

$$\gamma(h) = \frac{1}{2}h^{\theta};\ \text{para } 0 < \theta < 2 \tag{2.32}$$

Modelo efecto pepita puro: Formalmente se puede definir como:

$$\gamma(h) = S(1 - \delta(h)) \tag{2.33}$$

Modelo logarítmico (Modelo de Wijsian): Se define como:

$$\gamma(h) = k\log(h) \tag{2.34}$$

Puede ser de utilidad cuando el semivariograma experimental se comporta linealmente si se usa una escala logarítmica para las distancias.

2.1.14. Distribución de los errores de la estimación o residuales

Se define al error, r, como la diferencia entre el valor estimado y el valor real en un punto dado, entonces r_i es el error en la i-ésima estimación.

Si r_i es positivo, entonces se sobreestimó el valor verdadero; si r_i es negativo entonces se subestimó el valor verdadero. A estos errores también se les llama *residuales.*

La media de la distribución del error frecuentemente se refiere al sesgo. Lo ideal para cualquier método de estimación es producir estimadores insesgados. Además, la media no es la única medida para centrar. Idealmente, se desearía que la mediana y la moda de la distribución del error también fuera 0. Si se generan histogramas de la distribución de los errores se pueden observar tres casos:

1. Si la media es menor que cero, la distribución indica un sesgo negativo.
2. Si la media es mayor que cero, la distribución indica un sesgo positivo.
3. Si la media está centrada en cero, la distribución no tiene sesgo.

Una media de 0 puede ser el resultado de muchos subestimados combinados con pocos sobrestimados. Comúnmente es preferible tener una distribución más simétrica. La mediana de los errores sirve para verificar la simetría. Si tanto la media como la mediana son cercanas a 0, entonces no sólo los subestimados y los sobreestimados están en balance sino que también son simétricas sus magnitudes. Una diferencia apreciable entre la media y la mediana alerta que la magnitud de los sobreestimados seguramente no es la misma que la de los subestimados.

Además, se debe revisar la dispersión, se busca que la distribución de los errores tenga poca dispersión. Tanto la varianza como la desviación estándar son buenos criterios para evaluar la dispersión.

Las metas tanto de dispersión mínima como de que esté centrada a 0 no son independientes y habrá momentos en la práctica cuando se tenga que cambiar una por la otra. Dos estadísticas descriptivas que incorporan tanto el sesgo como la dispersión de la distribución de los errores son la media absoluta del error (MAE: *Mean Absolute Error*) y la media cuadrada del error (MSE: *Mean Squared Error*):

$$Media\ Absoluta\ del\ error = MAE = \frac{1}{n}\sum_{i=1}^{n}|r| \tag{2.35}$$

$$Media\ Cuadrada\ del\ error = MSE = \frac{1}{n}\sum_{i=1}^{n} r^2 \tag{2.36}$$

La MSE puede estar relacionada con otras estadísticas de la distribución de los errores:

$$MSE = varianza + (sesgo)^2 \tag{2.37}$$

Por otro lado, es deseable que esta distribución de los errores se cumpla para cualquier rango de los estimados, es decir, cualquier partición del conjunto de estimados. Entonces, si se realiza el mismo análisis para cada subconjunto, el sesgo debería seguir cercano a 0. Si se tiene una distribución de los errores insesgada en cada grupo, entonces se dice que es *condicionalmente insesgada*, sin embargo, es muy complicado obtenerla en la práctica.

Un conjunto de subconjuntos que son condicionalmente insesgados, es también globalmente condicional insesgado, pero cabe remarcar que el regreso no se cumple, no porque sea globalmente condicional insesgado va a cumplir que sus subconjuntos sean condicionalmente insesgados, inclusive en estos conjuntos insesgados pueden existir sobreestimaciones o subestimaciones en algún rango de valores. Una manera de revisar el sesgo condicional es graficar los valores estimados contra los residuales.

2.2. Análisis Exploratorio

El Análisis Exploratorio de Datos[8] (A.E.D.) es un conjunto de técnicas estadísticas cuya finalidad es examinar los datos previamente a la aplicación de cualquier técnica estadística, así como conocer su posible distribución en base a gráficos y estadísticos numéricos.

El AED consigue un entendimiento básico de sus datos y de las relaciones existentes entre las variables analizadas. Además, proporciona métodos sencillos para organizar y preparar los datos, detectar fallos en el diseño, tratamiento y evaluación de datos ausentes, identificación de casos atípicos y comprobación de los supuestos de la metodología a utilizar.

Este examen previo de los datos es un paso necesario que lleva tiempo, y que comúnmente puede ser descuidado por parte de los analistas de datos. Las tareas implícitas en dicho examen pueden parecer insignificantes y sin consecuencias a primera vista, pero son una parte esencial de cualquier análisis estadístico.

2.2.1. Estadísticas descriptivas

La estadística descriptiva se refiere a la recolección, presentación, descripción, análisis e interpretación de una colección de datos. Permite obtener un mayor conocimiento de la muestra acerca de características importantes como:

- Localización o Posición (Utiliza los cuantiles, percentiles, deciles)

[8]Más descripciones en Figueras y Gargallo, 2008. Ref. Bibliográfica [20]

- Centralización (Moda, media, y mediana)
- Dispersión (varianza, desviación estándar, coeficiente de variación, rango)
- Forma (asimetría, apuntamiento o curtosis)

En el tema de geoestadística, se busca que el centro o media, dispersión y coeficiente de asimetría sean lo más cercano a 0 que se pueda, ya que esto nos habla de un comportamiento distribucional.

En particular, para los casos que se realizarán se tomarán en cuenta las siguientes estadísticas:

1. El número total de datos de la muestra, ya que existe una gran influencia en la estimación dependiendo de la cantidad de datos que se conocen y es de gran importancia tomarlo en cuenta.
2. La mínima y máxima distancia entre ubicaciones, ya que es lo que proporciona el lag que sería más adecuado para el modelo.
3. La media y mediana que son medidas de centralización y proporcionan gran información acerca de los datos, nos proporcionan la distribución de las frecuencias y debido a que para cumplir uno de los supuestos de la metodología es necesario tener una muestra simétrica, las medidas de centralización son el primer paso para dar una idea de la distribución.
4. La varianza, desviación estándar, coeficiente de variación y rango son medidas de dispersión. Las medidas de dispersión permiten conocer hasta qué punto las medidas de tendencia central son representativas como síntesis de la información. El coeficiente de variación se define como el cociente entre la desviación típica y el valor absoluto de la media aritmética. El rango es la diferencia entre el valor máximo y el mínimo de las observaciones.
5. Los cuantiles dividen a la muestra en intervalos del mismo número de valores, por lo que son de gran importancia para la agrupación de datos y su importancia con la vecindad, así como el lag que se proporcionará.
6. La asimetría compara la forma que tiene la representación gráfica, bien sea el histograma o el diagrama de barras de la distribución, con la distribución normal. Muestra las características de forma de los datos.
7. La curtosis mide la mayor o menor cantidad de datos que se agrupan en torno a la moda.

Una vez que se conoce el comportamiento distribucional, se pueden utilizar distintos métodos para mejorarlo y poder llegar a una estimación más acertada con la menor cantidad de cálculos, sin embargo el comportamiento se analiza a través de diversos métodos, no sólo el análisis descriptivo sino también gráficos y pruebas no paramétricas.

2.2.2. Distribución espacial

Es el patrón que forma espacialmente la distribución de los datos y en particular resalta los outliers espaciales. También se refiere a los datos que tienen un valor muy diferente a los valores de los datos que lo rodean espacialmente. Para la distribución espacial se utiliza un gráfico de dispersión de los datos con un histograma a color que permite conocer la ubicación de las muestras y su valor.

2.2.3. Histograma y Q-Q plot

El histograma es una gráfica de barras que permite describir el comportamiento de un conjunto de datos en cuanto a su tendencia central, forma y dispersión. Está muy ligada con la información proporcionada en las estadísticas descriptivas y es el mejor gráfico para detectar normalidad de la muestra.

El Q-Q plot es un gráfico de cuantil contra cuantil para el diagnóstico de diferencias entre distribuciones de probabilidad. Comúnmente compara el cuantil de la distribución de un conjunto de datos contra el cuantil de la distribución normal. Sin embargo, también permite observar cuan cerca está la distribución de un conjunto de datos de alguna distribución ideal o comparar la distribución de dos conjuntos de datos cualquiera.

Obtener ambos gráficos es información muy relevante para evaluar si la muestra tiene distribución normal o no, así como observar la simetría en el histograma y sus estadísticas básicas.

2.2.4. Pruebas no paramétricas

Se utilizan pruebas no paramétricas para verificar la normalidad en los datos, buscando que sea lo más simétrica posible. Comúnmente se utiliza la prueba Kolmogorov-Smirnov que es una prueba que se utiliza para determinar la bondad de ajuste de dos distribuciones de probabilidad entre sí. Sin embargo, la prueba de Lilliefors conlleva algunas mejoras con respecto a la de Kolmogorov-Smirnov; y, en general, las pruebas Shapiro-Wilk o Anderson-Darling son alternativas más potentes.

2.2.5. Estacionariedad en el análisis exploratorio

En el estricto sentido, la estacionariedad requiere que todos los momentos (estadísticos) sean invariables ante traslación, pero esto no puede ser verificado en la parte experimental, usualmente sólo se puede contar con que dos momentos sean constantes, la media y la covarianza. Esto es la estacionariedad de segundo orden. Esto es que el valor esperado de $Z(x)$ no depende de $\underline{x}$ y que la función de covarianza de los valores existentes entre cualesquiera dos puntos x y $x + h$ depende del vector h, pero no de $\underline{x}$.

En la práctica, se utiliza un gráfico de los datos con respecto a sus distancias y se observa si existe un comportamiento de media constante. A menudo la aplicación de los supuestos no se satisface, por lo que es lógico enfocar la visión de estacionariedad a si hay tendencia. Resulta más práctico visualizar información

que indique que no se tenga estacionariedad con lo que para casos experimentales se reduce a ubicar si existe la tendencia.

2.2.6. Tendencia en el análisis exploratorio

En la práctica, en principio se realiza un gráfico de dispersión de la función aleatoria $Z(x)$ con respecto de la coordenada x y otro de la función aleatoria $Z(x)$ con respecto de la coordenada y, es decir, se realiza una exploración visual en cada una de las direcciones (x, y), para observar el comportamiento de los datos y dar a conocer si es posible que exista tendencia.

Posteriormente, se recuerda la definición de funciones aleatorias no estacionarias, las cuales son aquellas cuya esperanza matemática depende de $\underline{x}$:

$$E\left[Z(x)\right] = m(x) \tag{2.38}$$

Donde a $m(x)$ se le conoce como *función de deriva o tendencia.*

Si existe tendencia, el semivariograma de $Z(x)$ depende de x y por lo tanto se ve afectado gráficamente. Un indicador para la detección de tendencia es un semivariograma con comportamiento de h^2. En este caso, se descompone a la variable $Z(x)$ en:

$$Z(x) = m(x) + R(x) \tag{2.39}$$

Si la tendencia es lineal, $m(x)$ se descompone como:

$$m(x) = m_0 + m_1 x \tag{2.40}$$

Si la tendencia no es lineal, se prueba si es de segundo grado y así sucesivamente hasta que los residuos se puedan considerar estacionarios. Y se continúa la metodología sobre los residuos, ya que éstos son estacionarios.

Además, se pueden considerar a las funciones aleatorias no estacionarias como intrínsecas de orden k. Esto significa que se toman las diferencias de un orden k apropiado y éstas resultan ser estacionarias.

Por lo tanto, si no existe estacionariedad en la media se debe proceder asumiendo que $m(x)$ puede ser representado como un polinomio de orden finito k.

2.2.7. Anisotropía

La anisotropía geométrica está caracterizada por un variograma de muestra direccional que tiene aproximadamente el mismo umbral o meseta pero diferentes rangos (en el caso de un variograma lineal, la pendiente varía con respecto a la dirección).

La anisotropía zonal es en la cual el valor del umbral o meseta cambia con la dirección mientras que el rango se mantiene constante. En la práctica, rara vez se encuentra una anisotropía zonal pura; es más común encontrar una mezcla de anisotropías zonales y geométricas juntas. Cuando se combinan en el modelo,

tanto el umbral o meseta como el rango cambian con la dirección.

En la práctica se estudian 4 direcciones, estimando los semivariogramas y determinando los rangos para los mismos, luego se construye el gráfico direccional de los rangos para decidir si hay anisotropía geométrica presente o no.

En presencia de anisotropía geométrica, el gráfico direccional de los rangos forma una elipse, en la cual el eje menor B es el rango de la dirección de más rápida variación y A, el eje mayor, está en la dirección en que la variabilidad es más lenta. La relación $\lambda = A/B$ es una medida de anisotropía.[9]

En caso de anisotropía, si se desea diseñar una red óptima de muestreo, se debe hacer en forma rectangular haciendo coincidir los lados con las direcciones de los ejes principales y las longitudes de los mismos deben estar en la proporción λ, donde el lado menor le correspondería a la dirección del eje B.

De igual manera, los mapas de contorno del variograma de superficie o los mapas de superficie anisotrópicos son muy útiles para determinar las direcciones. Otra alternativa, es graficar los rangos experimentales para los diferentes variogramas de muestra direccionales en un diagrama de rosa.

La información cualitativa, como podrían ser las características de las rocas que se estudian, es usualmente muy útil para identificar los ejes de la anisotropía. En general, el conocimiento del origen del fenómeno de estudio puede ayudar a conocer o suponer las direcciones de anisotropía.

2.3. Análisis Estructural

En esta etapa se estudia la continuidad espacial de la variable. Se calcula el variograma experimental, o cualquier otra función que explique la variabilidad espacial, se ajusta a los datos un variograma teórico y se analiza e interpreta dicho ajuste al modelo paramétrico seleccionado.

El análisis estructural es una etapa especialmente crítica. La obtención de modelos geoestadísticos realistas conllevan un estudio riguroso del variograma o de cualquier función análoga que caracterice la variación espacial del atributo. En diversos trabajos, suelen usarse diversos algoritmos geoestadísticos sin analizar previamente las posibles estructuras espaciales, lo que conlleva a un modelo de estimación no acertado.

2.3.1. Variograma adireccional o también llamado omnidireccional

Se define como un variograma válido para todas las direcciones, o como aquel en el cual la tolerancia direccional es de 360°. Evidentemente, este variograma será función sólo de la distancia, h. Se puede considerar como un variograma

[9] Notas del Dr. Martín Díaz Viera, p.25. Ref. Bibliográfica [2]

medio para todas las direcciones. Sin embargo, el cálculo de un variograma omnidireccional no significa que la continuidad espacial sea idéntica en todas las direcciones, simplemente constituye el inicio del análisis estructural, sirviendo para determinar los parámetros relacionados con la distancia que generan los mejores resultados, ya que no depende de la dirección. Esos parámetros serán el incremento de la distancia y la tolerancia dimensional.

Si el muestreo se ha realizado de forma regular sobre el área experimental, la distancia entre muestras puede considerarse como incrementos de la distancia. Sin embargo, si el muestreo es aleatorio, se puede elegir inicialmente un incremento de la distancia que equivalga, de forma aproximada, al espaciamiento medio entre muestras adyacentes.

Para la tolerancia dimensional, generalmente se toma la mitad del incremento de la distancia. De forma práctica, se realizan pruebas con diversos valores de la distancia h, y con distintas tolerancias sobre los mismos. Aquellos que generen la mejor estructura en el variograma serán los seleccionados.

Si después de varios intentos, no se consigue obtener un variograma omnidireccional adecuado, no se puede esperar que los variogramas direccionales sean mejores, ya que el omnidireccional es el que más valores muestrales incluye. Una revisión de los gráficos de dispersión-h puede contribuir a encontrar los valores erráticos que causan los malos resultados. Si a pesar de todo, no se consigue un variograma omnidireccional óptimo, se debe emplear otra medida distinta de la continuidad espacial, como la función de covarianza o el correlograma.

En resumen, se estima una función que describa el grado de correlación espacial de la propiedad que se estudia. El variograma adireccional se estima tomando la dirección de 0° y una ventana de ±90°. El tamaño del intervalo *lag* se elige considerando una cantidad entre 10 y 25 intervalos para la mitad de la distancia máxima de separación de los puntos.

2.3.2. Variograma en 4 direcciones o variogramas direccionales

Conseguido el variograma omnidireccional, deben encontrarse los posibles patrones de anisotropía, calculando los variogramas direccionales. Para ello, puede ser de gran ayuda el conocimiento del fenómeno bajo estudio. Por ejemplo, si se analiza la distribución de un contaminante transportado por el aire, es fundamental conocer las direcciones de los vientos predominantes. Si el contaminante es transportado a través del agua subterránea, el conocimiento del acuífero puede ayudar a determinar las direcciones principales.

En resumen, se estima el variograma en 4 direcciones: 0°, 45°, 90°, 135° con ventanas de ±22,5°, y los intervalos se eligen con el mismo criterio que el variograma adireccional. Se verifica si existe anisotropía geométrica cuando los alcances de los variogramas en las direcciones son significativamente diferentes. Si la anisotropía es significativa, se determinan los alcances (radio de correlación) en las direcciones de mayor y menor valor, con lo cual se construyen los

modelos anisotrópicos.

2.3.3. Anisotropía en el análisis estructural

Frecuentemente, los variogramas muestrales o experimentales direccionales revelan grandes cambios en el rango o en el umbral o meseta conforme la dirección cambia. Cuando el variograma de superficie muestra que el rango cambia con la dirección mientras que el umbral o meseta permanece constante, es un tipo de anisotropía conocida como *anisotropía geométrica*. En el caso de la *anisotropía zonal*, el umbral o meseta cambia con la dirección mientras el rango permanece constante.

Dado un conjunto de variogramas de muestra, que muestran el rango y/o el umbral o meseta cambiando con la dirección, se comienza por identificar los ejes de anisotropía. Esto es usualmente determinando experimentalmente la dirección correspondiente al mínimo y al máximo del rango, o el máximo o mínimo del umbral o meseta en el caso de la anisotropía zonal.

Si los valores de la muestra son más continuos en una dirección que en otra, entonces la elipse de anisotropía estrá orientada con sus ejes mayores paralelos a la dirección de máxima continuidad. La anisotropía de la elipse es usualmente determinada por la anisotropía evidente en cierta medida de la continuidad espacial, típicamente el variograma muestral. Si no hay anisotropía evidente, la elipse de anisotropía se convierte en un círculo y la cuestión de orientación no es relevante.

En el caso de que exista anisotropía, la mejor opción para determinar las direcciones de anisotropía es la realización de unos pocos variogramas direccionales, 9 o 10, y la utilización del mapa anisotrópico o la elaboración de un diagrama de rosa. Para el diagrama de rosa, se traza un segmento en cada dirección elegida cuya longitud sea proporcional al rango, o a un valor próximo del variograma direccional del cual provenga. Tanto para el mapa anisotrópico como para el diagrama de rosa, los ejes mayor y menor de la elipse que mejor se ajuste a los extremos de los segmentos representarán las direcciones principales de anisotropía[10].

Establecidas las direcciones de máxima y mínima continuidad, se debe seleccionar la tolerancia direccional. Idealmente, debería ser lo menor posible. Lo que ocurre es que con tolerancias direccionales reducidas, el número de datos abarcados es muy pequeño dando lugar a valores erráticos en los variogramas.

En la práctica, se prueban varias tolerancias y se escoge la menor que genera los mejores resultados. Conviene indicar que los variogramas son muy sensibles a los datos anómalos (outliers), con valores muy grandes o muy pequeños. De forma práctica, si unos pocos puntos erráticos hacen que la forma de estas

[10] *"Having identified the axes of anisotropy, the next step is to put together a model that describes how the variograma changes as the distance and direction change"*, Isaaks 1989. Ref. Bibliográfica [1]

funciones se vea alterada, es necesario examinarlos cuidadosamente y comprobar que no son fruto de algún tipo de error.

2.3.4. Tendencia en el análisis estructural

Si se supone que $m(x)$ puede ser representado como un polinomio de orden finito k. Para continuar se debe conocer *a priori* el orden k del polinomio que mejor describe o explica la tendencia[11] y la función de semivarianzas o variograma γ de la función aleatoria $Z(x)$ sin tendencia.

Los supuestos anteriores conducen en la práctica a dos problemas:

1. El orden k del polinomio nunca es conocido, hay que adivinarlo.
2. El variograma γ tampoco es conocido y hay que estimarlo a partir de los residuales (datos -deriva), es decir, $R(x) = Z(x) - m(x)$.

El problema real cuando se utiliza esta descomposición para estimar, es que cuando existe una sola realización de la función aleatoria no estacionaria $Z(x)$ resulta imposible estimar el variograma. Lo que se puede hacer es intentar eliminar la deriva $m(x)$ y trabajar con los residuos $R(x) = Z(x) - m(x)$ los cuales son estacionarios o al menos intrínsecos. Esto significa que la deriva tiene que ser estimada a partir de los valores muestrales.

Por lo tanto, el proceso entero puede ser resumido como sigue:

1. Se necesita conocer el variograma $\gamma(h)$ de $Z(x)$, pero como éste no puede ser directamente estimado uno trata de estimar el variograma de los residuos $\gamma_R(h)$.
2. Para esto se requiere seleccionar un tamaño (un radio r_v) de la vecindad y dentro de ésta se supone un tipo de deriva $m(x)$, en general, se toma un polinomio en $\underline{x}$, de cierto orden k.
3. Los coeficientes del polinomio de la deriva son estimados haciendo una suposición del tipo del variograma $\gamma_R(h)$.
4. Con los coeficientes de la deriva podemos obtener los residuos $R(x_i)$ en los puntos muestrales x_i.
5. Se calcula el variograma experimental de los residuos $\gamma_R^*(h)$.
6. El variograma teórico esperado de los residuos es calculado $\gamma_R(h)$.
7. El variograma teórico supuesto al $\gamma_R(h)$ y el experimental resultante de los residuos $\gamma_R^*(h)$ son comparados en algún sentido de su bondad de ajuste según la norma $||\gamma_R^*(h) - \gamma_R(h)||$ que se elija.

Si el ajuste es razonable, entonces la vecindad tomada, el tipo de deriva y el variograma supuestos son correctos. En caso contrario, una de los parámetros del procedimiento es cambiado y el proceso comienza otra vez. El algoritmo es simple, pero desde el punto de vista práctico puede consumir una gran cantidad de tiempo y no ofrece ninguna garantía de que exista convergencia.

[11] Notas del Dr. Martín Díaz Viera, p.43. Ref. Bibliográfica [2]

2.3.5. Ajuste de los modelos

La modelación del variograma consiste en buscar una función analítica que represente adecuadamente los valores estimados del variograma. El proceso de modelación se reduce a determinar cuál modelo y con qué parámetros se ajusta mejor a los valores estimados del variograma[12].

Un compromiso entre la bondad de ajuste y la complejidad del modelo (número de parámetros) puede ofrecerlo el Criterio de Akaike (AIC), que se define como:

$$AIC = -2ln(máx.verosimilitud) + 2(núm.deparámetros) \tag{2.41}$$

Y se puede estimar usando:

$$AIC^* = \left\{ n \ln \frac{2\pi}{n} + n + 2 \right\} + n \ln R + 2p \tag{2.42}$$

Debido a que la cantidad $\left\{ n \ln \frac{2\pi}{n} + n + 2 \right\}$ es constante, independientemente del tipo de modelo, entonces para fines prácticos se calcula:

$$\widetilde{A} = n \ln R + 2p \tag{2.43}$$

Que es un estimador simplificado del Criterio de Información de Akaike.

Donde,
n es el número de valores estimados $\gamma^*(h_i), i = 1, \ldots, n$ del variograma muestral.
R es la suma residual de los cuadrados de las diferencias entre los valores experimentales $\gamma^*(h_i)$ y los del modelo ajustado $\gamma(h_i)$, es decir, $R = \sum_{i=1}^{n} (\gamma(h_i) - \gamma^*(h_i))^2$.
Mientras que p es el número de parámetros del modelo de variograma ajustado $\gamma(h)$.

Se considera que el modelo que presenta el menor AIC es el mejor. O también utilizando la bondad de ajuste (valor medio cuadrático del error). En consecuencia, se puede utilizar la menor suma de cuadrados de los errores (SCE).

La modelación del variograma continúa con un proceso de prueba y error de manera visual, modificando los parámetros hasta obtener un compromiso razonable según el criterio de Akaike o el criterio de la suma de cuadrados del error.

Cabe destacar, que la decisión usualmente depende del comportamiento del variograma de muestra cerca del origen. Si el fenómeno subyacente es bastante continuo, el variograma muestral tiene un comportamiento parabólico cerca del origen; en estas ocasiones el modelo Gaussiano será el que mejor se ajuste. Si el variograma de muestra tiene un comportamiento lineal cerca del origen, ya sea el modelo esférico o exponencial es preferible. Frecuentemente se puede ajustar una línea recta a los primeros puntos en el variograma muestral. Si la línea intersecta con el umbral en aproximadamente un quinto del rango, entonces

[12]Notas Dr. Martín Díaz Viera, p.27. Ref. Bibliográfica [2]

probablemente el modelo exponencial será un mejor ajuste que el esférico. Si se intersecta en aproximadamente dos tercios del rango, entonces probablemente el modelo esférico se ajustará mejor.

Si la mayoría de las características del variograma muestral pueden ser capturadas con un modelo simple, entonces se obtendrán soluciones que son tan precisas como las que se encuentran usando un modelo más complejo. Por ejemplo, si el modelo exponencial se ajusta al variograma muestral tan bien como dos modelos esféricos anidados, entonces es preferible utilizar el modelo exponencial.

Para decidir si una característica en particular del variograma muestral debe ser modelada, es conveniente considerar si la característica tiene una explicación. Si la información cualitativa acerca del origen del fenómeno explica o confirma una característica en particular del variograma de muestra, entonces vale la pena construir un modelo que incluya ésa característica. Si no existe explicación, entonces la característica puede ser falsa y no vale la pena modelarla.

2.3.6. Validación del modelo

Validación cruzada

La validación cruzada[13] es una técnica que permite comparar los valores estimados con los valores reales usando sólo la información disponible en la base de datos. Un estudio de validación cruzada puede ayudar a decidir entre diferentes procedimientos de ponderación, entre diferentes estrategias de búsqueda o entre diferentes modelos de variogramas. Desafortunadamente, los resultados de la validación cruzada son más comúnmente usados sólo para comparar distribuciones de la estimación de los errores o residuales de diferentes procedimientos de estimación. Tal comparación, especialmente si se comparan técnicas similares, usualmente queda muy lejos de indicar claramente qué alternativa es mejor. Sin embargo, la validación cruzada contiene importante información espacial y realiza un estudio cuidadoso de los residuos, con lo que puede dar una idea de donde tiene problema la estimación. Estas ideas pueden llevar a mejoras específicas del procedimiento de estimación, por lo que la validación cruzada es un paso preliminar importante antes de que los estimados finales sean calculados.

En la validación cruzada, el método de estimación se pone a prueba en las ubicaciones de la muestra que sí se tienen. Un dato de la muestra en una ubicación en particular es temporalmente removido de la muestra de datos; el valor de ésa ubicación es estimado utilizando los datos que quedan con el método seleccionado. Ahora se compara este valor con el valor que se había retirado de la muestra. Es decir, se imita el proceso de estimación pero en ubicaciones que sí se conocen para comparar los valores verdaderos con los valores estimados. El proceso se repite para todas las muestras. Entonces, se comparan los valores estimados contra reales y se utilizan las mismas herramientas gráficas estadísticas que se utilizan comúnmente.

[13]Información adicional en Isaaks 1989, p.351. Notas del Dr. Martín Díaz Viera, p.28. Ref. Bibliográfica [1][2]

Existen limitaciones para la validación cruzada que deberían ser tomadas en cuenta cuando se analizan los resultados del estudio de validación cruzada. Por ejemplo, es usual más no acertado el comparar valores estimados contra reales de algunas regiones muestreadas, sin embargo, el comportamiento en una región muestreada no necesariamente describe el comportamiento en las regiones no muestreadas.

En resumen, el método de validación cruzada o *leave one out* resulta atractivo por su sencillez y eficiencia. Consiste en sacar un elemento de la muestra y estimar el valor en ese punto usando Kriging con el modelo de variograma obtenido. De forma análoga se actúa para el resto de los elementos de la muestra. Como resultado se obtiene un mapa de las diferencias $Z(x_i) - Z^*(x_i), i = 1, \ldots, n$ entre el valor real y el estimado.

En consecuencia, si el modelo del semivariograma refleja adecuadamente la estructura espacial implícita en el conjunto de datos, entonces los valores estimados deben ser cercanos a los valores observados.

Esta *"cercanía"* puede ser caracterizada según las siguientes estadísticas:

1. $\frac{1}{n}\sum_{i=1}^{n}[Z(x_i) - Z^*(x_i)]$ es cercano a 0.
2. $\frac{1}{n}\sum_{i=1}^{n}[Z(x_i) - Z^*(x_i)]^2$ es pequeño
3. $\frac{1}{n}\sum_{i=1}^{n}\left[\frac{Z(x_i)-Z^*(x_i)}{\sigma_i}\right]$ es cercano a 1

Donde,
$Z(x_i)$ son los valores muestrales de la propiedad en x_i.
$Z^*(x_i)$ son los valores estimados de la propiedad en el punto x_i.
σ_i es la desviación estándar de la estimación en el punto x_i.

Análisis de los errores o residuales

Posteriormente a la validación cruzada, se realiza un análisis descriptivo de las diferencias $(Z - Z^*)$ entre los valores reales (Z) y los estimados (Z^*). Se revisan de manera combinada los siguientes criterios:

1. El valor medio de las diferencias debe ser cercano a cero.
2. La varianza normalizada de las diferencias debe ser próxima a la unidad.

Además, se realiza un gráfico de valores reales contra estimados, el cual debe ser cercano a una línea recta, lo que se interpreta como que las diferencias $(Z - Z^*)$ son cero o muy cercanas a cero.

Todos los procedimientos y herramientas estadísticas que se utilizaron con anterioridad para analizar el conjunto de datos, es decir, el resumen de estadísticas, el qq plot, el histograma, etc., se pueden utilizar ahora para comparar el conjunto de los valores estimados y los valores verdaderos que se generaron con la validación cruzada, es decir, se realiza un *análisis de residuales.*

En la práctica los residuos son más representativos sólo de algunas regiones o de un rango en particular de valores. Por ello, algunas de las conclusiones de la validación cruzada de residuos pueden ser aplicables a toda el área y otras no. Se prefiere que los estimados sean condicionalmente insesgados con respecto de cualquier rango de valores, también se prefiere que sean condicionalmente insesgados con respecto de su ubicación. En cualquier región se espera que el centro o la media, la dispersión y la asimetría de los residuales sean todas tan cercanas a cero como sea posible. Un mapa de contorno de los residuales puede revelar las áreas donde los estimados son consistentemente sesgados; mapas del movimiento de ventanas estadísticas pueden ser usadas para mostrar cómo la dispersión de los residuales varía a través del área.

En los residuales se puede observar sobreestimación o subestimación, se puede realizar un gráfico y marcar con un símbolo (+) para sobreestimación y uno de (−) para subestimación, también se le agrega sombreado para mostrar la magnitud del residual, y los símbolos más obscuros corresponden a los errores más grandes. En este tipo de gráfico se espera observar una buena mezcla de símbolos, es decir, sin regiones obvias de sobreestimación o de subestimación. Si se encuentran estas regiones, entonces se debe buscar la razón del sesgo.

2.4. Kriging

El Kriging[14] es una técnica de estimación local que ofrece el mejor estimador lineal insesgado de una característica desconocida que se estudia. La limitación a la clase de estimadores lineales es bastante natural ya que esto significa que solamente se requiere el conocimiento del momento de segundo orden de la función aleatoria (la covarianza o el variograma) y que en general en la práctica es posible inferir a partir de una realización de la misma.

Sea $Z(x)$ una función aleatoria, la cual está definida en un soporte puntual y es estacionaria de segundo orden, con:

1. Valor esperado
$$E\left[Z(x)\right] = m, \forall x \tag{2.44}$$
Donde m es una constante generalmente desconocida

2. Una función de covarianza centrada
$$C(h) = E\left[Z(x+h)Z(x)\right] - m^2 \tag{2.45}$$

3. Un variograma
$$Var\left[Z(x+h) - Z(x)\right] = 2\gamma(h) \tag{2.46}$$

Donde al menos uno de estos dos momentos de segundo orden se supone conocido. Y cuando solamente existe el variograma, entonces la función aleatoria $Z(x)$ se considera intrínseca.

[14]Kriging, Ecuaciones del Kriging ordinario y Clasificación de Kriging en notas Dr. Martín Díaz Viera, p.31, p.33 y p.37. Ref. Bibliográfica [2]

El estimador lineal Z_k^* considerado es una combinación lineal de n valores de datos tal que:

$$Z^*(x_k) = \sum_{i=1}^{n} \lambda_i Z(x_i) \tag{2.47}$$

Donde $Z_k^* = Z^*(x_k)$.

Los n coeficientes λ_i son calculados de manera tal que el estimador sea insesgado y que la varianza de la estimación sea mínima, entonces se dice que el estimador Z_k^* es óptimo.

2.4.1. Clasificación de Kriging

Los diferentes tipos de Kriging en base a la forma del estimador lineal se pueden clasificar en:

- Simple
- Ordinario
- Universal
- Residual

Por lo tanto, se resumen las ecuaciones de los tipos de Kriging en base a su estimador lineal de la siguiente manera:

Kriging Simple: kriging lineal con valores esperados conocidos.

Sistema de ecuaciones:

$$\left\{ \begin{array}{l} \sum_{j=1}^{n} \lambda_j \sigma_{ij}^{'}, i = 1, \ldots, n \\ \lambda_0 = m(x_0) - \sum_{i=1}^{n} \lambda_i m(x_i) \end{array} \right\} \tag{2.48}$$

Estimador:

$$Z_0^* = \lambda_0 + \sum_{i=1}^{n} \lambda_i Z(x_i) \tag{2.49}$$

Varianza de la estimación:

$$\sigma_{K_S}^2 = \sigma_{00}^{'} - \sum_{i=1}^{n} \lambda_i \sigma_{i0}^{'} \tag{2.50}$$

Donde,
$m(x_i) = E\left[(x_i)\right]$ es el valor esperado en el punto x_i.
$\sigma_{ij}^{'} = \sigma_{ij} - m(x_i)m(x_j)$ es la covarianza centrada.

Requisitos:

- Conocer $n+1$ valores esperados $m(x_i) = E\left[(x_i)\right], \forall i = 0, \ldots, n$ de la función aleatoria $Z(x)$.
- Conocer la función de covarianzas σ_{ij} de la función aleatoria $Z(x)$.

Kriging Ordinario: Kriging lineal con valor esperado estacionario pero desconocido.

Sistema de ecuaciones:

$$\left\{ \begin{array}{l} \sum_{j=1}^{n} \lambda_j \sigma_{ij}^{'} - \mu = \sigma_{0i}, i = 1, \ldots, n \\ \sum_{i=1}^{n} \lambda_i = 1 \end{array} \right\} \tag{2.51}$$

Estimador:

$$Z_0^* = \sum_{i=1}^{n} \lambda_i Z(x_i) \tag{2.52}$$

Varianza de la estimación:

$$\sigma_{K_0}^2 = \sigma_{00} - \sum_{i=1}^{n} \lambda_i \sigma_{i0} + \mu \tag{2.53}$$

Requisitos:

- Se requiere que el valor esperado $m(x_i) = E\left[(x_i)\right], \forall i = 1, \ldots, n$ de la función aleatoria $Z(x)$ sea constante.
- Conocer la función de covarianzas σ_{ij} o el semivariograma γ_{ij} de la función aleatoria $Z(x)$.

Kriging Universal: Kriging lineal en presencia de tendencia.

Sistema de ecuaciones:

$$\left\{ \begin{array}{l} \sum_{j=1}^{n} \lambda_j \sigma_{ij} - \sum_{l=1}^{L} \mu_l \phi_l(x_i) = \sigma_{0i}, i = 1, \ldots, n \\ \sum_{i=1}^{n} \lambda_i \phi_l(x_i) = \phi_l(x_0), l = 1, \ldots, L \end{array} \right\} \tag{2.54}$$

Estimador:

$$Z_0^* = \sum_{i=1}^{n} \lambda_i Z(x_i) \tag{2.55}$$

Varianza de la estimación:

$$\sigma_{K_U}^2 = \sigma_{00} - \sum_{i=1}^{n} \lambda_i \sigma_{i0} + \sum_{l=1}^{L} \mu_l \phi_l(x_0) \tag{2.56}$$

Requisitos:

- Conocer la forma de tendencia $m(x) = E\left[Z(x)\right]$, de la función aleatoria $Z(x)$ expresada mediante funciones conocidas $\phi_l(x)$, usualmente polinomios.
- Conocer la función de covarianzas σ_{ij} o el semivariograma γ_{ij} de la función aleatoria $Z(x)$ sin tendencia, es decir $[Z(x) - m(x)]$.

Kriging Residual: es análogo al Kriging Universal, ya que $Z(x)$ es no estacionaria, pero con la variante de que el orden de la tendencia $m(x)$ es conocida o se puede estimar, se utiliza la información de los residuos y posteriormente se aplica kriging ordinario.

2.4.2. Particularidades del kriging ordinario

En particular, el *kriging ordinario* es el tipo de kriging más común. Para la práctica se supone:

- Estacionariedad de segundo grado o la hipótesis intrínseca.
- Se necesitan suficientes observaciones para estimar el variograma.
- Los pesos del kriging ordinario cumplen con la condición de insesgamiento.
- Al igual que otros estimadores, el estimador del kriging ordinario es insesgado porque busca que la media de los residuos o el error, sea igual a cero.
- Este kriging es el mejor porque busca minimizar la varianza de los residuales o de los errores.

Lo que distingue al kriging ordinario[15] es que su procedimiento apunta a minimizar la varianza del error. Tanto la media de los residuos como la varianza son ambas desconocidas, por lo tanto, en situaciones prácticas no se puede garantizar que la media sea cero o que se pueda minimizar la varianza. Lo que se puede hacer es construir un modelo con los datos que se están estudiando y trabajar con el promedio del error y con la varianza del error para el modelo.

El Kriging ordinario utiliza un modelo probabilístico, en el cual el sesgo y la varianza del error pueden ser calculados y con ello se pondera sobre las muestras cercanas de tal forma que se asegura que el promedio del error del modelo sea exactamente cero, y que la varianza del error sea mínima.

Para derivar las ecuaciones del kriging ordinario se verifican las siguientes condiciones:

- *La condición de insesgamiento*

Para obtener un valor esperado del error igual a cero resulta suficiente imponer la condición:

$$E\left[Z_k^*\right] = E\left[\sum_{i=1}^{n} \lambda_i Z(x_i)\right] = m \tag{2.57}$$

Como m es el valor esperado de la función aleatoria $Z(x)$,

$$\sum_{i=1}^{n} \lambda_i E\left[Z(x_i)\right] = E\left[Z_k^*\right] \tag{2.58}$$

Y entonces,

$$\sum_{i=1}^{n} \lambda_i m = m \tag{2.59}$$

Y por lo tanto es suficiente con que:

$$\sum_{i=1}^{n} \lambda_i = 1 \tag{2.60}$$

[15] Para mayor información en el desarrollo y derivación de las ecuaciones de kriging ordinario, revisar p.278, *An Introduction to Applied Geostatistics*, Ref. Bibliográfica [1]

- *La condición de mínima varianza en la estimación*

Para satisfacer esta condición hay que minimizar la siguiente función:

$$F = \sigma_e^2 - 2\mu \left[\sum_{i=1}^{n} \lambda_i - 1 \right] \tag{2.61}$$

Donde,
σ_e^2 es la varianza de la estimación.
μ es un multiplicador de Lagrange.

Se remarca que la función F a minimizar, consta de la varianza de la estimación σ_e^2 e incluye la condición de insesgamiento de la estimación.

La varianza de la estimación se expresa de la siguiente manera:

$$\sigma_e^2 = Var\left[Z_k - Z_k^*\right] = E\left[(Z_k - Z_k^*)^2\right] \tag{2.62}$$

$$\sigma_e^2 = Var\left[Z_k\right] - 2Cov\left[Z_k, Z_k^*\right] + Var\left[Z_k^*\right] \tag{2.63}$$

Sustituyendo en esta última fórmula la expresión del estimador Z_k^*, se tiene:

$$\sigma_e^2 = Var\left[Z_k\right] - 2Cov\left[Z_k, \sum_{i=1}^{n} \lambda_i Z(x_i)\right] + Var\left[\sum_{i=1}^{n} \lambda_i Z(x_i)\right] \tag{2.64}$$

Desarrollando se obtiene:

$$\sigma_e^2 = \sigma_{Z_k}^2 - 2\sum_{i=1}^{n} \lambda_i \sigma_{Z_k Z_i} + \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j \sigma_{Z_i Z_j} \tag{2.65}$$

Si se encuentran las derivadas parciales de F respecto a los coeficientes desconocidos λ_i y con respecto a μ obtenemos el siguiente sistema de ecuaciones:

$$\left\{ \begin{array}{l} \frac{\delta F}{\delta \lambda_i} = -2\sigma_{Z_k Z_i} + 2\sum_{j=1}^{n} \lambda_j \sigma_{Z_i Z_j} - 2\mu = 0 \\ \frac{\delta F}{\delta \lambda_i} = \sum_{i=1}^{n} \lambda_i - 1 = 0 \end{array} \right\} \tag{2.66}$$

De una manera común se escribe:

$$\left\{ \begin{array}{l} \sum_{j=1}^{n} \lambda_j \sigma_{Z_i Z_j} - \mu = \sigma_{Z_k Z_i}, i = 1, \ldots, n \\ \sum_{i=1}^{n} \lambda_i = 1 \end{array} \right\} \tag{2.67}$$

La decisión del modelo de covarianza (o en todo caso, el modelo de variograma o correlograma) es un requisito para el kriging ordinario. Esto aporta una flexibilidad muy importante para personalizar el método de kriging ordinario.

En la práctica, el patrón de continuidad espacial elegido del modelo de función aleatoria es usualmente obtenido de las evidencias de continuidad espacial del conjunto de datos de la muestra.

Una vez que el variograma muestral ha sido calculado, se obtiene una función de él. Existen dos razones para que el variograma muestral no pueda usarse directamente en el kriging ordinario. Frecuentemente existen situaciones en las

que la distancia del punto a estimar a algún punto de la muestra es menor que la distancia entre cualquier par de valores disponibles de la muestra. Como la muestra no proporciona ningún par para estas distancias, se recae sobre una función que proporciona valores del variograma para todas las distancias y direcciones, incluso para aquellas que no fueran disponibles de la muestra. Segundo, el uso del variograma de muestra no garantiza la existencia y unicidad de la solución del sistema del kriging ordinario. Para garantizar que exista una y sólo una solución, se debe asegurar que en el sistema del kriging las matrices sean positivas definidas.

Debido a que el uso del variograma de muestra no garantiza que el sistema sea positivo definido, en la práctica se garantiza ajustando funciones que se conoce que son positivas definidas al variograma de muestra. En las muestras de ciencias de la tierra casi siempre existe un patrón de continuidad espacial aunque puede no ser evidente debido a un número insuficiente de datos, error al muestrear, valores erráticos, o posibles valores extremos.

La anisotropía es un elemento importante del patrón de continuidad espacial del modelo de función aleatoria. El uso de la tolerancia en la dirección puede causar que la evidencia de anisotropía en los variogramas de muestra sea más débil que en la que la información disponible es exhaustiva. Los variogramas de muestra y las funciones de covarianza muestran menos anisotropía que las muestras exhaustivas.

Aunque ajustar funciones a los variogramas de muestra es seguramente el enfoque más común para escoger el patrón de continuidad espacial para el modelo de función aleatoria, no debe ser visto como el único enfoque correcto. En todos los estudios que utilizan métodos de estimación geoestadísticos, los geoestadistas deben elegir el patrón de continuidad espacial.

El éxito del kriging ordinario sobre otros métodos de estimación es debido a que utiliza una distancia estadística personalizada en vez de una distancia geométrica y permite desagrupar la información disponible. El uso de un modelo de continuidad espacial que describe la distancia estadística entre puntos proporciona flexibilidad y una habilidad importante para personalizar el proceso de estimación hacia información cualitativa.

2.4.3. Dependencia de la estimación con kriging del modelo de variograma

Aún con una distribución homogénea de los datos, el ajustar una función al modelo de variograma implica tomar decisiones importantes. El variograma muestral no proporciona menos información que el espacio mínimo entre datos de la muestra. El efecto pepita y el comportamiento del variograma cerca del origen no pueden ser determinados por el variograma muestral. Sin embargo, los siguientes parámetros tienen el mayor efecto en los pesos del kriging ordinario y en el resultado del estimado.

1. *El efecto de Escala:*

   Si se tienen dos modelos de variograma donde la única diferencia entre ellos es la escala, es decir, uno es el doble del otro, entonces en los resultados se mostraría que no se afectaron ni los pesos del kriging ordinario ni el estimador del kriging ordinario, sin embargo, se afecta la varianza del kriging ordinario. Este efecto ocurre en cualquier tipo de reescalamiento, mientras que el estimador es el mismo, la varianza se incrementa con el mismo factor que fue usado en la escala del variograma.

2. *El efecto de Forma:*

   Se tienen dos variogramas con el mismo umbral (sill) pero con diferentes formas, ya que el segundo es el cuadrado del primero, entonces tiene un efecto parábola cerca del origen. En los resultados se mostraría que el segundo modelo aporta más peso a los tres valores que rodean el punto a estimar y los puntos restantes reciben menos peso, incluso pueden recibir peso negativo. Un comportamiento parabólico cerca del origen es indicativo de un fenómeno continuo así que el procedimiento de estimación utiliza mucho más las muestras cercanas.

3. *El efecto del Rango:*

   Si se tienen dos modelos de variogramas que difieren sólo en el rango, el segundo tiene el doble que el primero, entonces el cambio en el rango tiene un efecto relativamente menor en los pesos del kriging ordinario, aunque los cambios en los pesos son muy pequeños, los cambios en el estimador son notables.

   La varianza del kriging ordinario es menor debido a que el efecto de duplicar el rango hace que parezca que las muestras están el doble de cerca, en términos de distancia estadística. Si el rango se vuelve muy pequeño, entonces todas las muestras parecen igual de lejos del punto a estimar y de una a la otra, teniendo un resultado similar al que se obtiene con el modelo del efecto pepita puro (*pure nugget effect*).

4. *Influencia de la anisotropía en la estimación*

   La relación de anisotropía juega un papel importante. Los pesos del kriging ordinario son calculados en base a este modelo. La redistribución de los pesos es en base de la dirección de máxima continuidad, por eso, aunque una muestra se encuentre muy lejana como distancia geométrica puede tener mucho peso si cae dentro de esta dirección de máxima continuidad; si se escoge en una relación alta de anisotropía, se convierte en una de las muestras más influyentes.

   La posibilidad de escoger fuertes patrones de anisotropía para continuidad espacial en el modelo de función aleatoria permite personalizar el procedimiento de estimación. Por ejemplo, un geoestadista puede utilizar información cualitativa como la interpretación geológica de un depósito de minerales para escoger los ejes de anisotropía.

En muchos conjuntos de datos, la dirección de máxima continuidad no es únicamente en el área de interés, puede haber fluctuaciones locales en la dirección y el grado de anisotropía. En estas situaciones, la muestra de variogramas puede parecer sólo isotrópica porque no se puede clasificar el carácter ondulante de la anisotropía. Si la información cualitativa ofreciera una forma de identificar la dirección y el grado de anisotropía, entonces el método de estimación se beneficia enormemente de la decisión de basar el modelo de continuidad espacial en evidencia cualitativa en vez de en evidencia cuantitativa del variograma muestral.

Para métodos de estimación que pueden manejar cualquier número de muestras cercanas, el enfoque más común para decidir las muestras que contribuyen a la estimación es definir una vecindad de búsqueda en la que todas las muestras disponibles sean usadas. La vecindad de búsqueda usualmente es un elipse centrado en el punto a estimar. La orientación del elipse es definido por la anisotropía en el patrón de continuidad espacial.

Una vez que se decide una orientación y una relación de anisotropía para el elipse de búsqueda, todavía falta decidir qué tan grande será la vecindad. Debe ser suficientemente grande para que incluya muestras, pero esto depende de los datos. Si los datos están en una cuadrícula semiregular, se puede calcular qué tan grande debe ser el elipse para que se incluyan por lo menos las cuatro muestras más cercanas. En la práctica se busca que se incluyan por lo menos 12 muestras. Para datos de cuadrículas irregulares, la vecindad de búsqueda debe ser más grande que el promedio del espacio entre los datos de la muestra, que se pueden calcular como:

$$\textit{Promedio de espacio entre datos} \approx \sqrt{\frac{\textit{Área total cubierta por muestras}}{\textit{Número de muestras}}} \tag{2.68}$$

2.4.4. Gráfico de resultados

Una vez que se utilizó el análisis exploratorio de datos, el análisis estructural, la información cualitativa y los conocimientos como geoestadístico y se realiza la estimación por medio del Kriging, se necesitan visualizar los resultados.

Comúnmente, es más fácil identificar la información de los resultados con un gráfico en lugar de sólo utilizar los números, por lo que en la práctica se obtienen diversos gráficos que ayudan a la interpretación de los resultados como lo son:

- El gráfico de valores estimados contra residuales el cual permite verificar insesgamiento[16]
- El gráfico de estimación, el cual es un mapa que contiene la información de las estimaciones a color para una mejor apreciación.

[16] *"One way of checking for conditional bias is to plot the errors as function of the estimated values"*, Isaaks, p.264, 1989. Referencia Bibliográfica [1]

- El gráfico de errores permite verificar la información de los errores de la estimación visualmente.

2.5. Caso ideal

El caso ideal es el caso en el cual se cumplen todos los supuestos en los cuales está basada la metodología de estimación con Kriging. Además, es el caso en el cual el proceso de estimación resulta lo más sencillo y confiable posible.

2.5.1. Supuestos

Se habla de los supuestos del caso ideal cuando se contienen las siguientes características sobre la muestra:

1. Los datos se muestran completamente simétricos y tienen una distribución normal.
2. Media constante.
3. Homocedasticidad (varianza constante)
4. Al realizar el histograma y el gráfico de la distribución no se deben observar datos atípicos tanto espaciales como distribucionales.
5. Los datos no muestran ningún tipo de tendencia y la muestra es estacionaria (hipótesis intrínseca).
6. La matriz de covarianzas debe ser positiva definida.
7. El número de observaciones debe ser suficiente para cubrir el área total.
8. El tipo de muestreo debe ser una malla regular que contenga los límites de la región a estimar.
9. Los datos no deben presentar características influyentes para el caso ideal, sin embargo, si llegaran a presentarlas deberán ser visibles notablemente y su tratamiento deberá ser lo más sencillo posible.

Capítulo 3

Descripción de las características de los casos de estudio

Como ya se había mencionado en la introducción, el objetivo principal de esta tesis es revisar el análisis exploratorio de datos con el contexto de analizar la metodología y concentrar los conocimientos, así como ofrecer una guía metodológica sobre la aplicación práctica de cómo realizar el análisis geoestadístico para obtener resultados óptimos sobre todo cuando los datos se alejan del caso ideal. En el capítulo sobre la metodología de geoestadística, se mostró que cuando los datos no satisfacen los supuestos sobre los cuales fueron diseñados los estimadores (caso ideal), estos se degradan y pueden dar resultados completamente erróneos.

Es bien conocido cuales son las características de los datos que afectan al buen desempeño de un análisis geoestadístico. Entre otras, se pueden mencionar las desviaciones en el muestreo cuando este no es suficientemente homogéneo dentro del área de estudio, la cantidad de datos que pudiera no ser suficiente, la presencia de tendencia, anisotropía, etc. Con el propósito de investigar el impacto de estas características en el análisis geoestadístico se propuso generar muestras artificiales que fueran representativas de alguna de éstas. Como son muchas, para ser precisos, se escogieron los siguientes casos de estudio:

1. Tendencia
2. Estructura espacial anidada (modelo gaussiano con exponencial)
3. Anisotropía

Para darle un matiz realista a las muestras artificiales, tres simulaciones geoestadísticas fueron generadas a partir de datos reales de pluviómetros que corresponden a los casos arriba mencionados. La generación de las simulaciones geoestadísticas fue realizada aplicando el método Secuencial Gaussiano[1].

[1]Las simulaciones fueron generadas por el M. en C. Javier Méndez Venegas con información relacionada a su Tesis: *Modelación de la distribución espacial de la precipitación en el valle de la ciudad de México usando técnicas geoestadísticas.* Ref. Bibliográfica [25]

Los datos de niveles de lluvia acumulada utilizados en las simulaciones provienen de la red de pluviómetros del Sistema de Aguas del Departamento del Distrito Federal[2] y son manejados por la Comisión Nacional del Agua. Los pluviómetros tienen una densidad del orden de un pluviómetro por cada $30km^2$ (Figura 3.1). La red de pluviómetros se encuentra instalada en el área metropolitana de la Ciudad de México y reporta vía radio cada minuto la información de lluvia acumulada en ese intervalo a una computadora central.

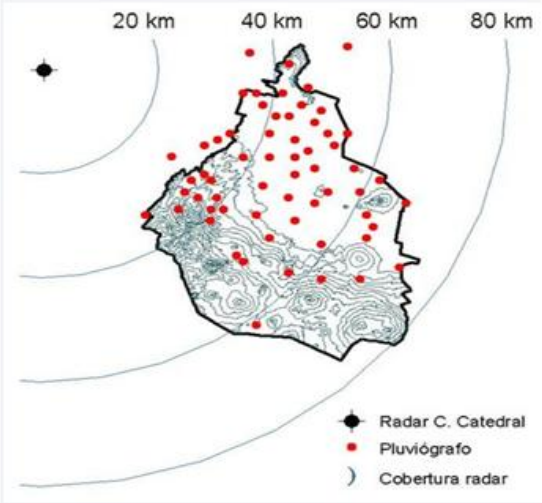

Figura 3.1: Mapa de pluviógrafos

El proceso de la simulación geoestadística de estas muestras artificiales no está dentro de los objetivos de esta tesis y por lo tanto, en lo sucesivo se considerarán simplemente como datos del tipo 1 para la simulación representando tendencia, datos del tipo 2 para la simulación representando una estructura anidada de modelo gauss-exponencial y datos del tipo 3 para la simulación representando anisotropía.

Los datos del tipo 1 representan uno de los tres casos de estudio a tratar, así como de igual manera los datos del tipo 2 representan el segundo caso de estudio y los datos del tipo 3 representan el tercer caso de estudio.

Los datos de las muestras están georeferenciados con respecto a un sistema de coordenadas UTM (Universal Transversal de Mercator) en una malla regular con una densidad de datos de 37x40, es decir, se tienen observaciones espaciadas a 1km por 1km, resultando un total de 1480 datos. Cada una de las bases de datos del tipo 1, 2 y 3 es una matriz de 1480x3, donde la primera columna se refiere a las coordenadas sobre el eje X con rango de valores de 469 a 505, la segunda columna se refiere a las coordenadas sobre el eje Y con rango de valores del 2121 al 2160 y la tercera columna se refiere al valor de la variable en la

[2]Información de la Tesis del M. en C. Javier Méndez Venegas: *Modelación de la distribución espacial de la precipitación en el valle de la ciudad de México usando técnicas geoestadísticas*. Ref. Bibliográfica [25]

ubicación referida por la pareja de coordenadas.

Las tres muestras de datos generadas serán tomadas como de referencia, es decir, por contar con una alta densidad de puntos en una malla regular son consideradas cercanas al caso de una muestra ideal.

Adicionalmente, con el fin de explorar la influencia del tipo de muestreo y de la cantidad de datos en el análisis geoestadístico, se proponen tres tipos de muestreo (regular, aleatorio y combinado) y tres tamaños de muestra (36, 100 y 400), respectivamente.

La combinación de los tres casos de estudios, los tres tipos de muestreo y tres tamaños de muestra, dan origen a un total de 27 escenarios.

Los casos de estudio se dividen jerárquicamente en base a la característica distintiva que presenta cada muestra de datos, el tipo de muestreo y el número de observaciones que se obtienen, tomando en cuenta siempre que cada escenario se realiza de manera independiente.

Para fines prácticos, se renombrará a cada uno de los escenarios con una clave que utiliza las iniciales del tipo de datos, el tipo de muestreo y el número de observaciones que contienen. Entonces, en base al tipo de datos se realiza la siguiente distinción:

1. Si los datos del escenario se obtienen de la simulación 1, se le nombrará como datos 1 ó D1
2. Si los datos del escenario se obtienen de la simulación 2, se le nombrará como datos 2 ó D2
3. Si los datos del escenario se obtienen de la simulación 3, se le nombrará como datos 3 ó D3

Así mismo, en base al tipo de muestreo se le asigna lo siguiente:

1. Si los datos del escenario son generados como muestreo de malla regular se le asignará MR
2. Si los datos del escenario son generados como muestreo aleatorio se le asignará MA
3. Si los datos del escenario son generados como muestreo combinado se le asignará MC

Por último, para el número de observaciones se asigna como sigue:

1. Si los datos del escenario tienen 36 observaciones, entonces se le asignará c036
2. Si los datos del escenario tienen 100 observaciones, entonces se le asignará c100
3. Si los datos del escenario tienen 400 observaciones, entonces se le asignará c400

En resumen, se tiene una clave que describe el escenario del cual se está hablando. Por ejemplo, si se habla del escenario de los datos 1 con muestreo de malla regular y 36 observaciones la clave será: "D1MRc036" y así sucesivamente para los 27 escenarios.

3.1. Características de la base de Datos 1

Los datos del tipo 1 son datos que contienen tendencia. La tendencia es un patrón de comportamiento de los elementos de un entorno particular durante un período, es decir, en este caso es la dirección que toman los valores respecto a cada una de las coordenadas y la importancia radica en que si existe tendencia entonces no se cumple con la hipótesis intrínseca, ya que no se tiene media constante lo cual indica que no cumple con los supuestos de la metodología mencionada.

La característica de tendencia varía debido a que el proceso para tratarla utiliza una transformación con un polinomio del cual el grado no es conocido, por lo cual se inicia con una transformación a un modelo de 1er grado y se verifica si es correcto. Si no lo es, se continúa con un polinomio de 2do grado y así sucesivamente hasta que el grado sea encontrado y los residuales sean estacionarios. Es de gran relevancia mencionar que este proceso es utilizado únicamente si la tendencia es influyente y muy significativa, ya que es posible que la tendencia sea mínima y no influya realmente sobre la estacionariedad, con lo cual no se estaría violando ningún supuesto sobre el cual están basados los estimadores y por lo tanto podría procesarse la información tal cual es recabada.

Es muy importante tratar la tendencia de manera adecuada, en principio porque esta característica no cumple con uno de los principales supuestos de la metodología el cual requiere de estacionariedad. Al no cumplir con el supuesto, el modelo de variograma elegido se vería afectado y la variabilidad en los resultados sería más alta de lo deseado, lo cual ocasionaría una menor confiabilidad de la necesaria. Sin embargo, si la tendencia es identificada y tratada adecuadamente, es posible llegar a una estimación confiable y exitosa.

Los escenarios presentados en el caso de estudio de los datos del tipo 1, resultan muy interesantes debido a que las diferencias en el tipo de muestreo o tamaño de muestra permitirán o no observar la característica mencionada y resaltar la diferencia en el proceso de estimación cuando se cuenta con suficiente información significativa. En particular, se mostrará el proceso completo de estimación para los escenarios de D1MRc100, D1MAc100 y D1MCc100 los cuales son los más representativos debido a que tienen un número suficiente de observaciones. En estos escenarios será posible destacar los detalles en cuanto a gráficos o estadísticos con los cuales se puede identificar si existe o no tendencia, así como el resultado gráfico de la estimación.

Así mismo, son conocidas las afectaciones que genera el tener una base de datos con información insuficiente o que no es significativa, por lo que para los escenarios con 36 observaciones se esperan varias complicaciones en el ajuste del modelo de variograma debido a que el proceso muestra que si no se cuenta con

suficiente información el variograma será calculado con poca confiabilidad y por lo tanto el modelo puede no ser el óptimo. También, la distribución de los datos no se observaría fácilmente, lo cual puede ocasionar que algunas características de la muestra no sean visibles. Además, los diferentes tipos de muestreos también influyen directamente sobre la distribución de los datos y podrían complicar o facilitar el proceso. Sin embargo, se espera que conforme aumente el número de observaciones dentro de los escenarios presentados se vuelva más fácil identificar las características de la muestra y facilitar el ajuste del modelo de variograma.

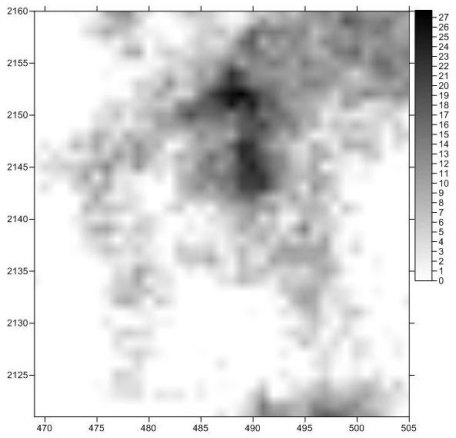

Figura 3.2: Base de Datos 1

3.2. Características de la base de Datos 2

Los datos del tipo 2 son datos que introducen una característica que influye de manera significativa en el ajuste del modelo, es decir, los datos del tipo 2 tienen una estructura espacial anidada la cual es un modelo Gaussiano con exponencial. Recordando de la metodología de geoestadística de este trabajo se menciona que la combinación lineal de semivariogramas con coeficientes positivos representa un semivariograma; es decir, la combinación lineal o el producto de semivariogramas también es un semivariograma y entonces es posible utilizar modelos combinados creados a partir de modelos ya conocidos, lo cual describe brevemente la estructura anidada[3]. En particular para el caso de los datos del tipo 2, la característica de la estructura anidada representa un modelo de variograma el cual en una parte tiene comportamiento de un modelo de variograma gaussiano y en otra parte tiene comportamiento de modelo de variograma exponencial.

Es interesante mostrar este tipo de datos ya que en la realidad las bases de datos no siempre se apegan a una distribución conocida y por consiguiente el variograma presentado puede no apegarse a una sola estructura de modelo de variograma conocida. Por lo tanto, es importante destacar que el variograma puede ser ajustado no sólo con un modelo conocido sino también con una com-

[3]Notas del Dr. Martín Díaz Viera, p.24. Ref. Bibliográfica [2]

binación de modelos conocidos.

Es importante tratar una estructura anidada ya que permitiría realizar una estimación óptima. Sin embargo, cabe destacar que dependiendo de qué modelo es el predominante podría o no observarse la influencia de esta característica en particular o incluso podría utilizarse otro modelo que se asemeje al variograma presentado y que siga siendo un modelo adecuado bajo las pruebas de residuales.

El variograma es el medio con el cual se puede identificar esta característica. Debido a ello no es de esperarse que una estructura anidada sea fácilmente observable dentro de los escenarios con 36 observaciones ya que no cuenta con suficiente información para calcular el variograma. Incluso aunque se tenga un mayor número de observaciones en los demás escenarios podría no ser visible dentro de los gráficos, estadísticas o el mismo variograma.

Los detalles más relevantes sobre los escenarios presentados dentro del caso de estudio de los datos del tipo 2, se refieren a mostrar el procedimiento de la metodología y cumplimiento de los supuestos o en su defecto el posible tratamiento sobre la base de datos para que los cumplan, así como el ajuste del modelo presentado y la variación con la cual este modelo es elegido en base a los diferentes tipos de muestreo, número de observaciones que contienen y la significancia en la información disponible.

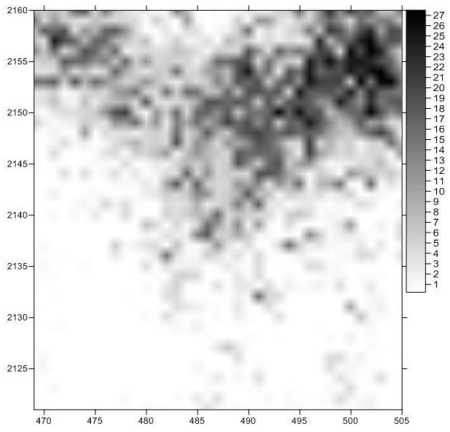

Figura 3.3: Base de Datos 2

3.3. Características de la base de Datos 3

Los datos del tipo 3 contienen anisotropía geométrica. Recordando de la metodología de geoestadística, la anisotropía es la propiedad general con la cual determinadas propiedades varían según la dirección en que son examinadas, es decir, algo anisótropo podrá presentar diferentes características según la dirección. En presencia de anisotropía geométrica, el gráfico direccional de los rangos forma una elipse, en el cual el eje menor B es el rango de la dirección de más rápida variación y A, el eje mayor, está en la dirección en que la variabilidad es

más lenta. La relación $\lambda = A/B$ es una medida de anisotropía. Si se desea diseñar una red óptima de muestreo, se debe hacer en forma rectangular haciendo coincidir los lados con las direcciones de los ejes principales y las longitudes de los mismos deben estar en la proporción λ, donde el lado menor le correspondería a la dirección del eje B.

Por lo tanto, los datos del tipo 3 introducen una característica muy importante e influyente para el ajuste del modelo porque depende de la dirección. En este caso de estudio se deberán introducir los modelos anisotrópicos, los cuales son modelos de variograma que se basan en la dirección del eje de mayor alcance para realizar la estimación. Es de gran importancia tratar esta característica adecuada y cuidadosamente porque destaca la significancia de las muestras cercanas a la zona que se desea estimar, y reasigna esta importancia en la dirección elegida.

Este caso de estudio es muy interesante debido a que es una característica que se encuentra fácilmente en la realidad. Incluso, en ocasiones, el ajuste del modelo de variograma y la elección de la dirección de mayor alcance puede ser mejorada a partir de información cualitativa ya sea climatológica, meteorológica, etc. dependiendo del estudio que se esté realizando. No obstante, la anisotropía es una característica difícil de identificar y a menudo es confundida con no estacionariedad, tendencia, asimetría u otras características que afectan la valoración de los datos.

Los escenarios ponen en evidencia la importancia de la cantidad de información con la que se cuenta (número de observaciones) y la significancia que transmite cada una de estas observaciones dependiendo de su ubicación (tipo de muestreo) para identificar adecuadamente la distribución de la muestra y características influyentes para la estimación. También es importante destacar si existe una diferencia considerable sobre la estimación cuando la dirección de mayor alcance no es la adecuada.

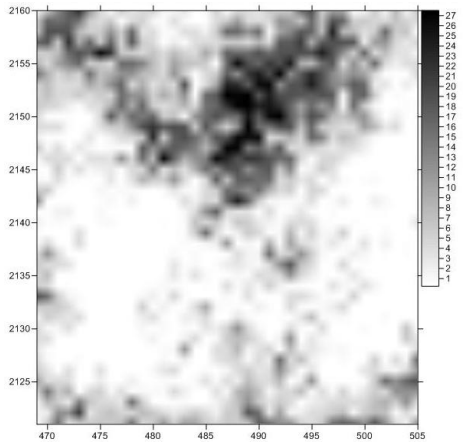

Figura 3.4: Base de Datos 3

3.4. Muestreo de malla regular

Como es conocido, normalmente los datos obtenidos como malla regular permiten observar mejor las características de la muestra. Sin embargo, es poco común en la práctica ya que la toma de muestras puede ser afectada por condiciones climatológicas, geográficas, o incluso depende del equipo, personal y tiempo disponible para realizarla. En particular, sería de esperarse que destacara la relevancia de este tipo de muestreo principalmente con bases de datos que contienen características influyentes las cuales para su detección dependen de la ubicación de las muestras.

El muestreo de malla regular se realiza mediante una submalla generada a partir de la subdivisión de las coordenadas del eje X y el eje Y, de tal manera que las observaciones sean equidistantes en su propio eje. Para los casos de 400 observaciones la base de datos no permite seleccionar suficientes muestras de manera equidistante, por lo que se realizó un ajuste con las primeras y últimas observaciones de la mitad del espaciamiento inicial.

Debido a que las 3 bases de datos originales del tipo 1, 2 y 3 tienen la misma dimensión, el muestreo de malla regular se realiza de manera análoga para los 3 tipos de base de datos.

A continuación, se detalla la obtención de las bases de datos de los escenarios referentes al muestreo de malla regular dependiendo del número de observaciones:

Para los escenarios con 36 observaciones se generó una malla de 6x6. Es decir, se subdividió el eje X a razón de 6 comenzando en la primera coordenada posible la cual es 469. De igual manera se subdividió el eje Y a razón de 7 comenzando también en la primera coordenada posible la cual es 2121. Con esto se tienen 6 coordenadas sobre el eje X y 6 coordenadas sobre el eje Y lo cual da un total de 36 pares de coordenadas. A cada una de estas parejas se le asigna el valor de la variable en ésa ubicación y con esto se conforma una matriz de 36x3, con lo cual se obtiene una malla regular de 36 observaciones.

Para los escenarios con 100 observaciones se generó una malla de 10x10, lo cual indica que se subdividió el eje X en razón de 4 comenzando en la primera coordenada posible la cual es 469. En el caso del eje Y también se subdividió a razón de 4 y de igual manera se inicia en 2121. Por lo tanto, se obtienen 10 coordenadas para el eje X y 10 coordenadas para el eje Y, con lo que se obtienen 100 pares de coordenadas a las cuales se les asigna el valor de la variable en la ubicación referida y con lo que se conforma una malla de 100 observaciones.

Para los escenarios con 400 observaciones se utiliza una malla de 20x20. Se subdividió el eje X a razón de 2, sin embargo, en este caso no alcanzan las coordenadas del eje X para obtener 20 datos, por lo que para la primera y última coordenada se utiliza la mitad del espaciamiento, es decir 1, y entonces la primer coordenada es 469, la segunda es 470 y a partir de la tercera ya se utiliza un espaciamiento de 2, hasta llegar a 504 que es la penúltima coordenada y se utiliza otra vez el criterio de la mitad del espaciamiento, por lo que la última

coordenada es 505. Con esto se obtienen 20 coordenadas sobre el eje X. Para el caso del eje Y, se cuenta con suficiente información para que el espaciamiento sea a razón de 2, comenzando en la primera coordenada posible que es 2121. De esta manera se obtienen 400 pares de coordenadas a las cuales se les asigna el valor de la variable en la ubicación referida y conforman una malla de 400 observaciones.

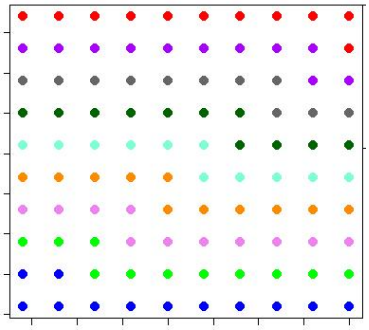

Figura 3.5: Muestreo de malla regular 100 observaciones

3.5. Muestreo aleatorio

El muestreo aleatorio muestra una base de datos más congruente con la realidad. El personal que realiza la toma de muestras normalmente no suele obtener los datos en ubicaciones de manera regular. Esto es debido a que las condiciones climatológicas y geográficas a menudo impiden llegar a ubicaciones específicas. Al realizar los estudios en estas bases de datos es importante utilizar toda la información cualitativa que se tenga a la mano, ya que existen regiones dentro del área de muestreo que pueden ser pobremente muestreadas en comparación con otras regiones del área en cuestión. Al no tener una base de malla regular puede ocasionar que algunas características de los datos se observen más acentuadas como la asimetría y tendencia, o también más escondidas como la anisotropía que depende de la dirección. Por lo tanto, el análisis exploratorio toma mayor importancia ya que permite mostrar las distintas características de los datos y un adecuado tratamiento de la información para llegar a un modelo de estimación más adecuado.

A continuación, se detalla la obtención de las bases de datos de los escenarios referentes al muestreo aleatorio dependiendo del número de observaciones:

Para realizar el muestreo aleatorio se redefine cada una de las tercias con los números del 1 al 1480, es decir, la matriz de 1480x3 se le asigna un vector columna con valores del 1 al 1480, por lo que cada renglón de la matriz inicial que contiene 3 entradas ahora está representado por un solo número. Después, con la ayuda de cualquier software estadístico, Excel o en particular para este caso el programa estadístico R; se obtiene un vector aleatorio del número de datos solicitados n, es decir de 36, 100 o 400 observaciones. Con ello se obtiene un vector de n entradas, el cual cada entrada es un número aleatorio dentro del intervalo [1,1480] sin reemplazo. El vector aleatorio es un subconjunto del

vector representativo. Por último, se le asigna la tercia correspondiente a cada valor del vector para convertir el vector columna aleatorio, en una matriz de nx3, siendo n el número de datos solicitados (36, 100 y 400).

Para generar las bases de datos de los escenarios con 36 observaciones se generó un vector columna de manera aleatoria y sin reemplazo de 36 entradas, cada entrada es un número aleatorio entre [1,1480]. Cada una de estas entradas corresponde a una tercia de la matriz de datos que se está utilizando (tipo 1, tipo 2 o tipo 3) por lo que se le asigna la tercia correspondiente para obtener la matriz resultante de 36x3, en la cual la primera columna muestra 36 coordenadas sobre el eje X, la columna 2 muestra 36 coordenadas sobre el eje Y, y la columna 3 muestra el valor de la variable en cada una de las ubicaciones del plano.

Para los escenarios con 100 observaciones se generó un vector columna de manera aleatoria y sin reemplazo de 100 entradas, cada una de estas entradas corresponde a una tercia de la matriz de datos que se está utilizando. Se le asigna la tercia correspondiente a cada uno de los 100 números aleatorios, y se obtiene como resultado una matriz de 100x3, en la cual cada renglón tiene la información de la coordenada X, la coordenada Y y el valor de la variable en ésa ubicación.

Para los escenarios con 400 observaciones se realiza de igual manera. Se genera un vector columna de manera aleatoria y sin reemplazo de 400 entradas. Luego se asigna la tercia correspondiente del vector que se renombró a cada uno de los 400 números aleatorios y con ello se obtiene la matriz resultante de 400x3, para obtener un muestreo de 400 datos aleatorios con sus respectivas coordenadas.

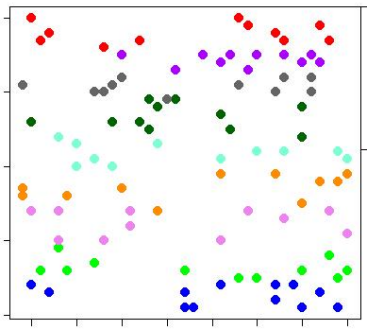

Figura 3.6: Muestreo aleatorio 100 observaciones

3.6. Muestreo combinado

Se le llama muestreo combinado ya que es una combinación del muestreo regular y el muestreo aleatorio. Este tipo de muestreo tiene la aleatoriedad que le aporta congruencia con las bases de datos en la práctica, pero sigue siendo de manera regular lo que permite tener mayor información sobre el área total y por lo tanto identificar características relevantes de cada una de las bases de

datos originales.

Para el muestreo combinado se realiza una subdivisión de los ejes de coordenadas en intervalos conformando una malla regular de cuadrantes y luego se elige aleatoriamente un dato dentro de cada uno de estos cuadrantes. Este muestreo permite obtener una base de datos semiregular que tiene información elegida aleatoriamente sobre cada uno de los cuadrantes, por lo que se tiene información sobre el área total.

A continuación, se detalla la obtención de las bases de datos de los escenarios referentes al muestreo combinado dependiendo del número de observaciones:

Para realizar el muestreo combinado con 36 observaciones, se subdivide el eje X a razón de 6, y el eje Y a razón de 7 igual que en el muestreo regular, sin embargo, los intervalos son abiertos del límite inferior para que no se tenga información repetida. Con esto se obtienen 6 intervalos del eje X y 6 intervalos del eje Y, con un total de 36 cuadrantes sobre el área total. Luego, se obtiene un vector de 6 entradas para cada uno de los intervalos del eje X, cada entrada es un número aleatorio dentro del intervalo que le corresponde. Con esto se obtienen 6 vectores aleatorios para cada intervalo con 6 entradas cada uno. Luego para el eje Y se generan también 6 vectores de 6 entradas cada uno para cada intervalo del eje Y. Ahora, se junta cada una de las entradas del vector del primer intervalo del eje X con la primera entrada de cada uno de los vectores de cada intervalo del eje Y. De esta forma se obtienen las primeras 6 parejas de coordenadas, es decir, se tienen 6 coordenadas aleatorias del primer intervalo de X con 6 coordenadas aleatorias de los 6 intervalos de Y, y así sucesivamente para todos los intervalos de X. Una vez realizado el proceso para todos los intervalos, el resultado son 36 parejas de coordenadas elegidas aleatoriamente sobre cada uno de los cuadrantes de la región. Por último, se le asigna el valor de la variable en la ubicación referida por la pareja de coordenadas, para finalmente obtener una matriz de 36x3, donde la primer columna son las coordenadas sobre el eje X, la segunda columna son las coordenadas sobre el eje Y y la tercera columna son los valores de la variable en cada ubicación del plano; siendo cada una de las tercias de datos elegidas de manera aleatoria sobre intervalos regulares.

Para generar las bases de datos de los escenarios con 100 observaciones se realiza el mismo procedimiento que en el caso de 36 observaciones. En este caso se subdividen el eje X y el eje Y a razón de 4, tomando en cuenta también que cada intervalo debe ser abierto en su límite inferior. De esta manera se tienen 10 intervalos sobre el eje X y 10 intervalos sobre el eje Y, dando un total de 100 cuadrantes sobre la región. Ahora se generan los vectores de 10 entradas para cada uno de los intervalos del eje X y 10 entradas para los intervalos del eje Y. Luego se junta cada una de las entradas del vector del primer intervalo del eje X con la primera entrada de cada uno de los vectores de cada intervalo del eje Y. Se obtienen 100 pares de coordenadas y a cada una se le asigna el valor de la variable en la ubicación referida por las coordenadas para finalmente obtener la matriz de 100x3, donde cada una de las entradas es información obtenida aleatoriamente de cada uno de los cuadrantes de la región dividida de manera regular.

Para las bases de datos de 400 observaciones también se realiza el mismo procedimiento mencionado. Sin embargo, en el caso del muestreo regular la primera y última coordenada sobre el eje X se tiene a un espaciamiento de la mitad, por lo que para el primer y último intervalo del muestreo combinado se realiza lo mismo, es decir, el primer intervalo sólo será la primer coordenada que es 469 y el último intervalo siempre será la última coordenada que es 505; del segundo intervalo al intervalo 19 del eje X todos serán a razón de 2 y abiertos del límite inferior. En el caso del eje Y los intervalos son a razón de 2 y abiertos del límite inferior, de tal forma que se obtienen 20 intervalos sobre el eje X y 20 intervalos sobre el eje Y, conformando 400 cuadrantes sobre toda la región. Luego se generan 20 vectores con 20 entradas para cada intervalo del eje X y 20 vectores con 20 entradas para cada intervalo del eje Y y se junta cada una de las entradas del vector del primer intervalo del eje X con la primera entrada de cada uno de los vectores de cada intervalo del eje Y. Con esto se obtienen 400 pares de coordenadas a las cuales se les asigna el valor de la variable en ésa ubicación del plano. Por lo tanto, se conforma una matriz de 400x3, siendo cada una de las entradas elegidas de manera aleatoria sobre cuadrantes de una malla regular.

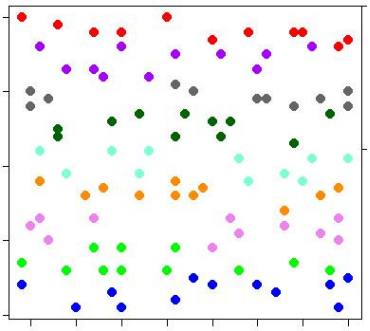

Figura 3.7: Muestreo combinado 100 observaciones

3.7. Bases de datos con 36 observaciones

Cuando se tiene una base de datos con tan pocas observaciones en un área que es proporcionalmente más grande es muy difícil identificar la distribución de los datos, así como características que influyen de manera significativa en el adecuado ajuste del modelo. Características como la anisotropía o tendencia son difíciles de observar por la misma razón de que no se cuenta con información suficiente para hacerlas visibles. Por ejemplo, una característica como la tendencia puede no resultar tan evidente o tan acentuada ya que los valores pueden ser muy extremos en algunas ubicaciones; también puede resultar muy difícil que se observe la anisotropía e incluso puede ser confundida con otra característica de los datos. Por lo tanto el análisis variográfico toma mucha importancia en estos escenarios, ya que es el indicador más común para determinar si existe o no tendencia o anisotropía. Sin embargo, al no tener información suficiente, el variograma es también pobremente estimado y en ocasiones resulta en un ajuste poco confiable. Es decir, el ajuste de un modelo adecuado donde la base de datos

solo contiene 36 observaciones se vuelve más complejo y menos confiable. En la práctica es común encontrarse algunas variables de estudio con estas bases de datos, ya que en ocasiones el equipo de medición es muy costoso, el personal es insuficiente o el periodo de tiempo para la toma de muestra es muy largo, por lo que resultan en bases con muy poca información.

3.8. Bases de datos con 100 observaciones

Para los casos que tienen 100 observaciones es importante remarcar que las observaciones son suficientes. Una base de datos con una cantidad suficiente de información permite observar las características y distribución de los datos y permite realizar un adecuado análisis exploratorio de datos. Al tener mayor número de datos, el variograma adireccional y los variogramas direccionales contienen más pares de datos para calcularse, por lo que se vuelven más confiables y mejor estimados, lo cual permite identificar características que son muy influyentes para el ajuste del modelo como la anisotropía y tendencia. Por lo tanto, el ajuste de un modelo adecuado se vuelve menos complejo y más confiable. Es muy importante tomar en cuenta todas las técnicas e indicadores conocidos para llegar al mejor ajuste del modelo, ya que aunque se tenga información que puede ser muy útil, cualquier estimación puede ser alterada por la decisión del modelo del geoestadístico, ésto debido a que el modelo es elegido en base al patrón de continuidad espacial, y éste es elegido por el geoestadístico.

3.9. Bases de datos con 400 observaciones

En estadística normalmente entre más datos se tengan más información se puede obtener de la muestra. El análisis exploratorio de datos toma mayor importancia en estos casos. El número de observaciones para este caso son una cantidad que permite mejorar la estimación, es decir, conforme se tienen más datos se puede utilizar más información para estimar. Una base de datos con una fuerte cantidad de información permite observar las características y distribución de los datos más fácilmente. Además, los variogramas son calculados con un número mayor de pares por lo que se vuelven más confiables. Sin embargo, cuando se tiene un número muy vasto de datos también es importante revisar la información para saber cuál es más representativa y si es útil o se incorporan más datos atípicos o valores extremos. Las características de gran influencia como tendencia, asimetría o anisotropía se vuelven más visibles, lo cual tiene un gran impacto en el ajuste del modelo que se propone y por ende en la estimación. Si existe anisotropía, en ocasiones las muestras más cercanas no son tan relevantes debido a los ejes de anisotropía, por lo que es importante revisar cuidadosamente el análisis variográfico. Durante el proceso de ajuste del modelo, el contar con mayor información permite que se tengan menos regiones del área total donde no se cuenta con datos. Cabe destacar que con un número grande de datos también la cantidad de cálculos aumenta proporcionalmente y entonces el tiempo que se necesita para realizar la estimación también aumenta. Algunos programas que con pocas muestras tardan menos de un minuto en realizar los cálculos pueden tardar numerosos minutos adicionales dependiendo del tamaño de muestra.

Capítulo 4

Aplicación de la metodología a los Datos del tipo 1

4.1. Proceso de estimación de los datos originales del tipo 1

Es importante destacar el proceso que se llevó a cabo para la estimación cuando se tienen todos los datos originales ya que este es el modelo al cual se desea llegar. Se inicia con el gráfico de la distribución espacial de los datos (Figura 4.1) y el (Cuadro 4.1) de sus estadísticas básicas.

Dentro de las estadísticas básicas se observa que la media es mayor que la mediana, lo cual indica una asimetría positiva que se confirma con el coeficiente de asimetría. Además, el 50 % de los datos se encuentran en el intervalo [0, 2.81], y el restante 50 % se encuentra en el intervalo (2.81, 27.73]. El histograma (Figura 4.2) muestra visualmente la asimetría mencionada, así como 24 datos atípicos y el q-q plot (Figura 4.3) confirma que la distribución no es normal.

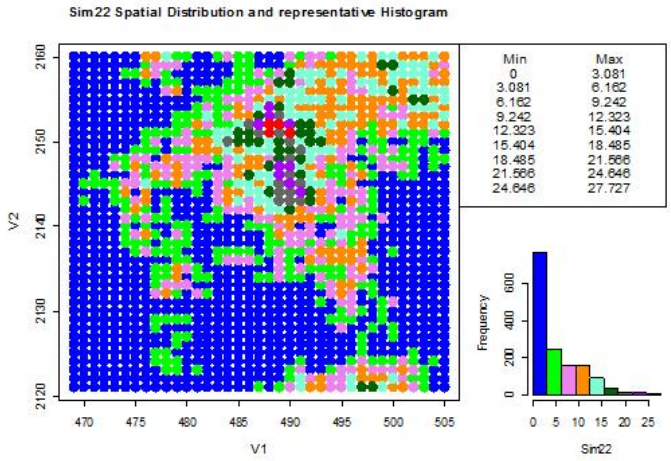

Figura 4.1: Distribución D1c1480

Nombre	**Estadísticas**
Número total	1480
Disctancia max	53.07541804
Distancia min	1
Media	4.601
Varianza	28.9057282
Desviación estándar	5.376404765
Coeficiente var	1.168486935
Rango min	0
1er cuantil	0
Mediana	2.815
3er cuantil	7.926
Máximo rango	27.73
Asimetría	1.191415392
Curtosis	3.899409767

Cuadro 4.1: Estadísticas básicas D1c1480

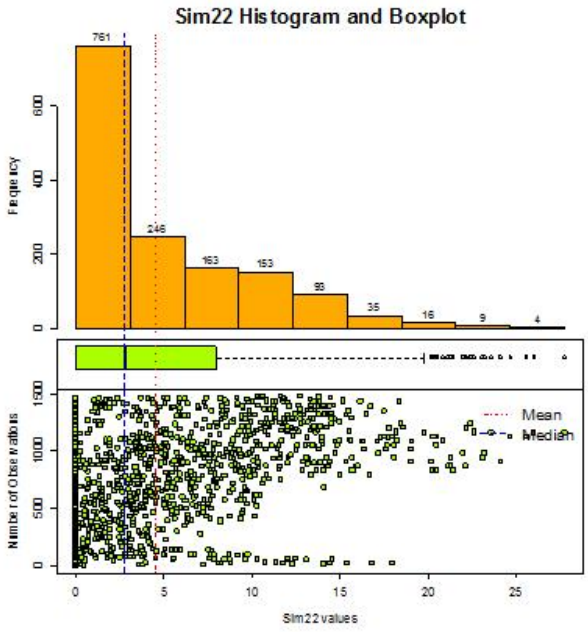

Figura 4.2: Histograma de D1c1480

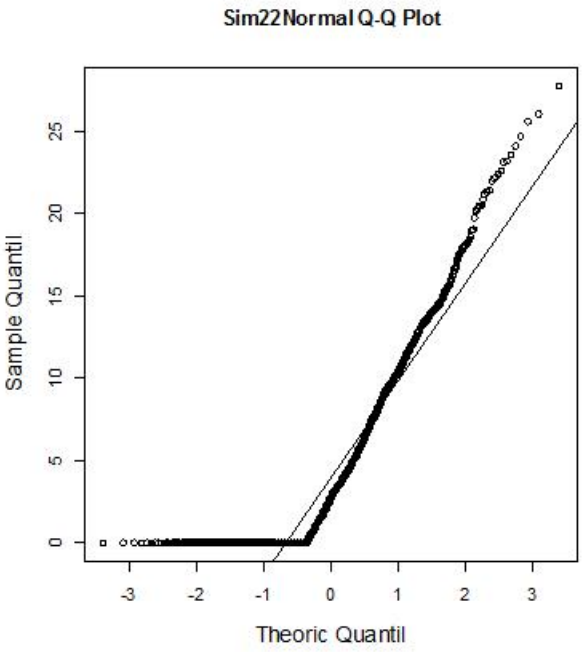

Figura 4.3: Q-Q plot de D1c1480

Debido a que no cumple con el supuesto de asimetría, se procede a realizar una transformación de raíz cuadrada para mejorar la simetría y reducir la escala, de manera que se continúe con la metodología. Se inicia nuevamente el análisis exploratorio de datos.

Dentro de las estadísticas básicas (Cuadro 4.2) una vez realizada la transformación, se observa que la media ahora es menor que la mediana y son muy cercanas, lo cual indica que la muestra se encuentra mucho más centrada. Además, el 75 % de la información se encuentra en el intervalo [0, 1.67] y el restante 25 % está en el intervalo (1.67, 5.26].

El histograma (Figura 4.5) permite observar que ya no se tienen datos atípicos y que la muestra se encuentra más centrada y simétrica. El q-q plot (Figura 4.6) confirma que no cumple normalidad, pero es suficiente con la simetría para continuar con la metodología.

Nombre	**Estadísticas**
Número total	1480
Distancia max	53.07541804
Distancia min	1
Media	1.594
Varianza	2.063296223
Desviación estándar	1.436417844
Coeficiente var	0.901418976
Rango min	0
1er cuantil	0
Mediana	1.678
3er cuantil	2.815
Máximo rango	5.266
Asimetría	0.252484284
Curtosis	1.747620078

Cuadro 4.2: Estadísticas básicas raízD1c1480

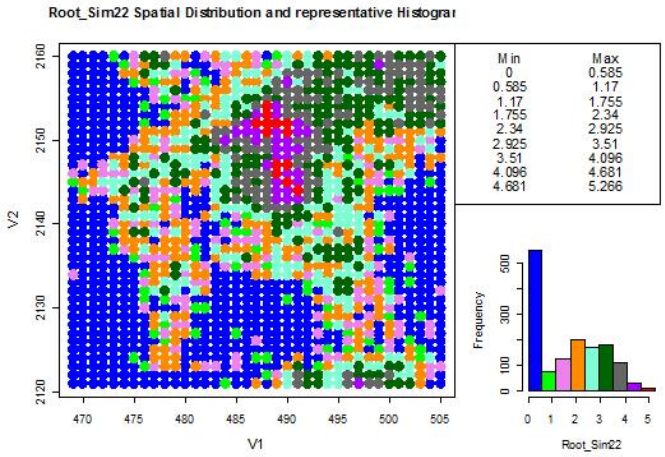

Figura 4.4: Distribución de raízD1c1480

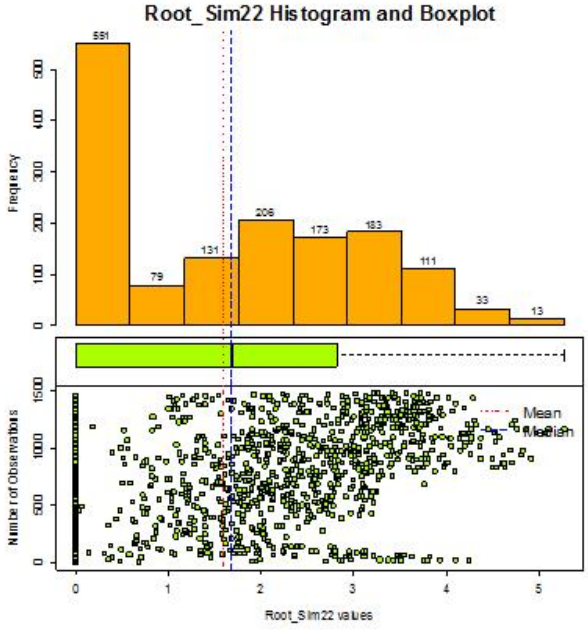

Figura 4.5: Histograma de raízD1c1480

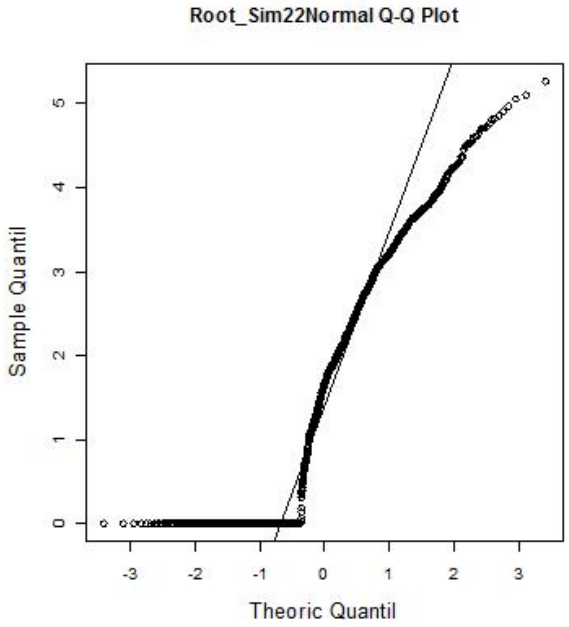

Figura 4.6: Q-Q plot de raízD1c1480

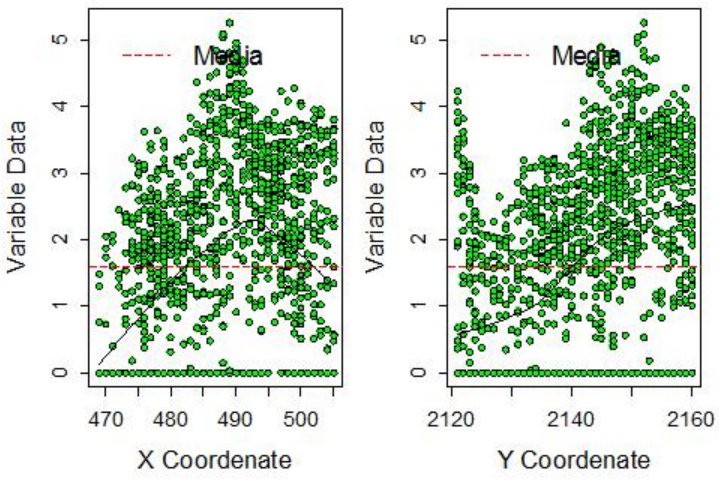

Figura 4.7: Gráfico respecto a las coordenadas raízD1c1480

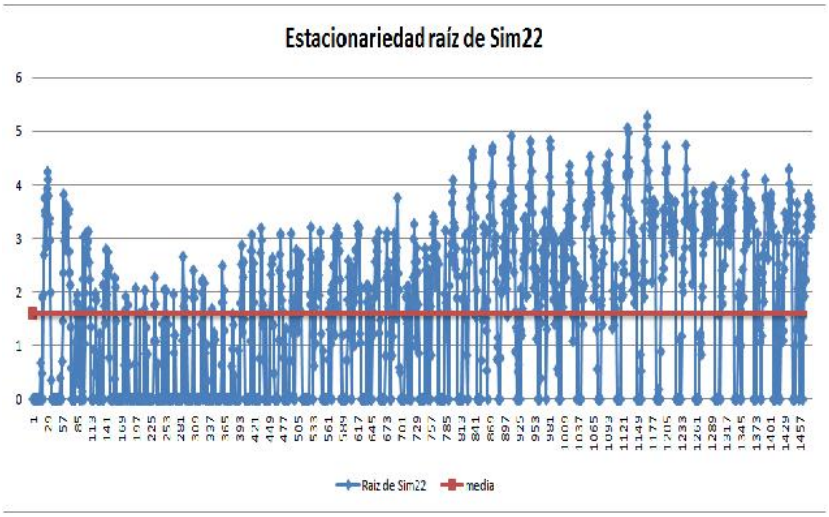

Figura 4.8: Gráfico de estacionariedad raízD1c1480

Se realiza el gráfico respecto a las coordenadas (Figura 4.7) y el gráfico de estacionariedad (Figura 4.8). Gráficos en los cuales no se observa tendencia ni indicios de que no cumpla estacionariedad. Por lo que se procede a realizar el análisis variográfico.

El variograma adireccional (Cuadro 4.3 Figura 4.9) confirma que no se tiene tendencia y el variograma en cuatro direcciones (Cuadro 4.4 Figura 4.10) no muestra diferencia significativa en los alcances como para indicar anisotropía. Utilizando los variogramas se realiza el mapa de anisotropía (Figura 4.11) el cual tampoco muestra diferencia con respecto a la dirección. Por lo que se elige utilizar únicamente variogramas isotrópicos.

Disctancia max	53.07541804
Distancia min	1
Dirección	0°
Tolerancia	90°
Intervalos	25
Distancia Lag	1

Cuadro 4.3: Variograma adireccional de raízD1c1480

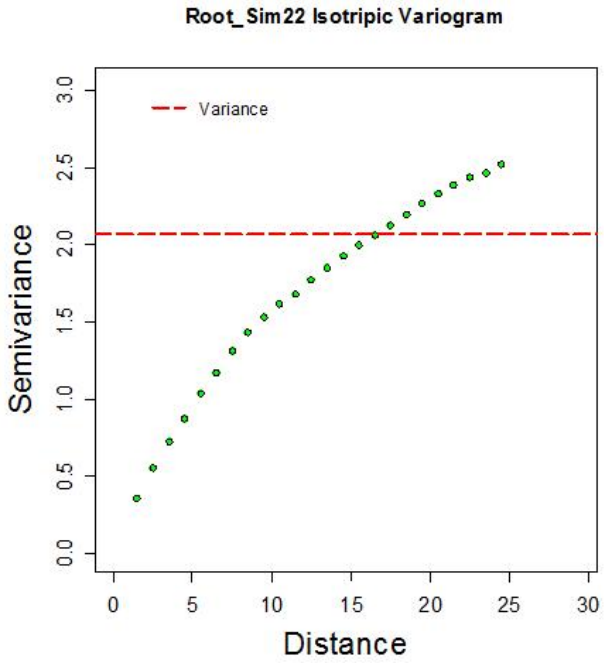

Figura 4.9: Variograma adireccional de raízD1c1480

Distancia max	53.07541804
Distancia min	1
Dirección	0°,45°,90°,135°
Tolerancia	22.5°
Intervalos	25
Distancia Lag	1

Cuadro 4.4: Variograma 4 direcciones raízD1c1480

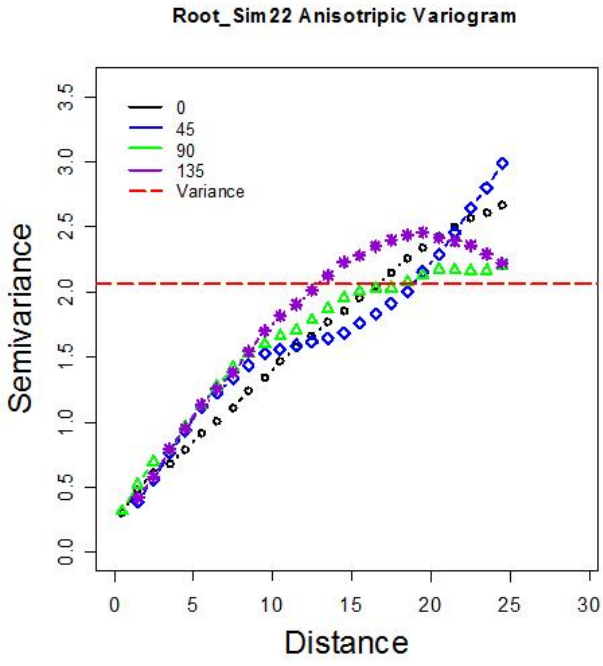

Figura 4.10: Variograma en 4 direcciones de raízD1c1480

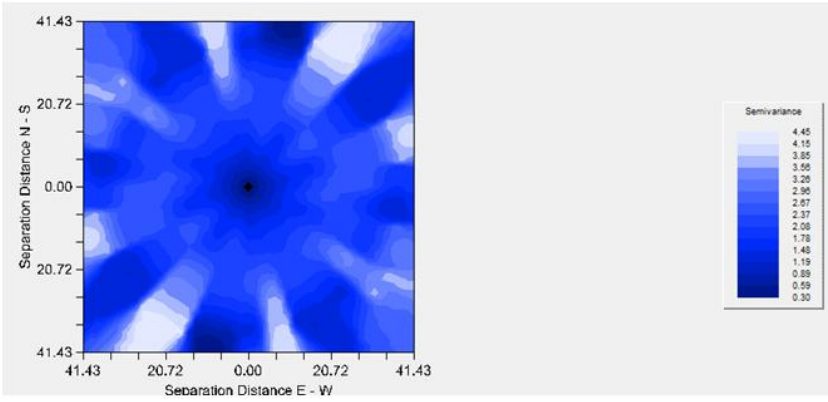

Figura 4.11: Mapa de anisotropía raízD1c1480

Una vez decidido que no existen indicios de tendencia o anisotropía, se generan las propuestas de modelos de variograma (Cuadro 4.5 Figura 4.12) y se realiza el ajuste visual sobre el modelo elegido. (Cuadro 4.6 Figura 4.13).

Modelo	Nugget	Sill+Nugget	Rango	SCE
Exponencial	0.15	3.15	15.98	0.0141
Esférico	0.38	2.5	26.69	0.0884
Gaussiano	0.71	2.49	13.13	3.30

Cuadro 4.5: Propuestas para modelos de variograma raízD1c1480

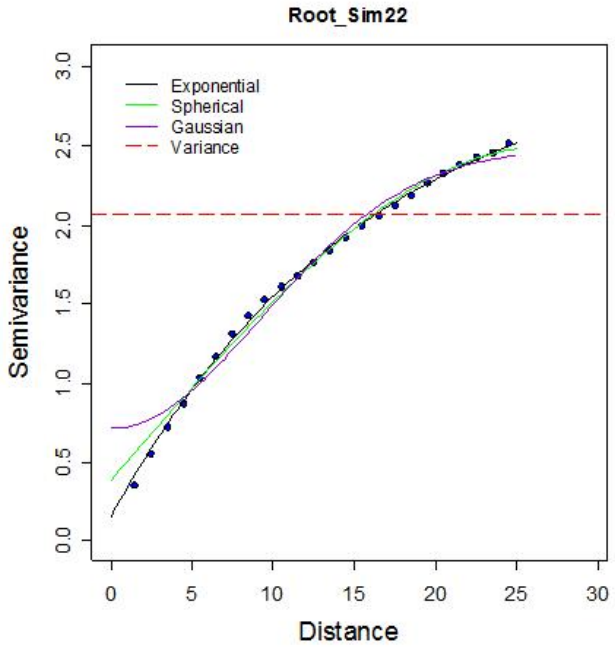

Figura 4.12: Propuestas de modelos de variograma de raízD1c1480

Modelo	Nugget	Sill+Nugget	Rango	SCE
Esférico	0.2	2.5	26	0.1214

Cuadro 4.6: Modelo de variograma elegido raízD1c1480

Una vez seleccionado el modelo de variograma, se verifica que sea adecuado mediante la validación cruzada (Cuadro 4.7). En el cuadro se muestra que la media de los errores es muy cercana a cero. Además, la distribución de los estimados es muy cercana a la distribución de la transformación de raíz cuadrada de los datos originales del tipo 1.

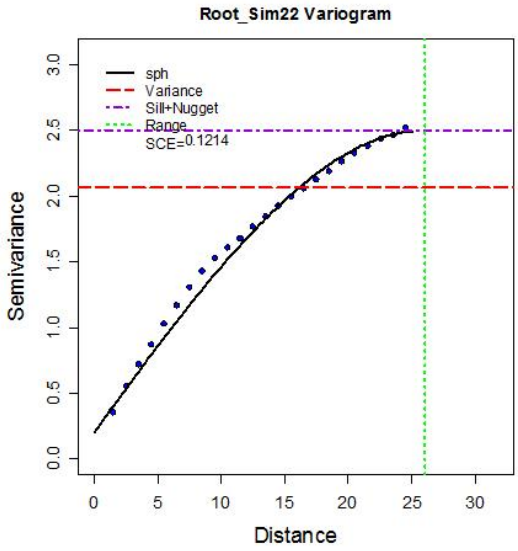

Figura 4.13: Modelo de variograma elegido para raízD1c1480

Nombre	**Raíz D1c1480**	**Estimados**	**Error**
Número total	1480	1480	1480
Distancia max	53.07541804	53.07541804	53.07541804
Distancia min	1	1	1
Media	1.594	1.594	-0.000859
Varianza	2.063296223	1.716307652	0.310478048
Desviación estándar	1.436417844	1.310079254	0.557205571
Coeficiente var	0.901418976	0.821692696	648.6713988
Rango min	0	-0.0716	-2.313
1er cuantil	0	1.3001	-0.2821
Mediana	1.678	1.475	0.01059
3er cuantil	2.815	2.649	0.2947
Máximo rango	5.266	4.797	1.805
Asimetría	0.252484284	0.339091721	-0.19562274
Curtosis	1.747620078	1.923159635	4.250578283

Cuadro 4.7: Validación cruzada raízD1c1480

También se realiza el gráfico de valores reales contra estimados (Figura 4.14), el histograma de errores (Figura 4.15) y el q-q plot de los residuales (Figura 4.16) los cuales muestran que se aproximan a una distribución normal. Por lo tanto el modelo es aceptado.

Una vez que el modelo es aceptado, se procede a realizar la estimación con kriging ordinario y se muestra el mapa de estimaciones (Figura 4.17) a una distancia de 1km por 1km.

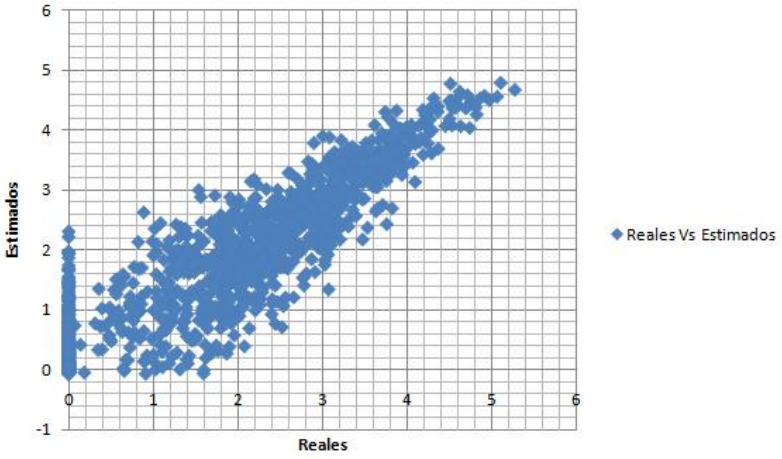

Figura 4.14: Valores reales contra estimados de raízD1c1480

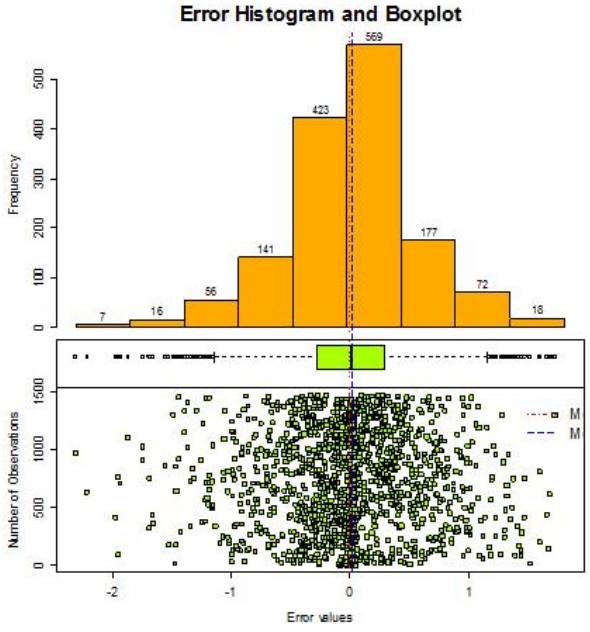

Figura 4.15: Histograma de residuales raízD1c1480

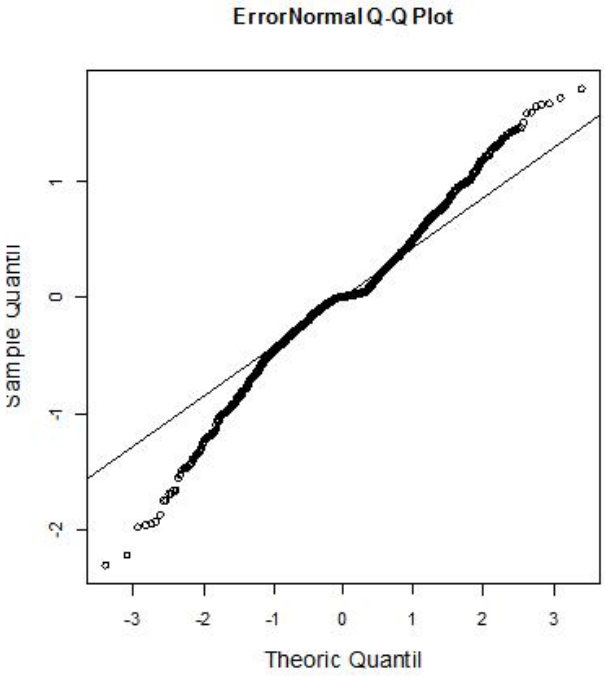

Figura 4.16: Q-Q plot de residuales raízD1c1480

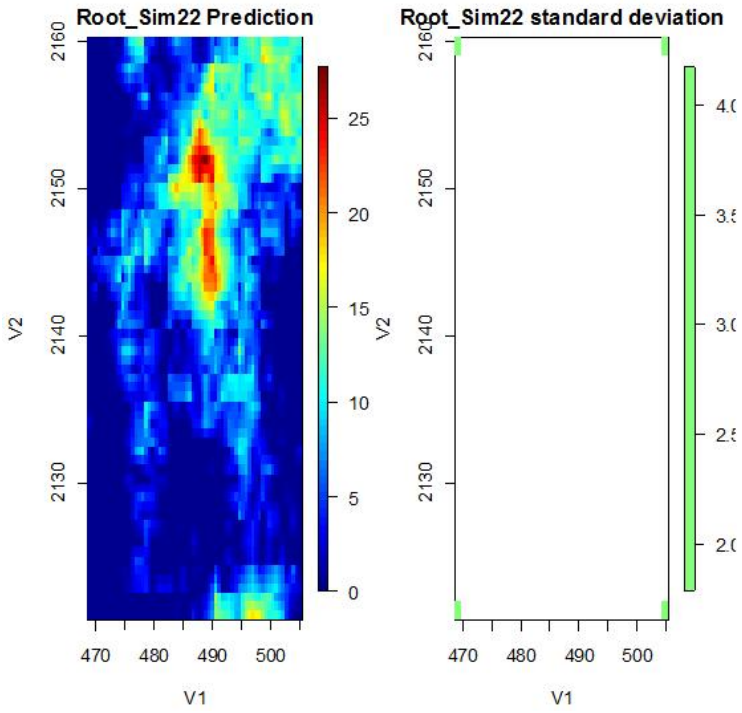

Figura 4.17: Mapa de estimación por kriging de raízD1c1480

4.2. Resumen de los modelos de datos 1

Es importante enfatizar las características de los escenarios realizados para los datos del tipo 1. Se muestra una tabla con el resumen de los variogramas ajustados para cada uno de los mencionados escenarios y se resaltan las características en común que tienen. Las siguientes observaciones ocurrieron en todos los escenarios:

1. La distribución de las muestras tiene asimetría positiva
2. En todos los escenarios de datos 1, debido a la asimetría, la muestra es tratada con una transformación de raíz cuadrada.
3. La información de cada una de las muestras no cumple normalidad.
4. La información de cada una de las muestras no presenta tendencia.
5. La información de cada una de las muestras no presenta indicios de no estacionariedad.
6. La información de cada una de las muestras no presenta indicadores de anisotropía.
7. En todos los escenarios los variogramas en 4 direcciones son calculados con el mismo número de intervalos y el mismo lag que en el cálculo de su respectivo variograma adireccional.
8. Los modelos de cada escenario pasaron el análisis de residuales, es decir, la media de los errores es cercana a cero, el gráfico de valores reales contra estimados se aproxima a una línea de 45°, el histograma y q-q plot son muy cercanos a la distribución normal.
9. En todos los escenarios se realizó la estimación por medio de Kriging ordinario.

Ahora se muestra un resumen (Figura 4.18) de la información de los modelos de variograma elegidos para cada escenario de Datos 1, así como la información relevante del análisis variográfico de los escenarios de Datos 1.

1. Se observa que los modelos ajustados en casi todos los escenarios son esféricos, excepto en el D1MAc400.
2. El nugget se encuentra en un intervalo de [0, 0.28] para todos los escenarios de Datos 1.
3. La meseta se encuentra en un intervalo de [2.34, 3.63] para todos los escenarios de Datos 1.
4. Y el alcance varía en un intervalo de [21.35, 35] para todos los escenarios de Datos 1.

Cabe destacar que el modelo exponencial tiene el mínimo de alcance y el máximo en el nugget, esto puede ser por la forma que tiene el modelo en sí mismo con respecto al esférico.

ANÁLISIS ESTRUCTURAL Y VARIOGRÁFICO						
Escenario	Variograma adireccional	Variograma en 4 direcciones	Modelo	Nugget	Meseta (Sill+Nugget)	Alcance
DATOS 1	**inter 25, lag 1**	**lag 1**	**Esférico**	**0.2**	**2.5**	**26**
D1MRc036	inter 10, lag 6	lag 6	Esférico	0.1	2.7	28
D1MRc100	inter 10, lag 4	lag 4	Esférico	0.18	2.45	25
D1MRc400	inter 13, lag 2	lag 2	Esférico	0.17	2.4	25
D1MAc036	inter 11, lag 2	lag 2	Esférico	0	2.34	22.4
D1MAc100	inter 10, lag 3	lag 3	Esférico	0.1872	2.4764	30
D1MAc400	inter 10, lag 3	lag 3	Exponencial	0.28	3.63	21.35
D1MCc036	inter 10, lag 3	lag 3	Esférico	0.14	3.5	35
D1MCc100	inter 10, lag 3	lag 3	Esférico	0.27	2.68	27
D1MCc400	inter 13, lag 2	lag 2	Esférico	0.26	2.52	26

Figura 4.18: Modelos de variograma por escenario de DATOS1

Por otro lado, es importante mencionar que los modelos ajustados se realizaron de manera completamente independiente, lo cual resalta la notoria similitud que existe entre los modelos que se utilizaron para realizar la estimación en cada uno de los escenarios de Datos 1.

4.3. Comparación respecto a los datos originales del tipo 1

La importancia de los diferentes escenarios radica en que en la práctica no se conoce la cantidad de datos, el tipo de muestreo o las características que pueden presentar las bases de datos con las cuales se tiene que trabajar. Por lo tanto, es importante destacar las dificultades que se presentan cuando varían las características mencionadas. Las dificultades más relevantes presentadas en los escenarios fueron las siguientes:

1. *Escenarios con 36 observaciones D1MRc036, D1MAc036 y D1MCc036* El ajuste del variograma fue casi completamente visual para los escenarios de malla regular y aleatorio, debido a que el variograma está calculado pobremente y por lo tanto no permite definir bien los parámetros del variograma, por lo que la experiencia cuenta mucho al realizar el ajuste visual. En el escenario del muestreo combinado no fue completamente visual pero si tuvo mucha importancia la decisión personal en base a la experiencia. En estos escenarios se muestra que aún cuando se tiene poca información, se pueden observar las características importantes de la muestra y si se adapta adecuadamente el modelo, el mapa de la estimación con Kriging resulta muy cercano al obtenido con los datos originales del tipo 1. En los escenarios de 36 observaciones, el muestreo de malla regular no fue tan efectivo e incluso fue complejo para adecuarle un modelo, esto debido a que las muestras se encontraban a grandes distancias de separación. El muestreo aleatorio también tuvo sus complicaciones debido a que dentro de la región había partes que estaban pobremente muestreadas, por lo que se hacía evidente la falta de información. El muestreo combinado

resultó ser el más sencillo para ajustarle un modelo y el más cercano al mapa de datos originales, mientras que el escenario del muestreo aleatorio y de malla regular cumplen con las partes estimadas con valores altos pero cuando los valores disminuyen se pierde respecto al mapa de datos originales del tipo 1.

2. *Escenarios con 100 observaciones D1MRc100, D1MAc100 y D1MCc100*
   Los escenarios con 100 observaciones son representativos dentro de la base de datos del tipo 1. Las bases de este tipo son más comunes que las demás y el número de muestras permite realizar los cálculos con suficiente rapidez y confiabilidad. Para estos escenarios, se observa en los gráficos la similitud que existe con los datos originales del tipo 1. Desde el análisis exploratorio de datos se observan características en común como asimetría. El escenario D1MCc100 es el muy parecido a los datos originales del tipo 1. El escenario D1MRc100 tiene una zona estimada que no coincide con los datos originales del tipo 1 pero en conjunto si se asemeja. El escenario D1MAc100 es el que mejor coincide con los datos originales del tipo 1. Es importante remarcar que en ocasiones no se necesita un modelo igual, sino más bien el modelo que mejor se ajuste a los datos y a la información obtenida.

3. *Escenarios con 400 observaciones D1MRc400, D1MAc400 y D1MCc400*
   Principalmente, en los escenarios con 400 observaciones el tiempo para realizar la estimación aumentó considerablemente respecto a los demás. La información obtenida en los mapas de estimación resultante es más acertada pero menos continua sobre los mapas del kriging de los demás escenarios. Conforme aumentó la cantidad de información, el tipo de muestreo resultó ser menos importante, debido a que se cuenta con mucha más información sobre toda la región. Por lo tanto, los muestreos de malla regular y combinado resultaron en mapas muy parecidos a los de los datos originales del tipo 1, al igual que el de muestreo aleatorio excepto por una pequeña zona que resultó mal estimada. Cuando se tienen tantas muestras también se observa que los variogramas son muy parecidos entre los distintos tipos de muestreo y se asemejan al variograma de los datos originales del tipo 1, por lo que el ajuste visual del modelo es similar al de los datos originales del tipo 1 y por consiguiente la estimación también lo es.

A continuación se muestra el procedimiento a seguir con 3 escenarios representativos. Se fija en 100 el número de observaciones y se realiza el análisis exploratorio de datos, el análisis estructural y la estimación con Kriging para los 3 tipos de muestreo.

Por lo tanto, los casos representativos son:

- D1MRc100
- D1MAc100
- D1MCc100

4.4. Base de datos del tipo 1 con muestreo de malla regular y 100 observaciones (D1MRc100)

Dentro de la metodología mencionada se comienza por el análisis exploratorio. Para la base D1MRc100 se muestra el gráfico de dispersión de los datos (Figura 4.19), así como las estadísticas básicas (Cuadro 4.8). Dentro de las estadísticas básicas se puede destacar que la media es mayor que la mediana por lo que indica asimetría positiva, lo cual también se puede observar en el coeficiente de asimetría. Por otro lado, se muestra que la distribución de los datos se encuentra en un 75 % dentro del intervalo [0, 6.961], y el 25 % restante se encuentra en el intervalo [7, 20.16]. También cabe destacar que la desviación estándar no es tan grande.

Nombre	**Estadísticas**
Número total	100
Distancia max	50.91168825
Distancia min	4
Media	4.293
Varianza	28.93933882
Desviación estándar	5.379529609
Coeficiente var	1.253214998
Rango min	0
1er cuantil	0
Mediana	1.774
3er cuantil	6.961
Máximo rango	20.16
Asimetría	1.119417428
Curtosis	3.199724284

Cuadro 4.8: Estadísticas básicas D1MRc100

Se realiza el histograma (Figura 4.20) en el cual se puede apreciar que la distribución de los datos no es simétrica ya que una gran parte de información se encuentra dentro del primer intervalo. Además, el gráfico de caja y brazos muestra la existencia de 4 datos atípicos y que el 3er cuantil contiene sólo 3 intervalos de los 9 existentes. También se realiza un q-q plot para verificar normalidad (Figura 4.21), el cual debido a la asimetría, confirma la no normalidad en los datos.

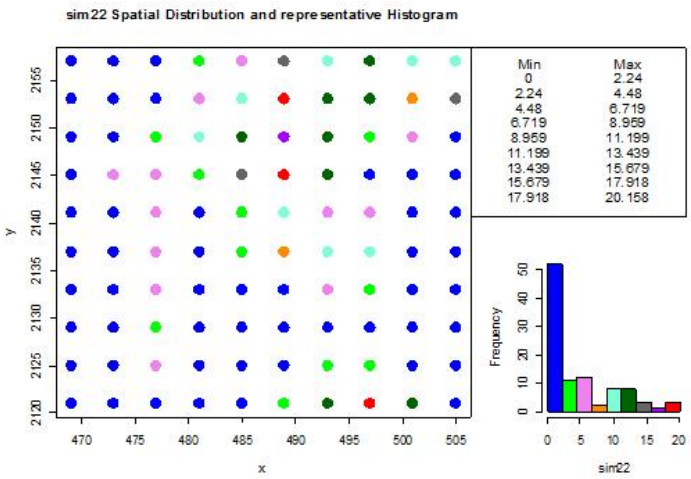

Figura 4.19: Distribución de D1MRc100

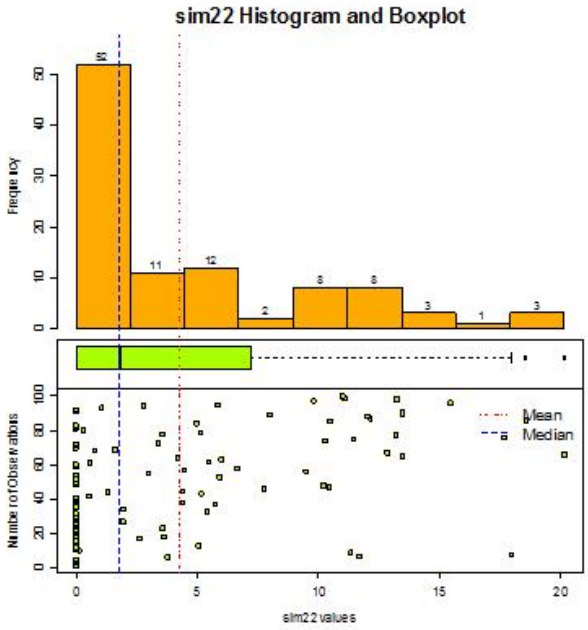

Figura 4.20: Histograma de D1MRc100

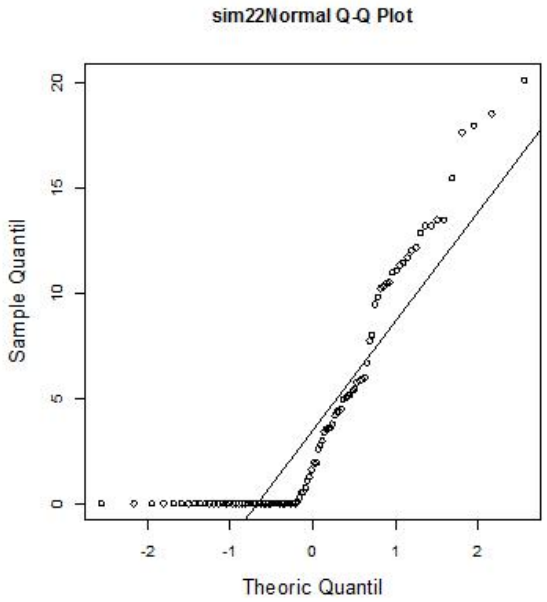

Figura 4.21: Q-Q plot de D1MRc100

Sin embargo, para la metodología es suficiente con tener una muestra simétrica, por lo que se procede a realizar una transformación para mejorar la simetría de los datos. Se utiliza una transformación de raíz cuadrada y se realiza nuevamente el análisis exploratorio comenzando con la dispersión de los datos (Figura 4.22) y las estadísticas básicas (Cuadro 4.9). En las estadísticas básicas se observa que la media y mediana son muy cercanas. La media sigue siendo mayor que la mediana pero el coeficiente de asimetría es muy cercano a cero. También la desviación estándar se redujo.

Lo siguiente es mostrar el histograma (Figura 4.23), en él se observa una mejora considerable sobre la simetría de la muestra como se observó en las estadísticas, además el diagrama de caja y brazos muestra que ya no hay datos atípicos. Aún cuando el histograma no se asemeja a una distribución normal, se realiza el q-q plot (Figura 4.24) el cual confirma que no cumple con normalidad. Sin embargo, es posible continuar con la metodología debido a que la muestra es suficientemente simétrica.

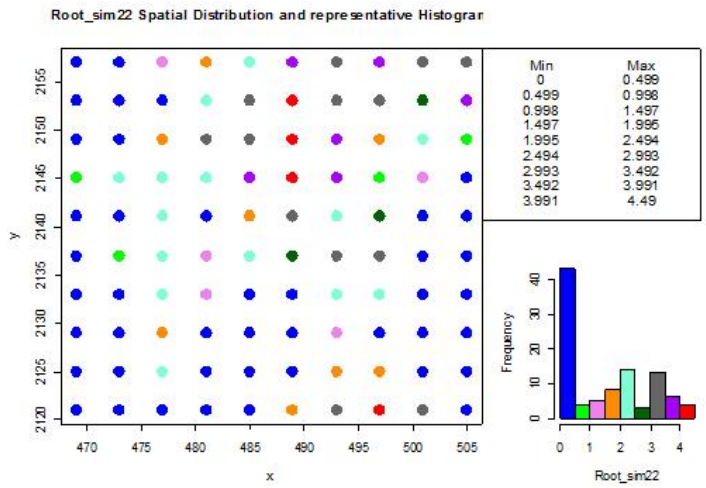

Figura 4.22: Distribución de raízD1MRc100

Nombre	Estadísticas
Número total	100
Distancia max	50.91168825
Distancia min	4
Media	1.463
Varianza	2.173521996
Desviación estándar	1.474286945
Coeficiente var	1.007614336
Rango min	0
1er cuantil	0
Mediana	1.331
3er cuantil	2.637
Máximo rango	4.49
Asimetría	0.383671925
Curtosis	1.677310986

Cuadro 4.9: Estadísticas básicas raízD1MRc100

Se realiza el gráfico con respecto a las coordenadas (Figura 4.25) para observar si existe alguna tendencia. A pesar de que se observa una posible tendencia con respecto a la variable Y, se procede con la metodología y se confirma o anula si existe una tendencia mediante el análisis variogáfico. También se realiza el gráfico de estacionariedad de la media (Figura 4.26).

Una vez que se observa que la simetría y estacionariedad se muestran de manera aceptable y la tendencia no es marcada se procede a realizar el análisis variogáfico. En este análisis se puede verificar la característica de tendencia.

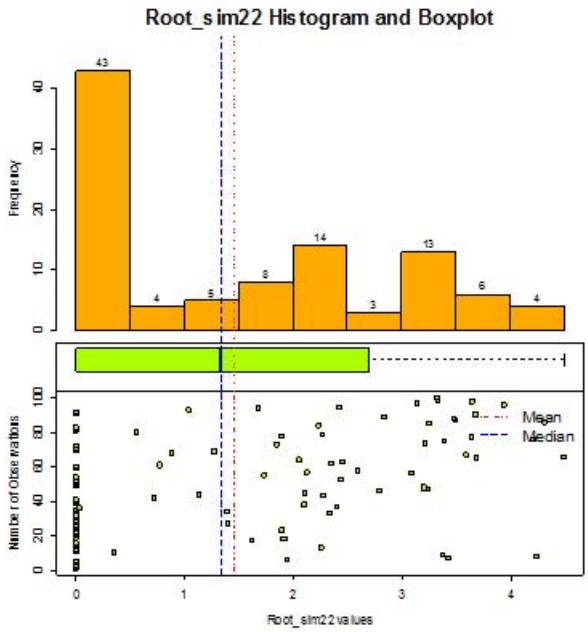

Figura 4.23: Histograma de raízD1MRc100

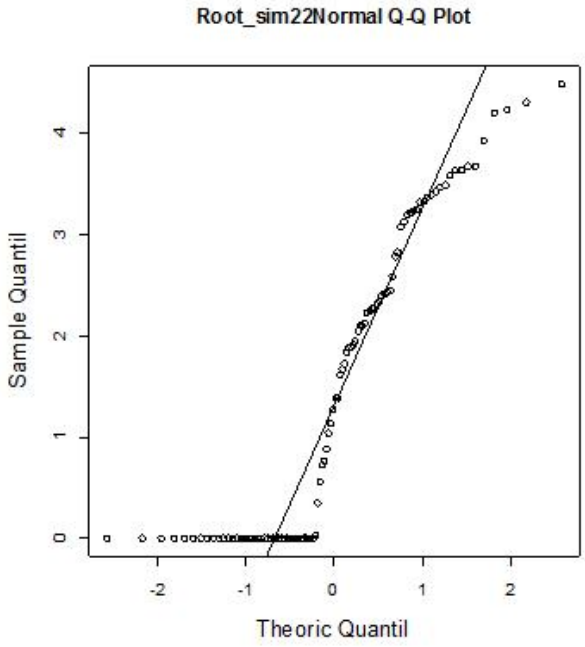

Figura 4.24: Q-Q plot de raízD1MRc100

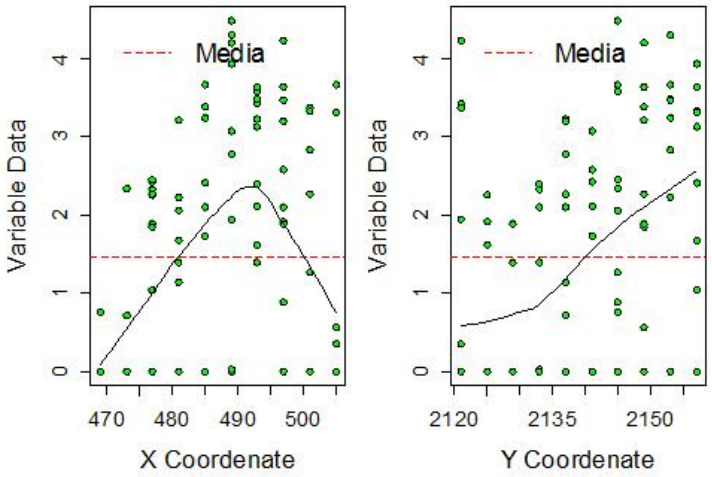

Figura 4.25: Gráfico con respecto a las coordenadas

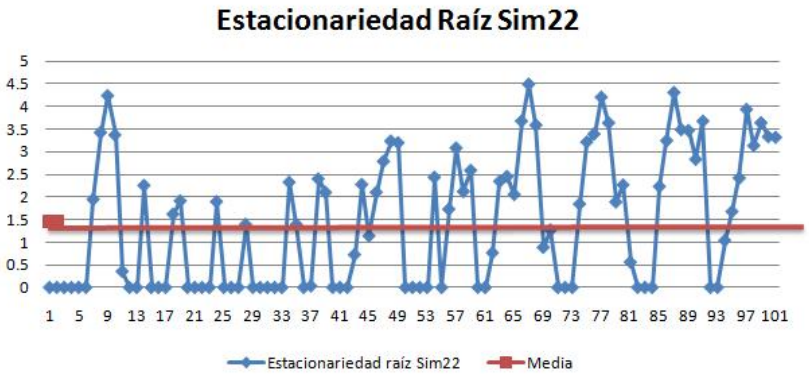

Figura 4.26: Gráfico de estacionariedad raízD1MRc100

Se inicia con un variograma adireccional (Figura 4.27) y sus datos de cálculo (Cuadro 4.10) son elegidos como se menciona en la metodología. En el variograma adireccional no se muestra un comportamiento de h^2, por lo que no se realiza ningún tratamiento para tendencia. Luego se realizan los variogramas en 4 direcciones (Cuadro 4.11 Figura 4.28), en los que se puede observar que no hay indicios de anisotropía ya que los rangos no se encuentran a diferentes distancias.

Disctancia max	50.91168825
Distancia min	4
Dirección	0°
Tolerancia	90°
Intervalos	10
Distancia Lag	4

Cuadro 4.10: Variograma adireccional de raízD1MRc100

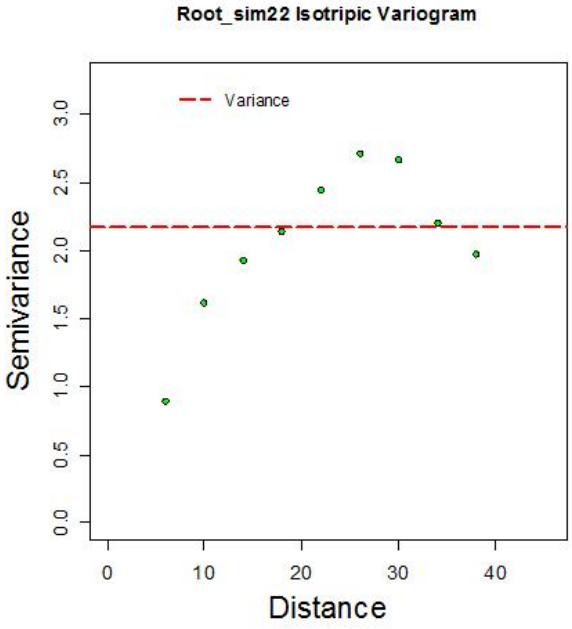

Figura 4.27: Variograma adireccional de raízD1MRc100

Disctancia max	50.91168825
Distancia min	4
Dirección	0°,45°,90°,135°
Tolerancia	22.5°
Intervalos	10
Distancia Lag	4

Cuadro 4.11: Variograma 4 direcciones raíz de Datos 1

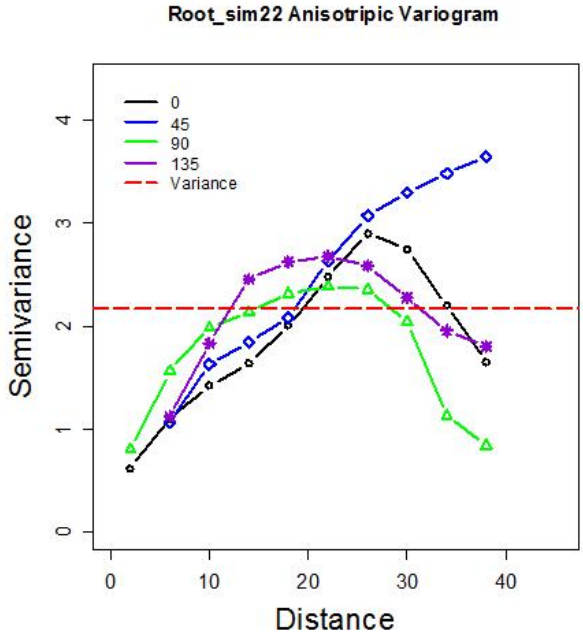

Figura 4.28: Variograma en 4 direcciones de raízD1MRc100

Adicionalmente, a partir de los variogramas mostrados se genera el mapa de anisotropía (Figura 4.29) y se observa que conforme se aleja del centro no se muestran elipses, esto concluye que no se tiene anisotropía. En el Cuadro 4.12 y Figura 4.30 se muestran las propuestas para modelo y se procede a realizar el ajuste. Se observa que el modelo que mejor ajusta es el esférico debido a la forma que tiene el variograma, también presenta la menor suma de cuadrados de los errores y se realiza un ajuste visual para mejores resultados.

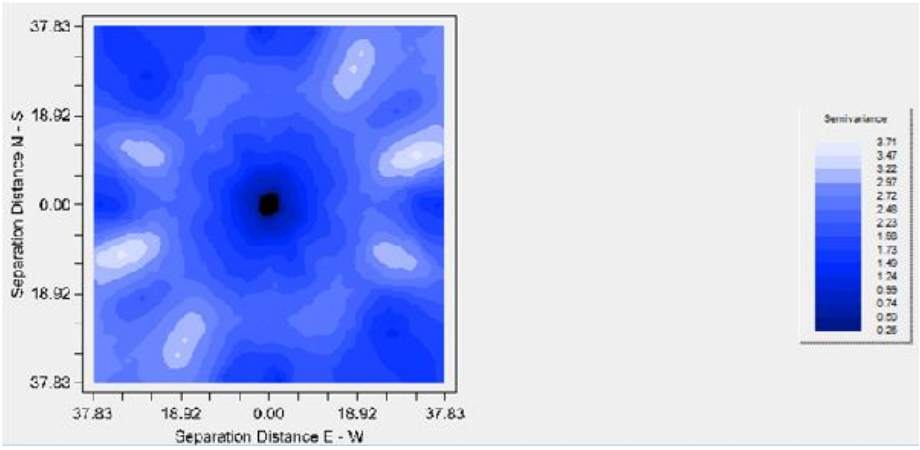

Figura 4.29: Mapa de anisotropía raízD1MRc100

Modelo	Nugget	Sill+Nugget	Rango	SCE
Exponencial	0	2.62	10.14	0.659
Esférico	0.21	2.48	24.99	0.456
Gaussiano	0.53	2.47	11.95	1.68

Cuadro 4.12: Propuestas para modelos de variograma raízD1MRc100

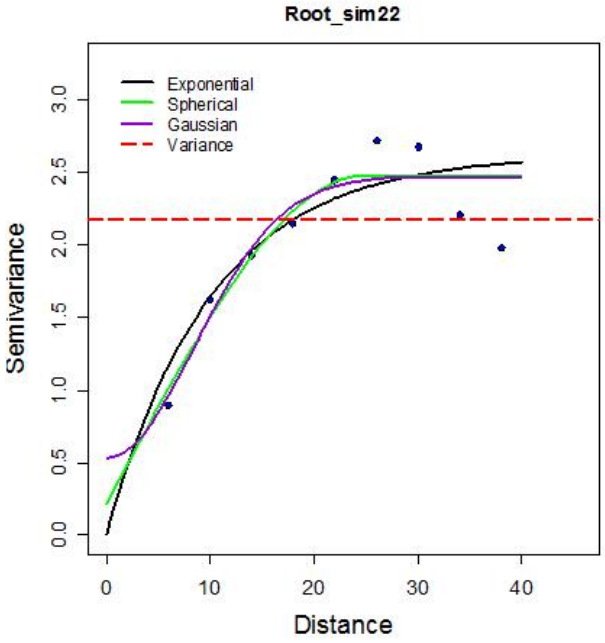

Figura 4.30: Propuestas de modelos de variograma raízD1MRc100

Modelo	Nugget	Sill+Nugget	Rango	SCE
Esférico	0.18	2.45	25	0.441

Cuadro 4.13: Modelo de variograma elegido

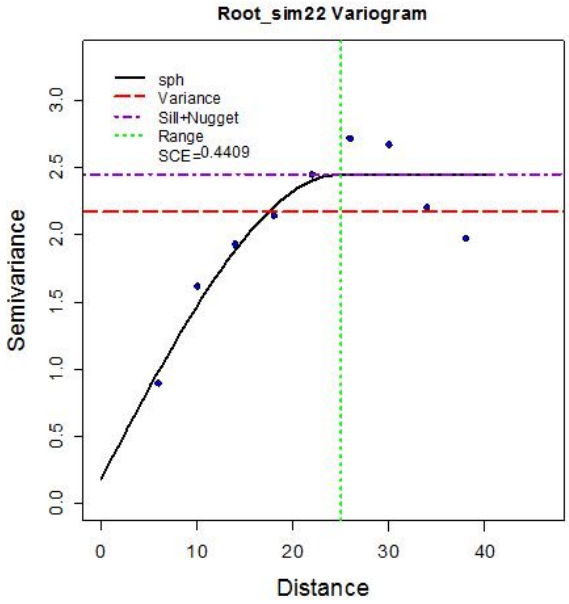

Figura 4.31: Modelo de variograma elegido para raízD1MRc100

Una vez seleccionado el modelo de variograma (Cuadro 4.13 Figura 4.31) se realiza la validación cruzada (Cuadro 4.14) para verificar si el modelo es adecuado. Dentro de las estadísticas de la validación cruzada destaca que la media del error es muy cercana a cero, la media de los estimados y la media de los datos transformados también son muy cercanas y la varianza también es cercana a uno; lo cual son criterios de selección de modelos. Se realiza el análisis de residuales con el gráfico de valores reales contra estimados (Figura 4.32), el histograma (Figura 4.33) y el q-q plot (Figura 4.34).

Nombre	raízD1MRc100	Estimados	Error
Número total	100	100	100
Distancia max	50.91168825	50.91168825	50.91168825
Distancia min	4	4	4
Media	1.463	1.47	-0.007151
Varianza	2.173521996	1.45629545	0.59332559
Desviación estándar	1.474286945	1.20677067	0.77027631
Coeficiente var	1.007614336	0.82076657	107.715923
Rango min	0	-0.2651	-2.095
1er cuantil	0	0.4695	-0.4872
Mediana	1.331	1.28	0.01489
3er cuantil	2.637	2.383	0.5144
Máximo rango	4.49	4.147	1.986
Asimetría	0.383671925	0.44713804	-0.01629247
Curtosis	1.677310986	2.07744488	2.90687488

Cuadro 4.14: Validación cruzada RaízD1MRc100

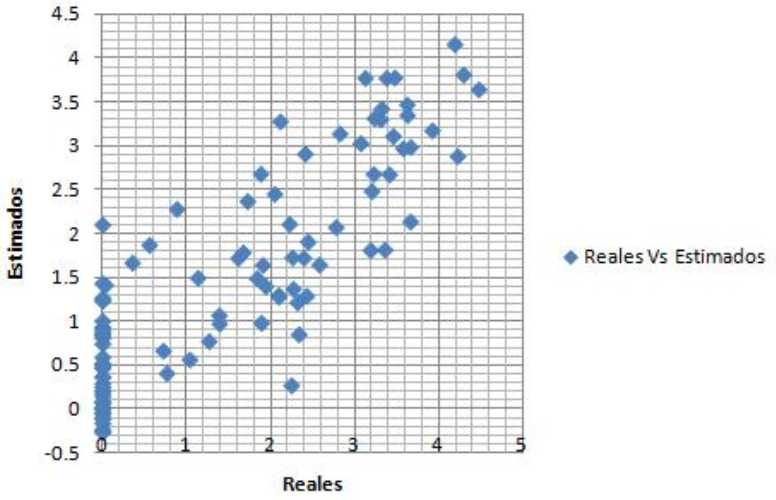

Figura 4.32: Valores reales contra estimados

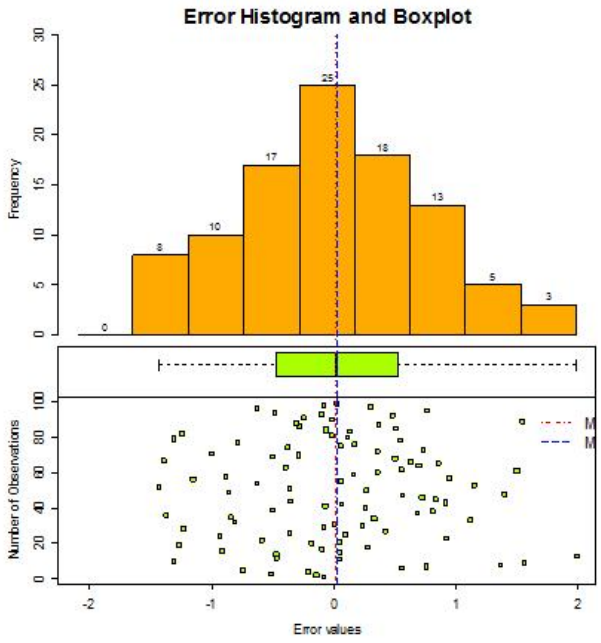

Figura 4.33: Histograma de residuales de raízD1MRc100

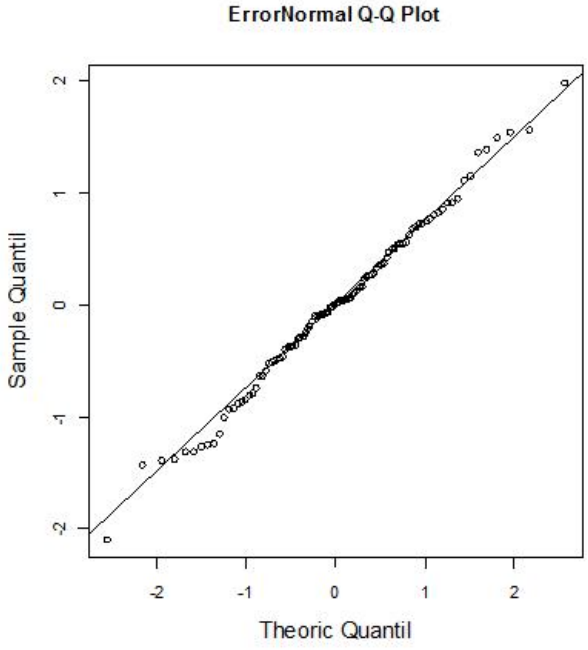

Figura 4.34: Q-Q plot de residuales de raízD1MRc100

Se observa que en la Figura 4.32 los valores se aproximan a una recta de 45° y se distribuyen aleatoriamente alrededor del cero. Con el histograma se observa que tienen una distribución simétrica y muy cercana a la normal, lo que también se aprecia en las estadísticas del error (Dentro del Cuadro 4.14). La media y mediana son muy cercanas a cero y entre sí. En el q-q plot se corrobora que la distribución es muy cercana a la normal, por lo tanto cumplen con los supuestos de los residuales, y el modelo es adecuado. Se procede a realizar la estimación con kriging y se obtiene un mapa de valores de la región estimada a una distancia de 1km por 1km.

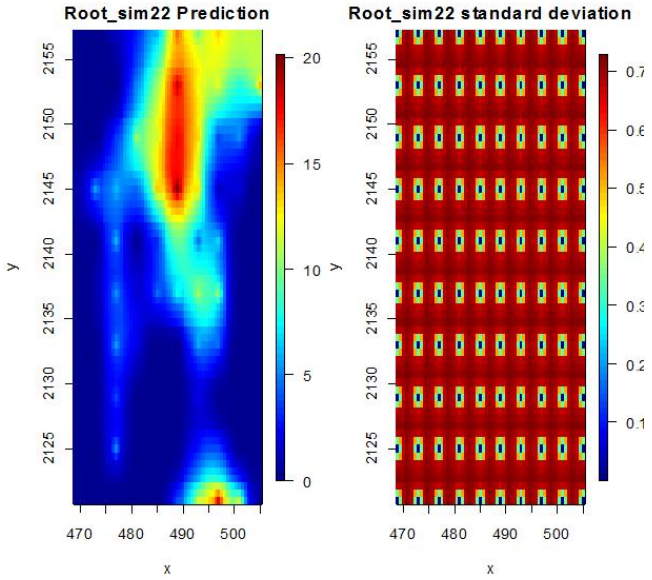

Figura 4.35: Mapa de estimaciones con kriging de raízD1MRc100

4.5. Base de datos del tipo 1 con muestreo aleatorio y 100 observaciones (D1MAc100)

La importancia del caso D1MAc100 radica en que el muestreo aleatorio es más congruente con la realidad. La distribución de los datos es asimétrica y tiene dificultades ya que el 50 % de la información se encuentra en el intervalo [0, 3.64] lo cual indica que el otro 50 % de los datos se encuentra en el intervalo (3.65, 14.71] lo cual es un intervalo mucho más grande. Esto se muestra en la distribución espacial de los datos (Figura 4.36) y sus estadísticas básicas (Cuadro 4.15). Además, dentro de las estadísticas básicas también se observa que la media es mayor que la mediana.

Para continuar con el análisis exploratorio de datos es importante observar el histograma (Figura 4.37), el cual corrobora la asimetría positiva que se ve en las estadísticas básicas. Observando el gráfico de caja y brazos (Figura 4.37) no se tienen datos atípicos. Se realiza el q-q plot (Figura 4.38) y se observa que no cumple normalidad.

Nombre	Estadísticas
Número total	100
Distancia max	51.73973328
Distancia min	1
Media	4.865
Varianza	23.98971281
Desviación estándar	4.897929441
Coeficiente var	1.006696056
Rango min	0
1er cuantil	0
Mediana	3.647
3er cuantil	9.475
Máximo rango	14.71
Asimetría	0.572633124
Curtosis	1.920425366

Cuadro 4.15: Estadísticas básicas D1MAc100

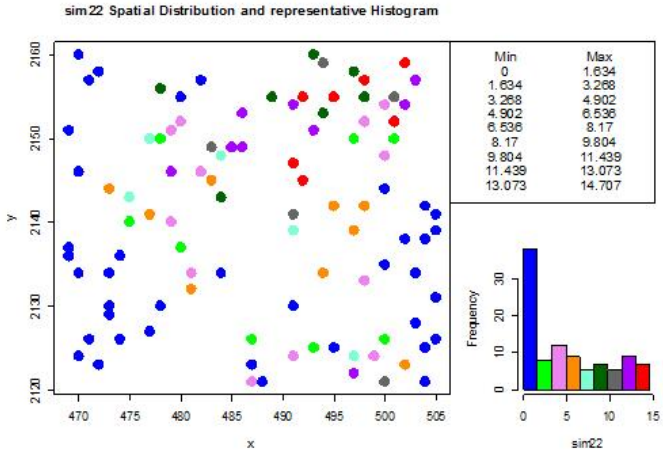

Figura 4.36: Distribución de D1MAc100

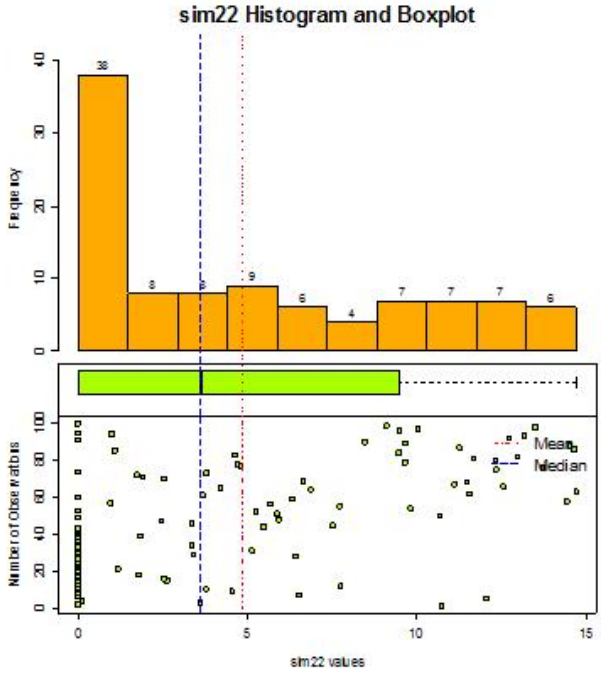

Figura 4.37: Histograma de D1MAc100

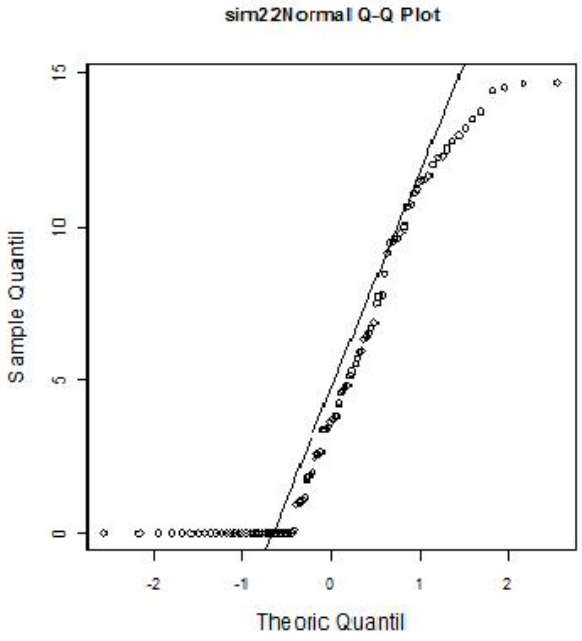

Figura 4.38: Q-Q plot de D1MAc100

Debido a que se tiene una distribución asimétrica, se le aplica una transformación de raíz cuadrada a la variable para disminuir la escala y observar si mejora la asimetría por lo que se vuelve a realizar el análisis exploratorio de datos.

Se muestra la distribución espacial (Figura 4.39) y las estadísticas básicas (Cuadro 4.16). En las estadísticas básicas la media ahora es menor que la mediana pero son mucho más cercanas. La asimetría ahora es negativa pero muy cercana a cero. La escala se redujo de manera que el 50 % de la información se encuentra de [0, 1.91] y el otro 50 % entre (1.91, 3.83]. La varianza también se redujo considerablemente.

El histograma y gráfico de caja y brazos (Figura 4.40) no muestra datos atípicos y se nota una distribución más centrada. Sin embargo, el q-q plot (Figura 4.41) muestra que la distribución no cumple normalidad.

Nombre	**Estadísticas**
Número total	100
Distancia max	51.73973328
Distancia min	1
Media	1.709
Varianza	1.963891432
Desviación estándar	1.401389108
Coeficiente var	0.819946761
Rango min	0
1er cuantil	0
Mediana	1.91
3er cuantil	3.078
Máximo rango	3.835
Asimetría	-0.043965415
Curtosis	1.488674398

Cuadro 4.16: Estadísticas básicas raízD1MAc100

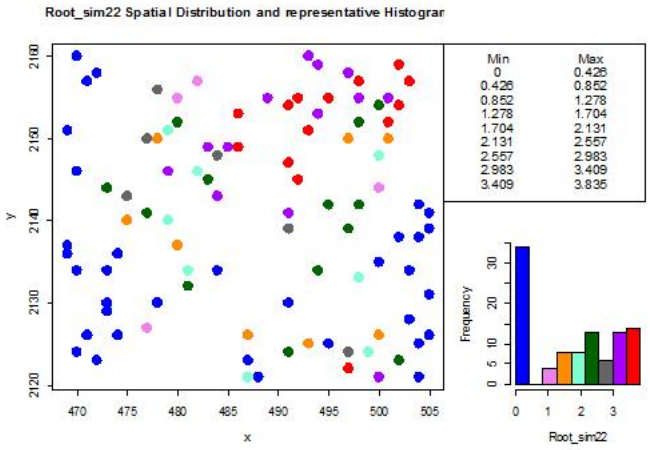

Figura 4.39: Distribución de raízD1MAc100

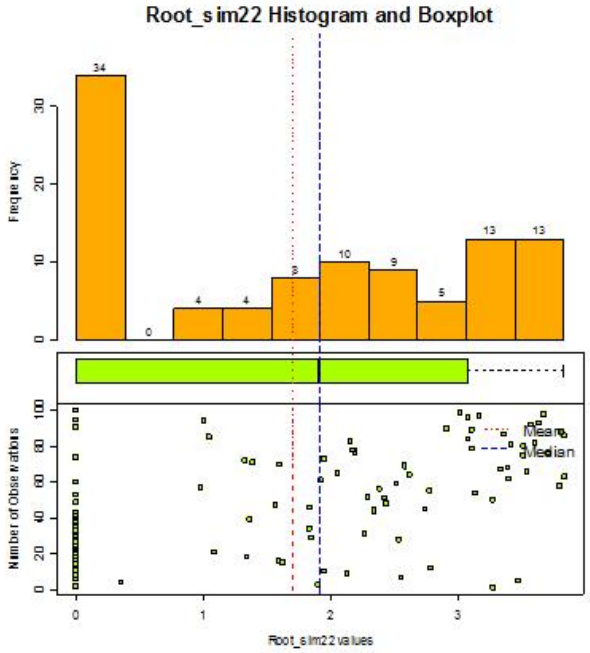

Figura 4.40: Histograma de raízD1MAc100

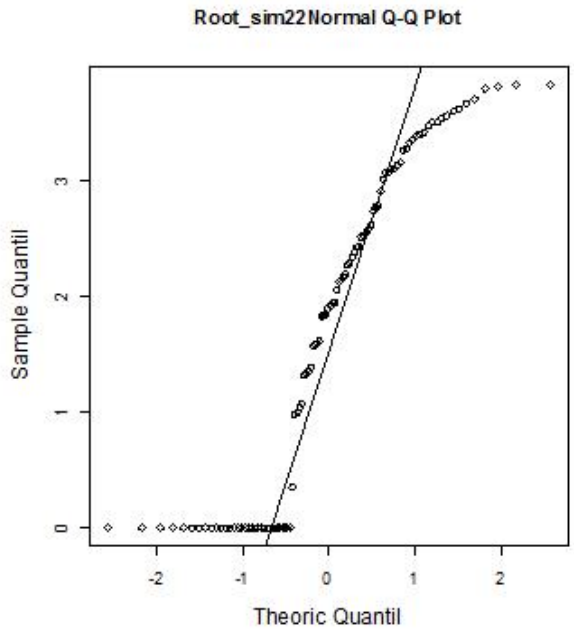

Figura 4.41: Q-Q plot de raízD1MAc100

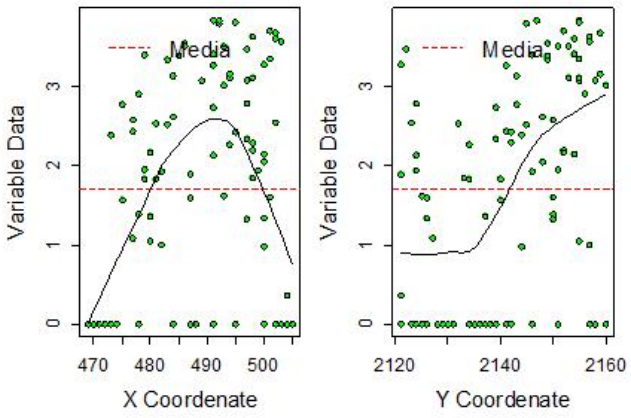

Figura 4.42: Gráfico con respecto a las coordenadas de raizD1MAc100

Dentro de la metodología es suficiente con que la muestra sea simétrica, por lo que se procede a realizar un gráfico con respecto a las coordenadas (Figura 4.42) para verificar si hay tendencia. El gráfico muestra una posible tendencia con respecto al eje Y, pero se verifica más adelante con el análisis variográfico. Se realiza un gráfico de los datos (Figura 4.43) para observar la estacionariedad de la muestra y se observa que no hay indicadores contundentes de que no la cumpla.

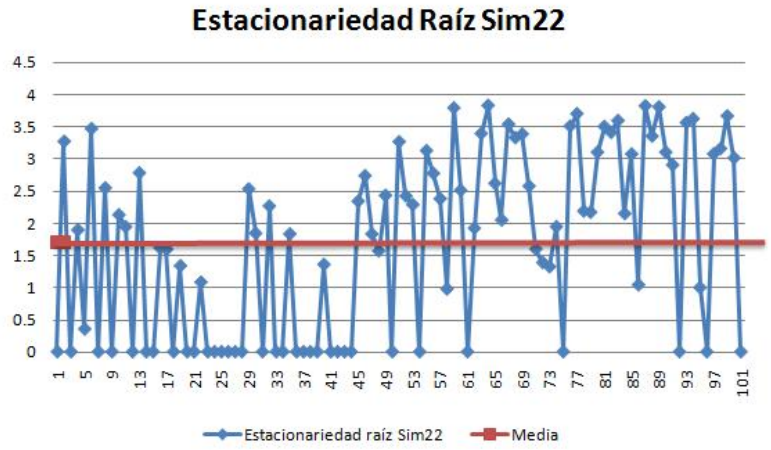

Figura 4.43: Gráfica de estacionariedad

Una vez realizado el análisis exploratorio de datos, se procede a realizar el análisis variográfico. Se calcula el variograma adireccional (Cuadro 4.17 Figura 4.44) con un lag de 3. En el variograma no se muestra comportamiento de h^2, por lo que se confirma que no tiene tendencia significativa. Se continúa con los variogramas en 4 direcciones (Cuadro 4.18 Figura 4.45) donde no se observan distintos alcances respecto a la dirección. Luego, se genera un mapa de anisotropía (Figura 4.46) con el que se corrobora la no existencia de anisotropía.

Disctancia max	51.73973328
Distancia min	1
Dirección	0°
Tolerancia	90°
Intervalos	10
Distancia Lag	3

Cuadro 4.17: Variograma adireccional de raízD1MAc100

Disctancia max	51.73973328
Distancia min	1
Dirección	0°,45°,90°,135°
Tolerancia	22.5°
Intervalos	10
Distancia Lag	3

Cuadro 4.18: Variograma 4 direcciones raízD1MAc100

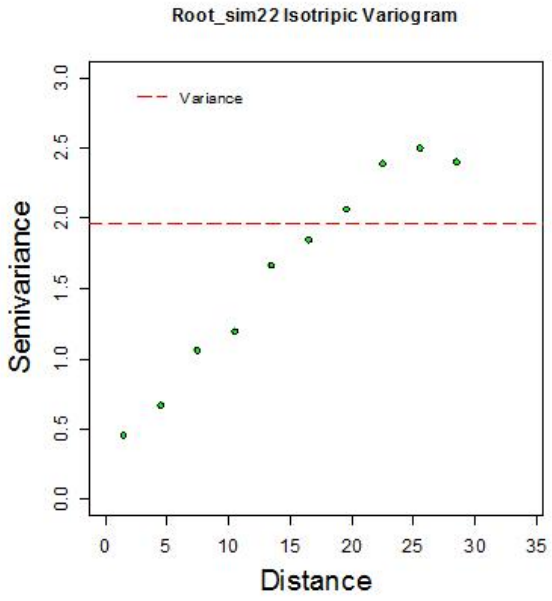

Figura 4.44: Variograma adireccional de raízD1MAc100

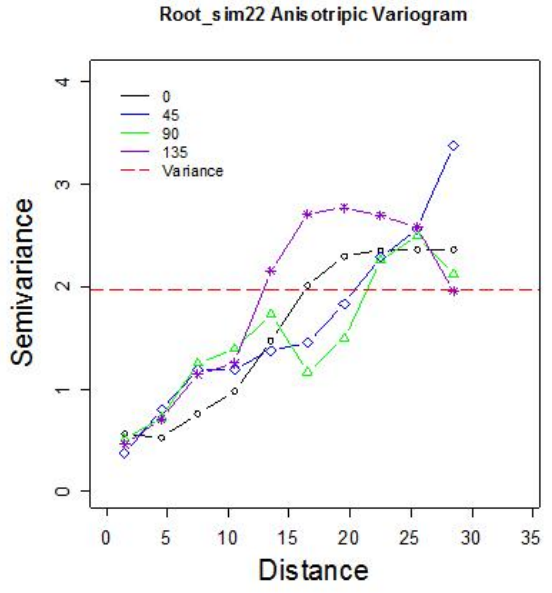

Figura 4.45: Variograma en 4 direcciones de raízD1MAc100

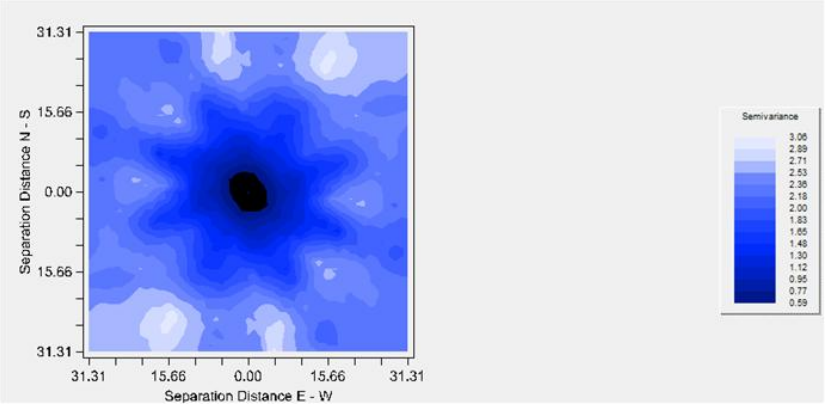

Figura 4.46: Mapa de anisotropía de raízD1MAc100

Una vez que se corrobora que no existe tendencia y no existe anisotropía se continúa con la metodología, por lo que se muestran las propuestas de modelos isotrópicos (Cuadro 4.19 Figura 4.47). Dentro de las propuestas, la menor suma de cuadrados del error la tiene el modelo esférico, además es el que mejor ajusta por la forma del variograma, por lo que sobre ése se realiza el ajuste visual para mejorarlo y llegar finalmente al modelo de variograma elegido (Cuadro 4.20 Figura 4.48).

Modelo	**Nugget**	**Sill+Nugget**	**Rango**	**SCE**
Exponencial	0.05	3.56	22.6	0.109
Esférico	0.14	2.5	30	0.0689
Gaussiano	0.54	2.57	15.65	1.24

Cuadro 4.19: Propuestas para modelos de variograma raízD1MAc100

Modelo	**Nugget**	**Sill+Nugget**	**Rango**	**SCE**
Esférico	0.1872	2.4764	30	0.0624

Cuadro 4.20: Modelo de variograma elegido raízD1MAc100

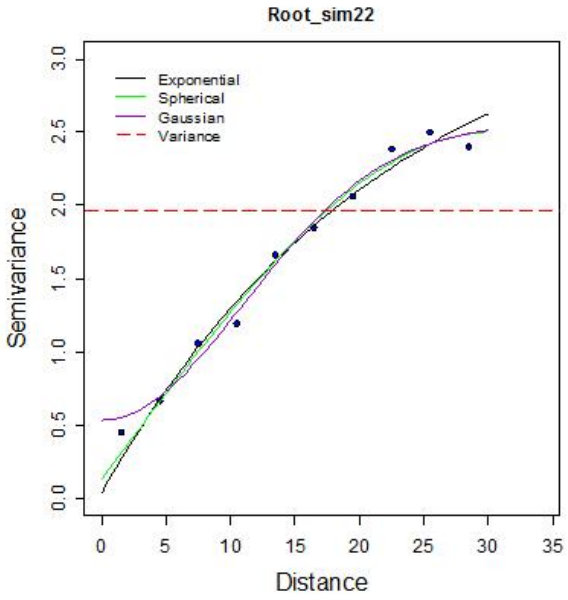

Figura 4.47: Propuestas de modelos de variograma raízD1MAc100

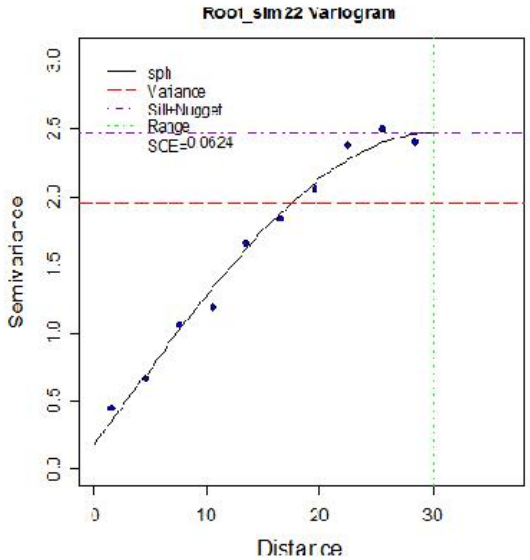

Figura 4.48: Modelo de variograma elegido para raizD1MAc100

Para validar el modelo se realiza la validación cruzada (Cuadro 4.21) . Dentro del cuadro de validación cruzada, se observa en las estadísticas del error que la media es casi cero y la varianza es cercana a uno, los cuales son indicadores de un modelo adecuado. En el análisis de residuales se realiza el gráfico de valores reales contra estimados (Figura 4.49), donde los datos se ajustan a una línea de 45° y están dispersos alrededor del cero. Para verificar normalidad en los residuales se realiza el histograma (Figura 4.50) el cual muestra una distribución muy cercana a la normal y se confirma con el q-q plot (Figura 4.51) que también es una línea de 45°.

Nombre	**raízD1MRc100**	**Estimados**	**Error**
Número total	100	100	100
Distancia max	51.73973328	51.73973328	51.73973328
Distancia min	1	1	1
Media	1.709	1.719	-0.009433
Varianza	1.963891432	1.4213816	0.62719924
Desviación estándar	1.401389108	1.19221709	0.79195911
Coeficiente var	0.819946761	0.69373208	83.9522092
Rango min	0	-0.1967	-2.095
1er cuantil	0	0.5729	-0.3922
Mediana	1.91	1.732	0.006917
3er cuantil	3.078	2.735	0.4921
Máximo rango	3.835	3.568	2.11
Asimetría	-0.043965415	-0.09690414	-0.24092822
Curtosis	1.488674398	1.68066249	3.35803346

Cuadro 4.21: Validación cruzada raízD1MAc100

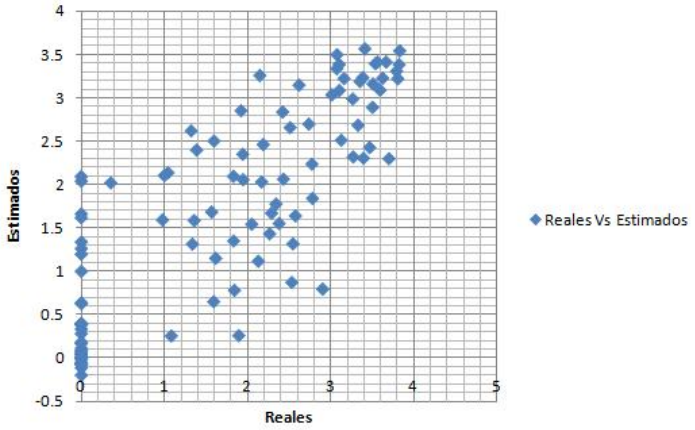

Figura 4.49: Gráfico de valores reales contra estimados de raízD1MAc100

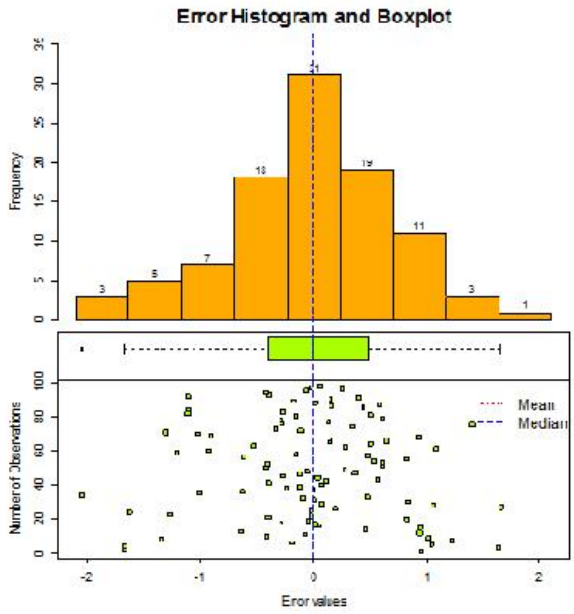

Figura 4.50: Histograma de residuales de raízD1MAc100

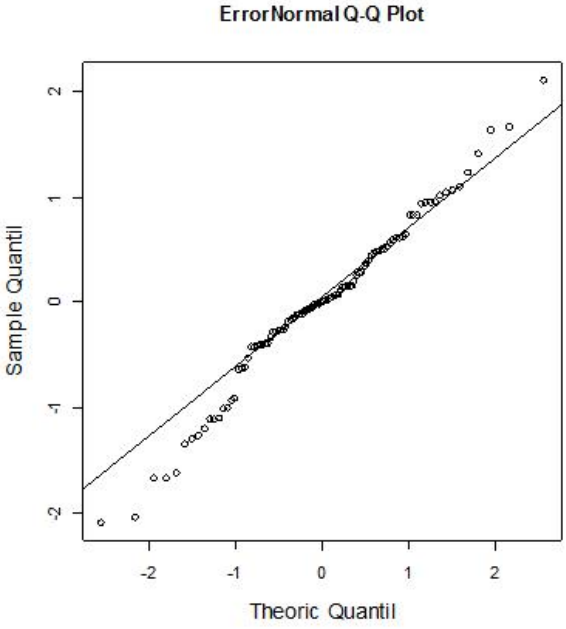

Figura 4.51: Q-Q plot de residuales de raízD1MAc100

Una vez aceptado el modelo, se procede a realizar la estimación con Kriging y se muestra el mapa de valores estimados (Figura 4.52) espaciados a 1km por 1km.

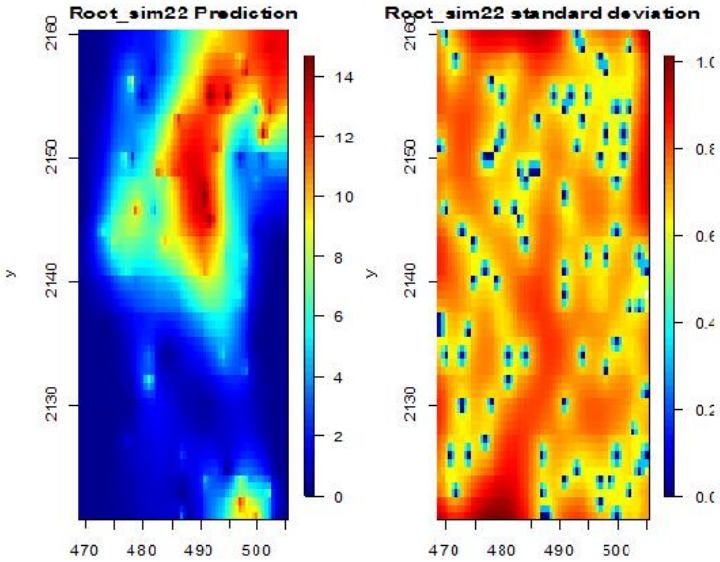

Figura 4.52: Mapa de estimaciones con kriging de raízD1MAc100

4.6. Base de datos del tipo 1 con muestreo combinado y 100 observaciones (D1MCc100)

Para el escenario de D1MCc100, se tiene un muestreo combinado el cual permite tener suficientes observaciones obtenidas aleatoriamente sobre toda la región. Se inicia el análisis exploratorio de datos con el gráfico de la distribución espacial de los datos (Figura 4.53) y sus estadísticas básicas (Cuadro 4.22).

Dentro de las estadísticas básicas se observa que la media es mayor que la mediana, lo que indica una asimetría positiva y se corrobora con el coeficiente de asimetría. El 75 % de los datos se encuentra en un intervalo de [0, 8.18] mientras que el restante 25 % se encuentra en un intervalo de (8.18, 21.29].

Se procede a realizar el histograma (Figura 4.54) y q-q plot (Figura 4.55). El histograma muestra una asimetría positiva y el gráfico de caja y brazos muestra un dato atípico. Además, se observa que el 50 % de la información se encuentra sobre el primer intervalo del histogrma. El q-q plot muestra que no cumple con normalidad.

Nombre	Estadísticas
Número total	100
Distancia max	51.623638
Distancia min	1
Media	4.797
Varianza	31.8878283
Desviación estándar	5.646930875
Coeficiente var	1.177165993
Rango min	0
1er cuantil	0
Mediana	2.484
3er cuantil	8.183
Máximo rango	21.29
Asimetría	1.046364688
Curtosis	3.097987317

Cuadro 4.22: Estadísticas básicas D1MCc100

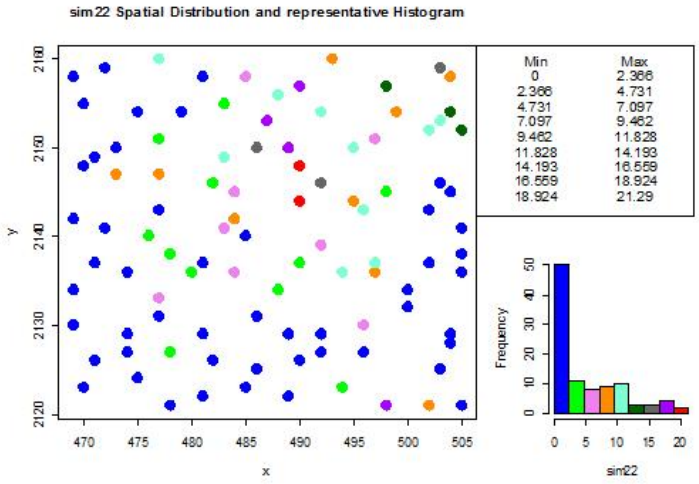

Figura 4.53: Distribución de D1MCc100

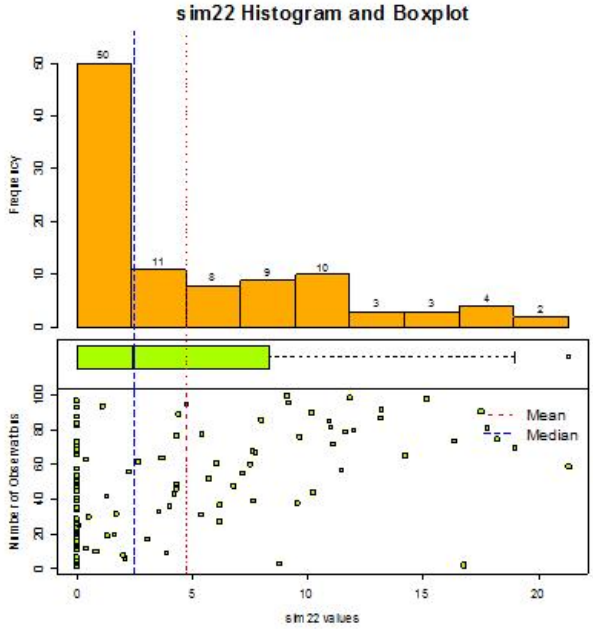

Figura 4.54: Histograma de D1MCc100

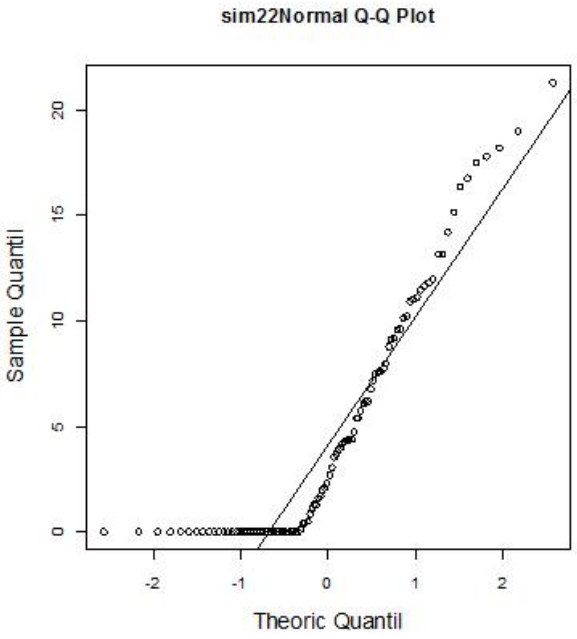

Figura 4.55: Q-Q plot de D1MCc100

Debido a que no cumple con normalidad, se realiza una transformación de raíz cuadrada para disminuir la escala y verificar si mejora la asimetría y distribución de los datos. Se realiza nuevamente el análisis exploratorio de datos. Dentro de las estadísticas básicas (Cuadro 4.23), se observa una mejora considerable sobre la asimetría de la muestra. La media sigue siendo mayor que la mediana pero ahora son muy cercanas. El 75 % de los datos se encuentran en el intervalo [0, 2.86], mientras que el otro 25 % se encuentra en el intervalo (2.86,4.61]. Además la varianza se redujo considerablemente.

Se realiza el histograma (Figura 4.57) y aunque la muestra no se observa normal, la asimetría se redujo considerablemente y la distribución se encuentra más centrada. Además el gráfico de caja y brazos no muestra datos atípicos. El q-q plot (Figura 4.58) no cumple normalidad, pero es suficiente con que la muestra sea simétrica para continuar con la metodología.

Por lo tanto, se realiza el gráfico de la variable con respecto a sus coordenadas (Figura 4.59) para observar alguna tendencia y el gráfico de datos (Figura 4.60) para observar estacionariedad.

Nombre	**Estadísticas**
Número total	100
Distancia max	51.623638
Distancia min	1
Media	1.609
Varianza	2.22934971
Desviación estándar	1.493100703
Coeficiente var	0.927767721
Rango min	0
1er cuantil	0
Mediana	1.575
3er cuantil	2.86
Máximo rango	4.614
Asimetría	0.261890838
Curtosis	1.652025286

Cuadro 4.23: Estadísticas básicas raízD1MCc100

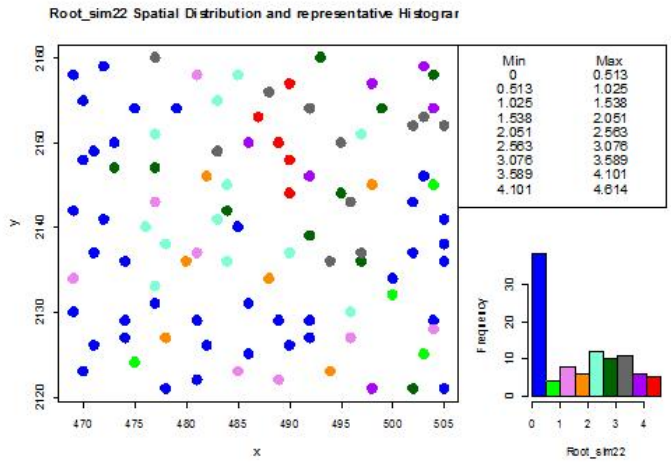

Figura 4.56: Distribución de raízD1MCc100

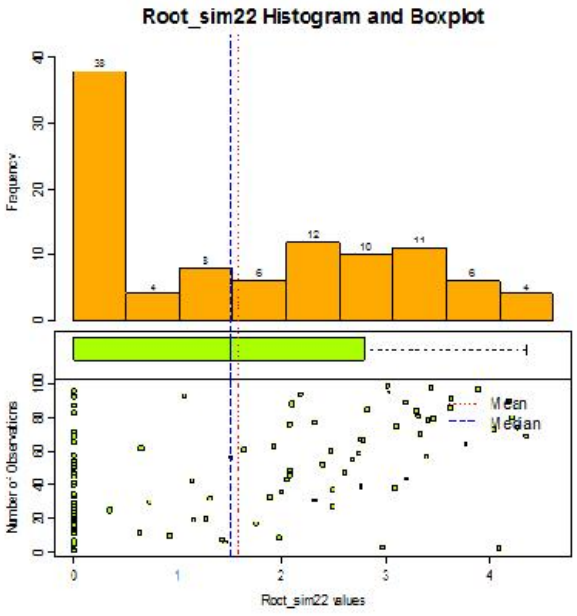

Figura 4.57: Histograma de raízD1MCc100

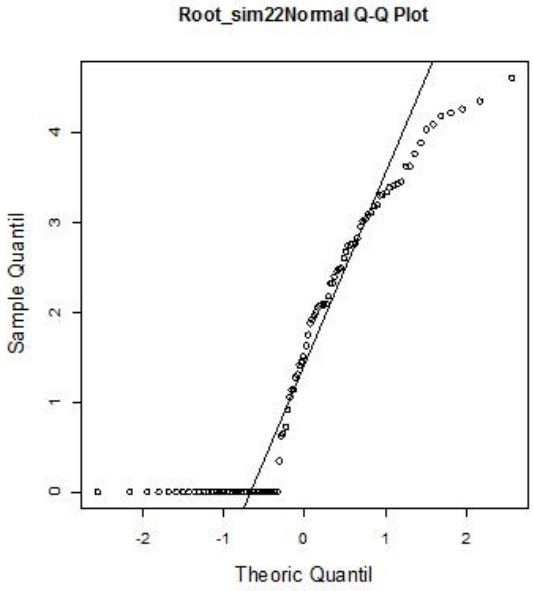

Figura 4.58: Q-Q plot de raízD1MCc100

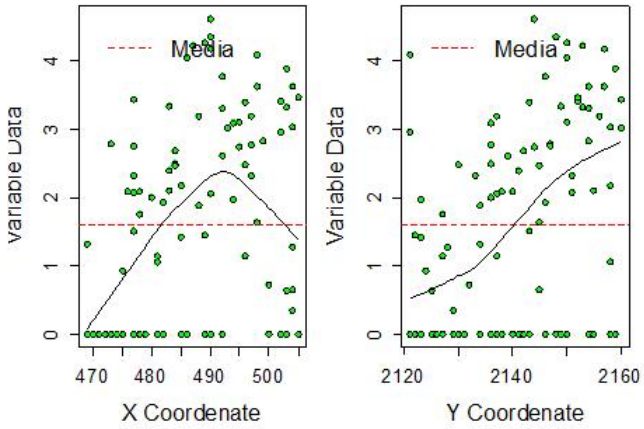

Figura 4.59: Gráfico respecto a las coordenadas de raízD1MCc100

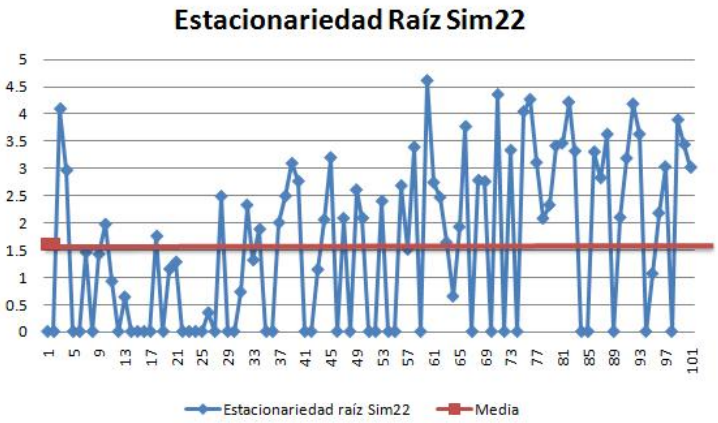

Figura 4.60: Gráfico de estacionariedad de raízD1MCc100

En el gráfico con respecto a las coordenadas (Figura 4.59) se observa una ligera tendencia con respecto al eje Y, la cual se verifica con el variograma. El gráfico de estacionariedad (Figura 4.60) no tiene indicios contundentes de que no se cumpla, por lo que se continúa con el análisis variográfico. Se realiza el variograma adireccional.

El variograma adireccional (Cuadro 4.24 Figura 4.61) no muestra un comportamiento de h^2, no se confirma la tendencia y por lo tanto no se realiza ningún proceso para corregirla. Se grafican los variogramas en 4 direcciones y se verifican los alcances para revisar si existe anisotropía.

Disctancia max	51.623638
Distancia min	1
Dirección	0°
Tolerancia	90°
Intervalos	10
Distancia Lag	3

Cuadro 4.24: Variograma adireccional de raízD1MCc100

Disctancia max	51.623638
Distancia min	1
Dirección	0°,45°,90°,135°
Tolerancia	22.5°
Intervalos	10
Distancia Lag	3

Cuadro 4.25: Variograma 4 direcciones raízD1MCc100

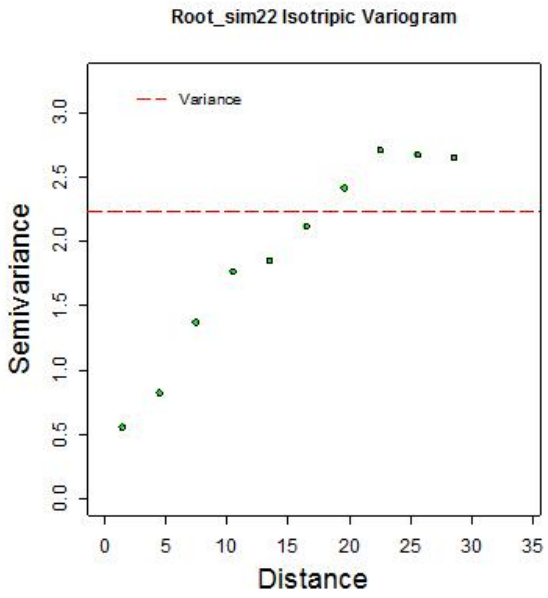

Figura 4.61: Variograma adireccional de raízD1MCc100

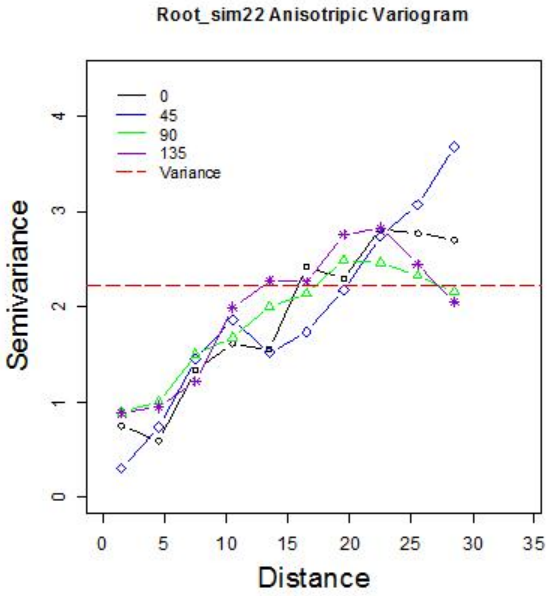

Figura 4.62: Variograma en 4 direcciones de raízD1MCc100

En el variograma en 4 direcciones (Cuadro 4.25 Figura 4.62) se observa que los alcances no son a una distancia diferente, por lo que no hay indicios de anisotropía. También se realiza el mapa de anisotropía, el cual no muestra elipses, con lo que se confirma que no hay anisotropía y se proponen sólo modelos isotrópicos.

Dentro de las propuestas de modelo de variograma (Cuadro 4.26 Figura 4.64), el modelo que mejor se adapta es el modelo esférico, por lo que se le realiza un ajuste visual para mejorarlo y llegar al modelo elegido (Cuadro 4.27 Figura4.65)

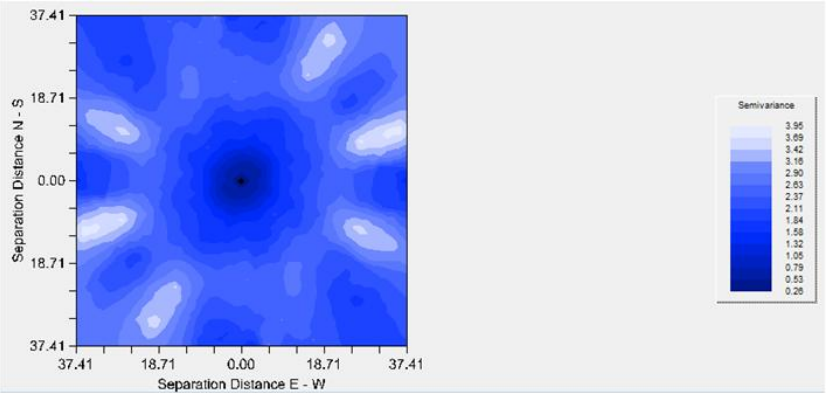

Figura 4.63: Mapa de anisotropía de raízD1MCc100

Modelo	**Nugget**	**Sill+Nugget**	**Rango**	**SCE**
Exponencial	0.07	3.22	14.68	0.11
Esférico	0.37	2.69	27.53	0.0699
Gaussiano	0.8	2.75	14.41	1.8

Cuadro 4.26: Propuestas para modelos de variograma raízD1MCc100

Modelo	**Nugget**	**Sill+Nugget**	**Rango**	**SCE**
Esférico	0.27	2.68	27	0.0809

Cuadro 4.27: Modelo de variograma elegido raízD1MCc100

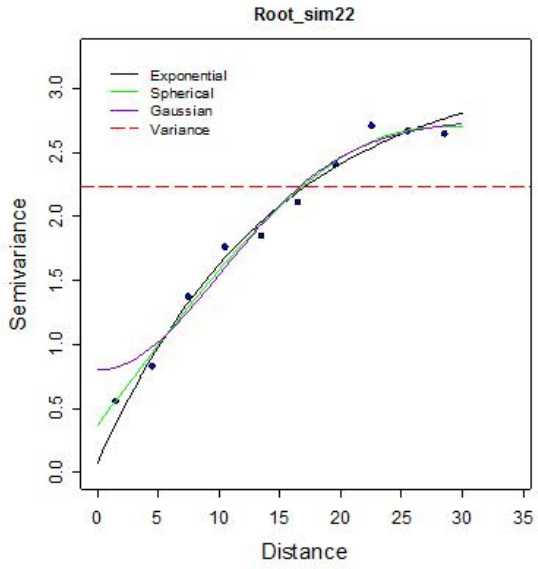

Figura 4.64: Propuestas de modelo de variograma para raízD1MCc100

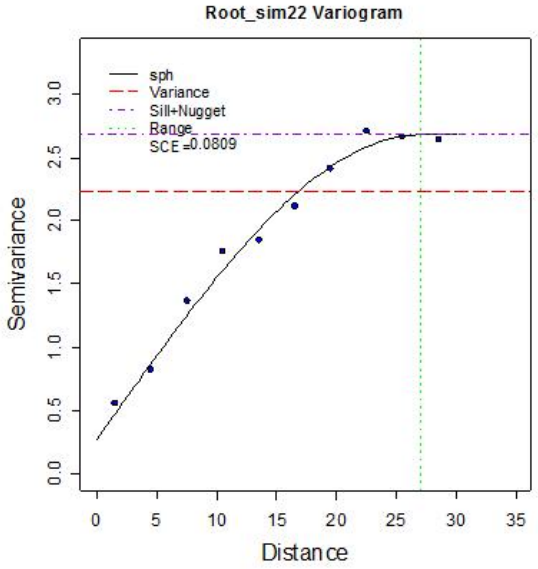

Figura 4.65: Modelo de variograma elegido para raízD1MCc100

Una vez elegido el modelo se verifica que cumpla los supuestos de la metodología realizando la validación cruzada y el análisis de residuales para saber si es adecuado. Dentro del Cuadro 4.28, las estadísticas del error muestran que la media es casi cero y la varianza es muy cercana a uno, lo cual es ideal. Además se grafican los valores reales contra estimados (Figura 4.66) y se observa que se apegan a una línea de 45° y se distribuyen de manera aleatoria alrededor del cero.

Nombre	**raízD1MRc100**	**Estimados**	**Error**
Número total	100	100	100
Distancia max	51.623638	51.623638	51.623638
Distancia min	1	1	1
Media	1.609	1.614	-0.004343
Varianza	2.22934971	1.46120092	0.81926101
Desviación estándar	1.493100703	1.20880144	0.90513038
Coeficiente var	0.927767721	0.74909106	208.393209
Rango min	0	-0.1621	-2.598
1er cuantil	0	0.5395	-0.5727
Mediana	1.575	1.248	-0.1874
3er cuantil	2.86	2.61	0.6173
Máximo rango	4.614	4.152	3.007
Asimetría	0.261890838	0.40349022	0.40032574
Curtosis	1.652025286	1.88681068	3.83613942

Cuadro 4.28: Validación cruzada raízD1MCc100

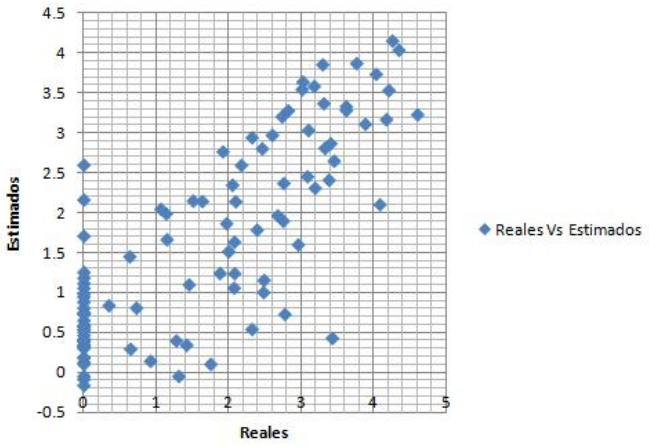

Figura 4.66: Valores reales contra estimados

El histograma de errores (Figura 4.67) muestra una distribución cercana a la normal, y también se observa un q-q plot de los errores (Figura 4.68) suficientemente adecuado para que cumpla con los supuestos.

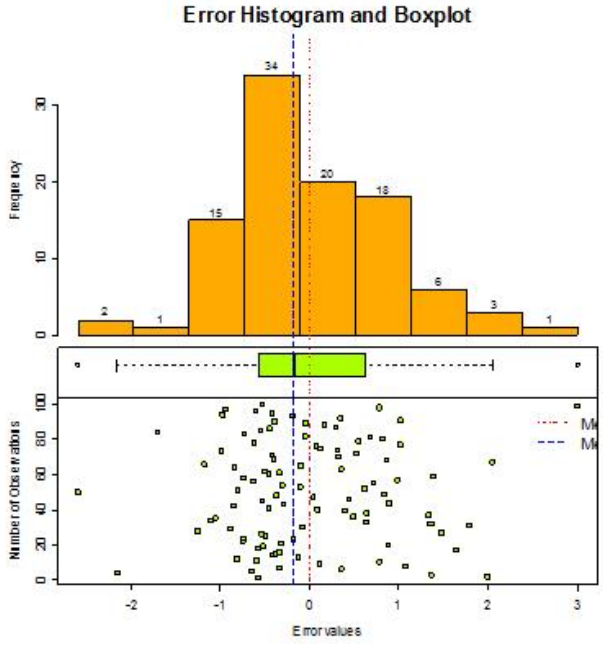

Figura 4.67: Histograma de residuales de raízD1MCc100

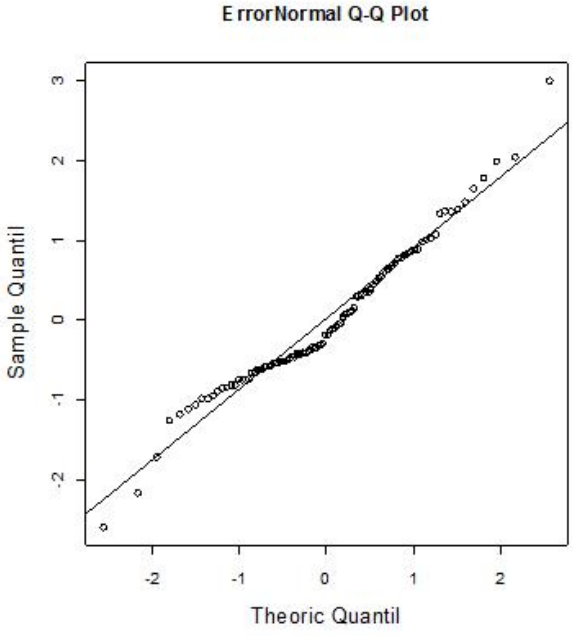

Figura 4.68: Q-Q plot de residuales de raízD1Mcc100

Por lo tanto, una vez que el modelo fue elegido, es adecuado y cumple con los supuestos de la metodología se procede a realizar la estimación con Kriging y se obtiene un mapa de valores estimados a 1km por 1km (Figura 4.69).

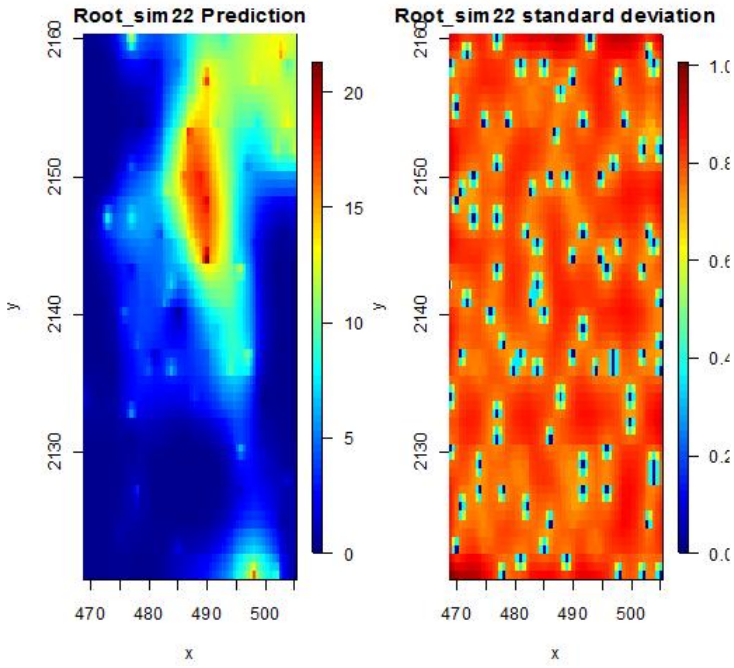

Figura 4.69: Mapa de estimaciones con kriging de raízD1MCc100

4.7. Conclusiones de los modelos de datos 1

- La base de datos del tipo 1 no presenta tendencia o anisotropía, por lo que los modelos son isotrópicos para todos los escenarios.
- Todos los modelos presentaron asimetría positiva.
- Es necesaria una transformación de raíz cuadrada en todos los casos de estudio para disminuir la asimetría que presentan los datos y cumplir con los supuestos del modelo.
- No fue necesario eliminar los datos atípicos debido a que la transformación reduce la escala y permite ya no existan datos atípicos.
- El lag de distancia para calcular los variogramas varía dependiendo del espaciamiento que existe entre muestras.
- Casi todos los modelos son ajustados con un modelo esférico y parámetros del variograma muy cercanos.
- Se realizó un ajuste visual con criterio propio para los modelos de variograma presentados.

- Se realizó validación cruzada en todos los casos y los modelos se aceptaron mediante el criterio de media de los errores cercana a cero y varianza de los errores cercana a 1, así como residuales que asemejen una distribución normal.
- En todos los casos se estimó con Kriging ordinario.
- El modelo D1MAc100 es el que mejor mapa de estimación presenta.
- Cuando se tienen 400 observaciones, el muestreo utilizado pasa a 2do grado de importancia debido a que la información cubre una gran parte de la región.
- Los escenarios con 36 observaciones son mucho más complejos para ajustarles un modelo adecuado y en ocasiones se logra únicamente con ajuste visual.
- Los escenarios con 100 observaciones son los más representativos, ya que tienen una cantidad suficiente de información para estimar de manera adecuada y no tienen tantos datos que la estimación requiera de mucho tiempo.
- Los modelos son adecuados dependiendo de la precisión con la que se realice el ajuste del modelo, así como la información que se tenga a la mano.

Capítulo 5

Aplicación de la metodología a los Datos del tipo 2

5.1. Proceso de estimación de los datos originales del tipo 2

Dentro de las recomendaciones para la aplicación de los procesos de la metodología de geoestadística, es importante mostrar el ajuste al cual se pretende llegar para cada uno de los escenarios propuestos dentro del grupo de datos del tipo 2 (D2c1480). Debido a ello, se muestra el procedimiento de ajuste del modelo de variograma y su estimación considerando el total de los datos originales del tipo 2.

En principio, se muestra la dispersión de los datos del tipo 2 (Figura 5.1 y sus estadísticas básicas (Cuadro 5.1). En el gráfico de dispersión de los datos se muestran las regiones donde las observaciones muestran un valor mayor. Debido a que son los datos originales sería ideal obtener para cada uno de los escenarios estimaciones muy cercanas a esta distribución de datos. Por lo que será la referencia contra la cual se compararán los escenarios de los datos del tipo 2.

Dentro de las estadísticas básicas destaca que la asimetría es positiva, tanto por el coeficiente de asimetría como por la media que es mayor que la mediana. Además se muestra que el 50 % de la información se encuentra en un intervalo de [0.25, 1.6) y el restante 50 % se encuentra entre (1.6,27.5].

Se realiza el histograma y se observa la notable asimetría que existe así como los 174 datos atípicos. También con el q-q plot es contundente que no se tiene una distribución normal.

Nombre	Estadísticas
Número total	1480
Distancia max	53.07541804
Distancia min	1
Media	4.993
Varianza	42.79093157
Desviación estándar	6.54147782
Coeficiente var	1.310229811
Rango min	0.25
1er cuantil	0.25
Mediana	1.588
3er cuantil	6.553
Máximo rango	27.5
Asimetría	1.458596438
Curtosis	4.092784453

Cuadro 5.1: Estadísticas básicas D2c1480

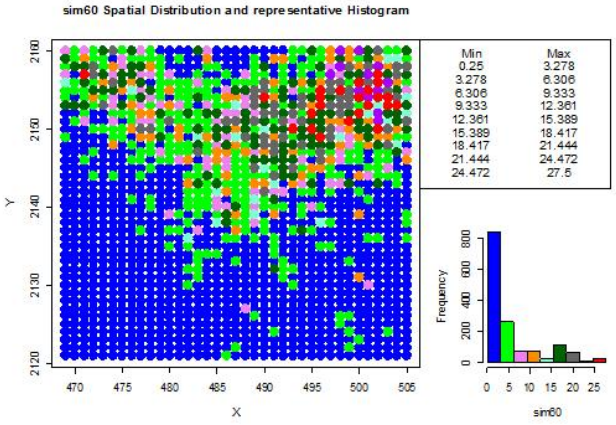

Figura 5.1: Distribución de D2c1480

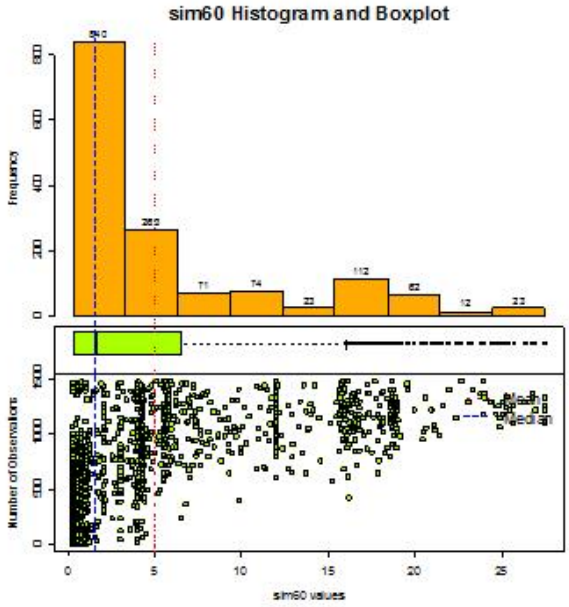

Figura 5.2: Histograma de D2c1480

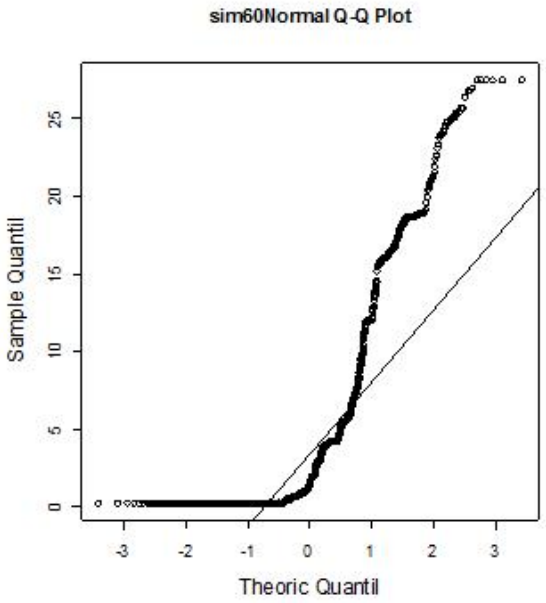

Figura 5.3: Q-Q plot de D2c1480

Debido a que la asimetría puede generar conflictos en la estimación, es necesaria una transformación que reduzca la escala y mejore la distribución de los datos, así como los valores atípicos que se observan. Se realiza una transformación logarítmica y se vuelve a realizar el análisis exploratorio de datos.

Dentro de las estadísticas básicas (Cuadro 5.2) se reduce la escala y el 75 % de los datos se encuentran en el intervalo de [-1.386, 1.88], mientras que el restante 25 % se encuentra en el intervalo (1.88, 3.315], la dispersión de los datos (Figura 5.4) permite ver las regiones donde se encuentran los valores más altos de la variable. La media sigue siendo mayor que la mediana pero son más cercanas, es decir, la distribución se encuentra más centrada.

El histograma (Figura 5.5) y q-q plot (Figura 5.6) muestran que la asimetría es considerablemente menor y la distribución no es normal pero es suficientemente simétrica. La media y mediana son más cercanas y centradas y ya no se tienen datos atípicos.

Nombre	**Estadísticas**
Número total	1480
Distancia max	53.07541804
Distancia min	1
Media	0.4862
Varianza	2.7107032585
Desviación estándar	1.648342375
Coeficiente var	3.390052669
Rango min	-1.386
1er cuantil	-1.386
Mediana	0.4625
3er cuantil	1.88
Máximo rango	3.314
Asimetría	0.142770105
Curtosis	1.474603909

Cuadro 5.2: Estadísticas básicas logD2c1480

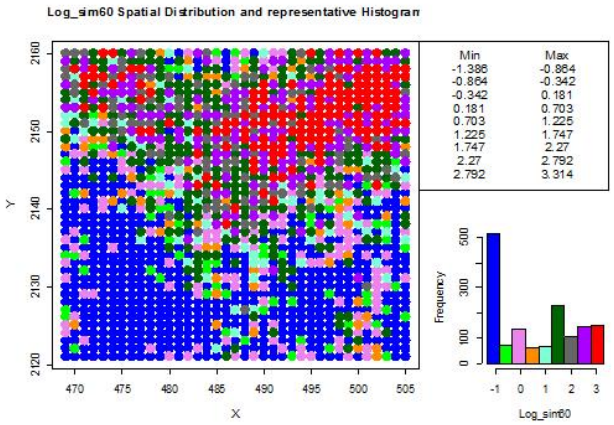

Figura 5.4: Distribución de logD2c1480

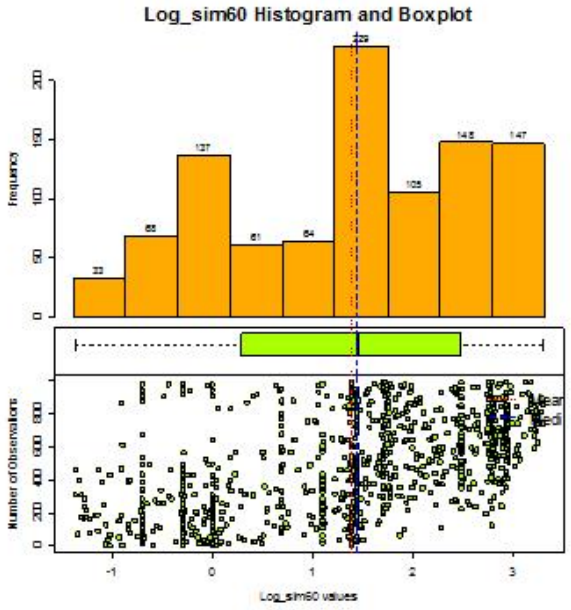

Figura 5.5: Histograma de logD2c1480

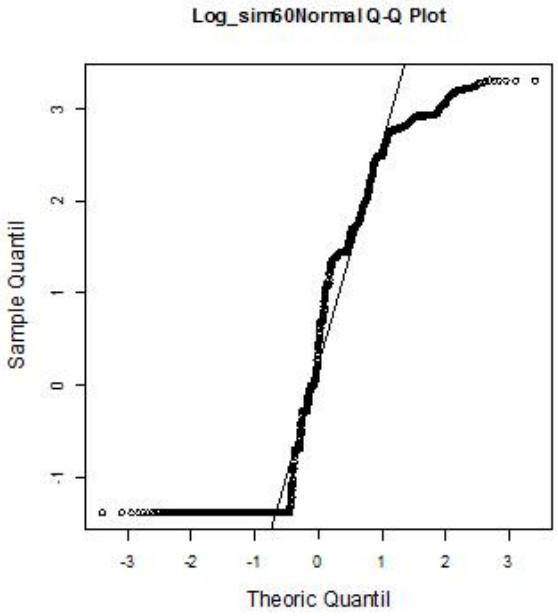

Figura 5.6: Q-Q plot de logD2c1480

Ahora se observa el grafico con respecto a las coordenadas (Figura 5.7) el cual muestra si existe tendencia. Se muestra una tendencia considerable con respecto a la variable y. Se revisa también el gráfico para estacionariedad (Figura 5.8) en el cual se observa también que no cumple una media constante. Por lo tanto, la muestra tiene tendencia y no cumple completamente estacionariedad.

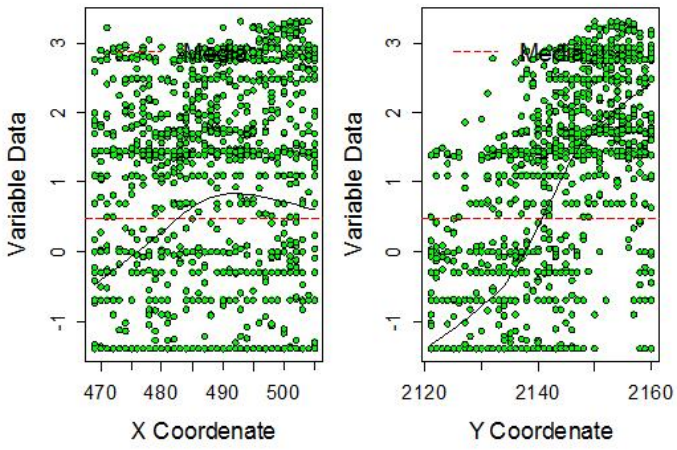

Figura 5.7: Gráfico con respecto a las coordenadas de logD2c1480

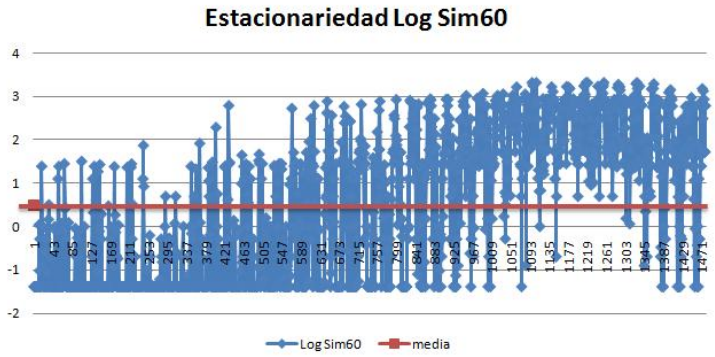

Figura 5.8: Gráfico de estacionariedad de logD2c1480

Posteriormente, es necesario realizar el análisis variográfico para decidir si se corrige o no la tendencia. Se inicia con el variograma adireccional (Cuadro 5.3 Figura 5.9).

Debido a que se cuenta con una base de datos tan grande, es posible calcular los puntos del variograma con el máximo de intervalos (25) y el mínimo del lag manteniendo bastante confiabilidad en los puntos calculados. Se observa que el variograma sí muestra una tendencia lineal.

Sin embargo, también es necesario revisar el variograma en 4 direcciones (Cuadro 5.4 Figura 5.10) para verificar que la tendencia observada no sea producto de una posible anisotropía.

Disctancia max	53.07541804
Distancia min	1
Dirección	0°
Tolerancia	90°
Intervalos	25
Distancia Lag	1

Cuadro 5.3: Variograma adireccional de logD2c1480

Disctancia max	53.07541804
Distancia min	1
Dirección	0°,45°,90°,135°
Tolerancia	22.5°
Intervalos	25
Distancia Lag	1

Cuadro 5.4: Variograma 4 direcciones logD2c1480

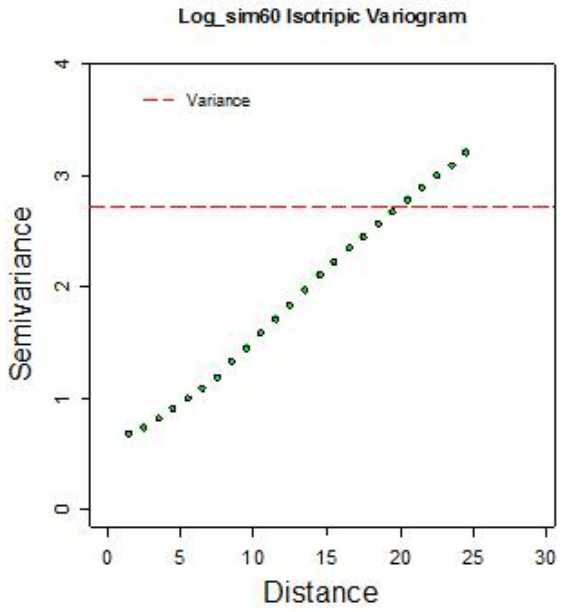

Figura 5.9: Variograma adireccional de logD2c1480

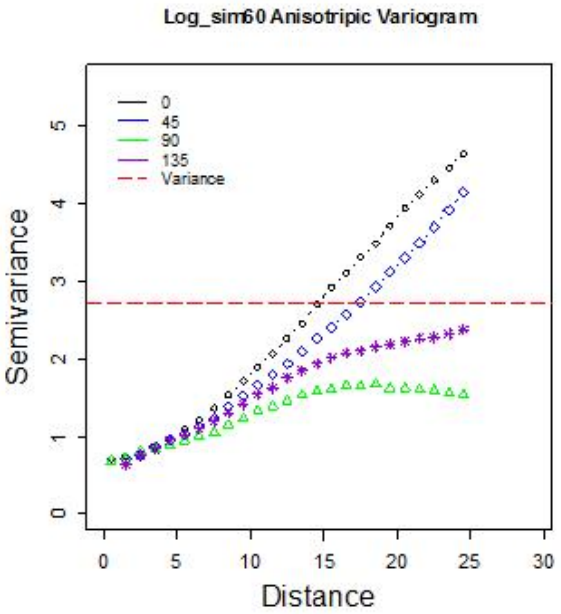

Figura 5.10: Variograma en 4 direcciones de logD2c1480

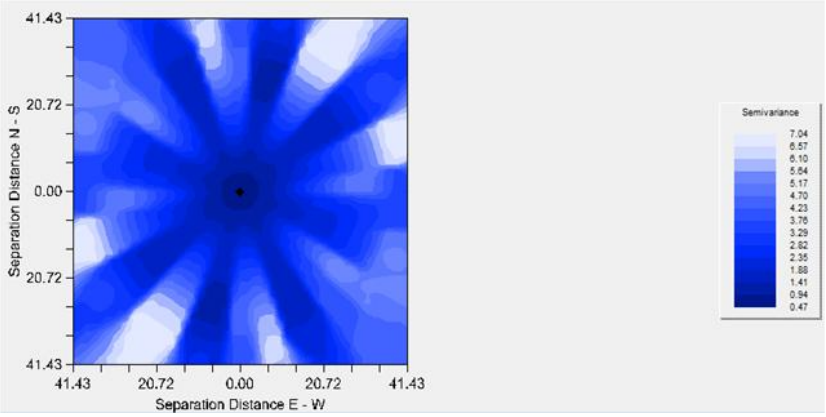

Figura 5.11: Mapa de anisotropía de logD2c1480

El mapa de anisotropía (Figura 5.11) confirma que la linealidad no depende de la dirección por lo que se procede a corregir la tendencia de 1er grado mediante el modelo de tendencia mencionado en la metodología de geoestadística.

Debido a que no es recomendable realizar varias transformaciones, se inicia nuevamente con los datos originales del tipo 2 (sim60) y se realiza la transformación al modelo de tendencia. Se procede a iniciar nuevamente el análisis exploratorio.

Coeficiente	**Valor**
Intercepto	-793.2596
Coeficiente de x	0.138419578
Coeficiente de y	0.341435124

Cuadro 5.5: Modelo de Tendencia de 1er grado modificada D2c1480

El histograma (Figura 5.12) y q-q plot (Figura 5.13) mejoran considerablemente y muestran que la distribución de los datos es adecuada y no presenta asimetría significativa.

El gráfico con respecto a las coordenadas (Figura 5.14) ya no muestra tendencia y la estacionariedad es estabilizada (Figura 5.15). Por lo tanto, se procede a realizar el análisis variográfico.

Nombre	Estadísticas
Número total	1480
Distancia max	53.07541804
Distancia min	1
Media	1.43x10⁻16
Varianza	25.06070295
Desviación estándar	5.006066614
Coeficiente var	3.5x10¹6
Rango min	-11.68
1er cuantil	-3.244
Mediana	-0.7208
3er cuantil	2.189
Máximo rango	18.43
Asimetría	0.763274258
Curtosis	3.778914668

Cuadro 5.6: Estadísticas básicas modelo sin tendencia D2c1480

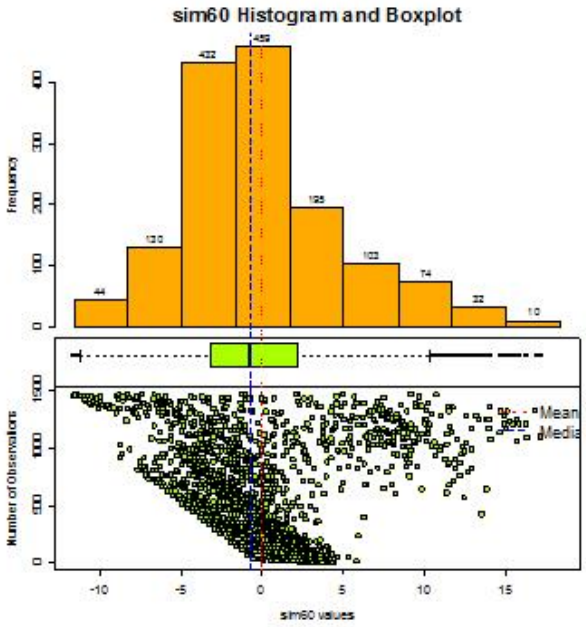

Figura 5.12: Histograma del modelo sin tendencia de D2c1480

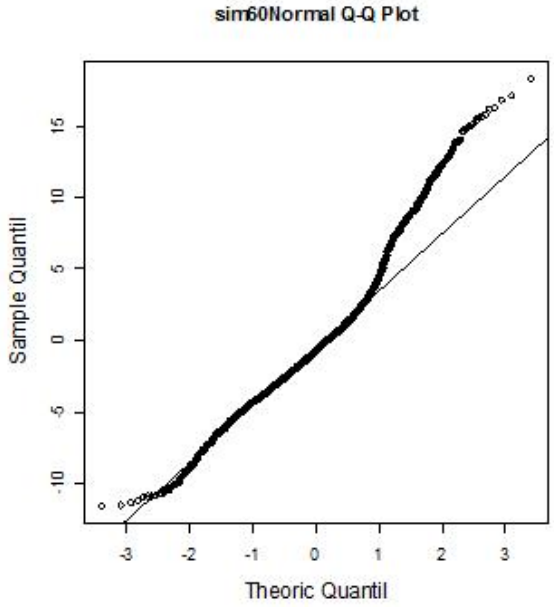

Figura 5.13: Q-Q plot del modelo sin tendencia de D2c1480

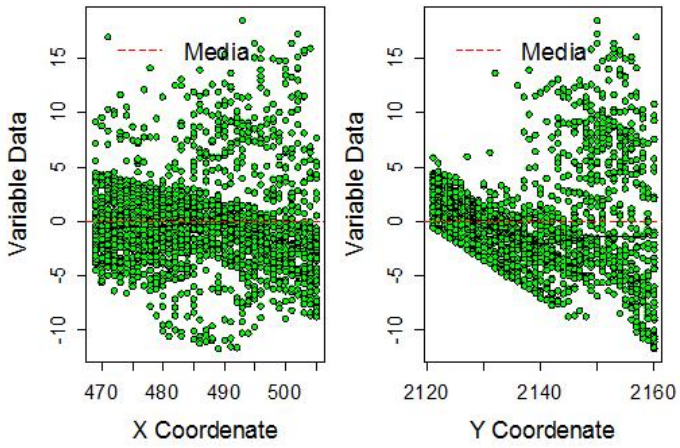

Figura 5.14: Gráfico respecto a las coordenadas, modelo s/tendencia D2c1480

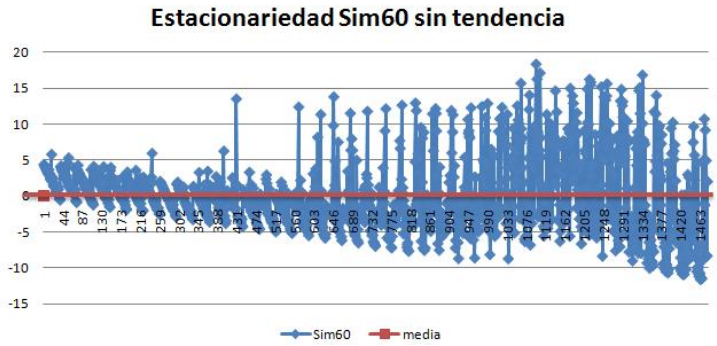

Figura 5.15: Estacionariedad del modelo sin tendencia D2c1480

El variograma adireccional (Cuadro 5.7 Figura 5.16) es adecuado y muestra un buen comportamiento para el ajuste. Se verifica nuevamente la anisotropía con el variograma en 4 direcciones (Cuadro 5.8 Figura 5.17) y el mapa de anisotropía (Figura 5.18).

Disctancia max	53.07541804
Distancia min	1
Dirección	0°
Tolerancia	90°
Intervalos	25
Distancia Lag	1

Cuadro 5.7: Variograma adireccional del modelo sin tendencia D2c1480

Disctancia max	53.07541804
Distancia min	1
Dirección	0°,45°,90°,135°
Tolerancia	22.5°
Intervalos	25
Distancia Lag	1

Cuadro 5.8: Variograma 4 direcciones modelo sin tendencia D2c1480

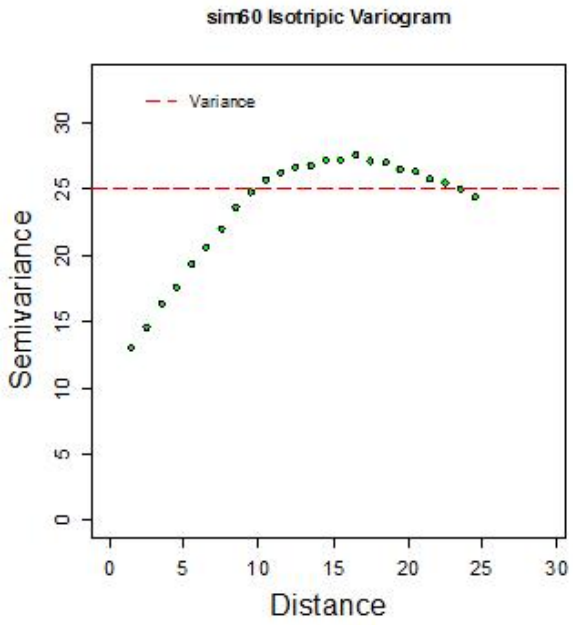

Figura 5.16: Variograma adireccional sin tendencia D2c1480

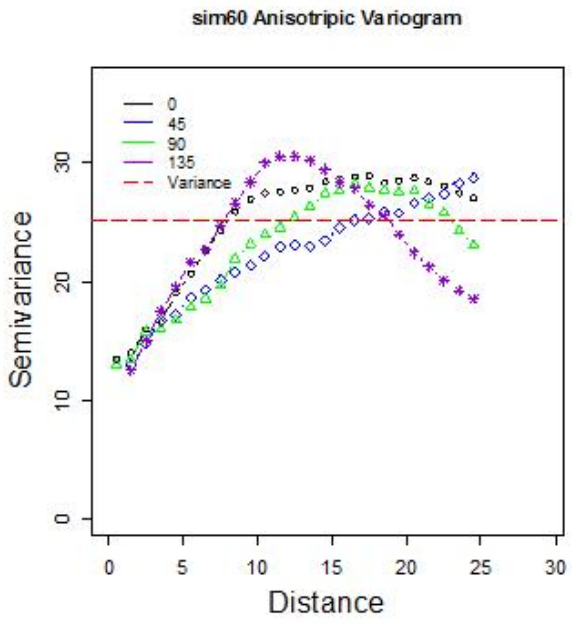

Figura 5.17: Variograma en 4 direcciones sin tendencia D2c1480

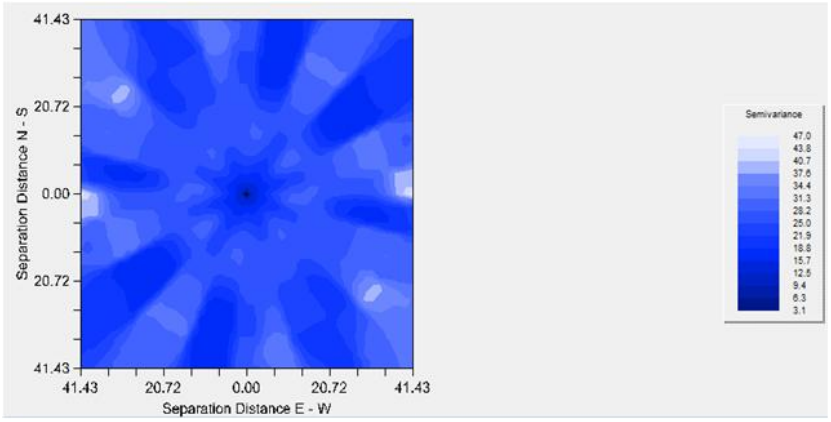

Figura 5.18: Mapa de anisotropía modelo sin tendencia D2c1480

Debido a que no se observa anisotropía, se procede a realizar el ajuste del modelo presentando las propuestas (Cuadro 5.9 Figura 5.19). Con la menor suma de cuadrados del error y observando visualmente el comportamiento de los puntos es adecuado aceptar el modelo esférico como el que mejor se adapta. Se realiza un mínimo ajuste visual sobre la propuesta y se obtiene el modelo de variograma elegido (Cuadro 5.10 Figura 5.20).

Modelo	**Nugget**	**Sill+Nugget**	**Rango**	**SCE**
Exponencial	2.59	26.71	4.01	32.6
Esférico	9.54	26.42	13.18	11.9
Gaussiano	12.34	26.46	6.57	194

Cuadro 5.9: Propuestas de modelos de variograma sin tendencia D2c1480

Modelo	**Nugget**	**Sill+Nugget**	**Rango**	**SCE**
Esférico	9.8	26.4	13.2	11.8

Cuadro 5.10: Modelo de variograma elegido sin tendencia D2c1480

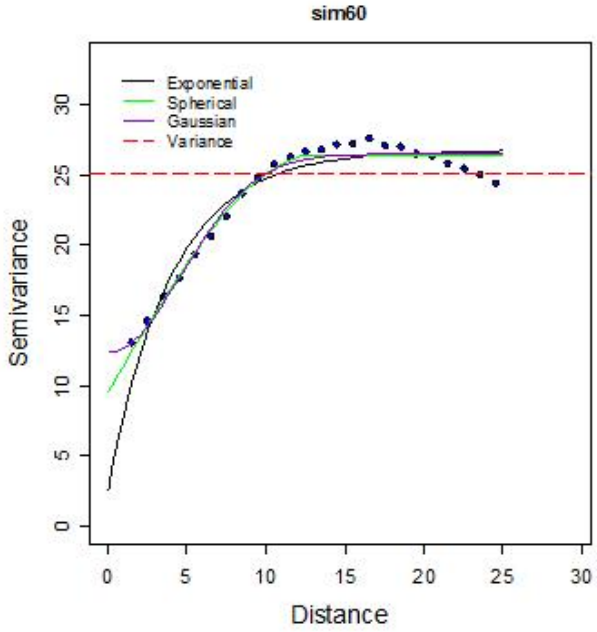

Figura 5.19: Propuestas de modelos de variograma sin tendencia D2c1480

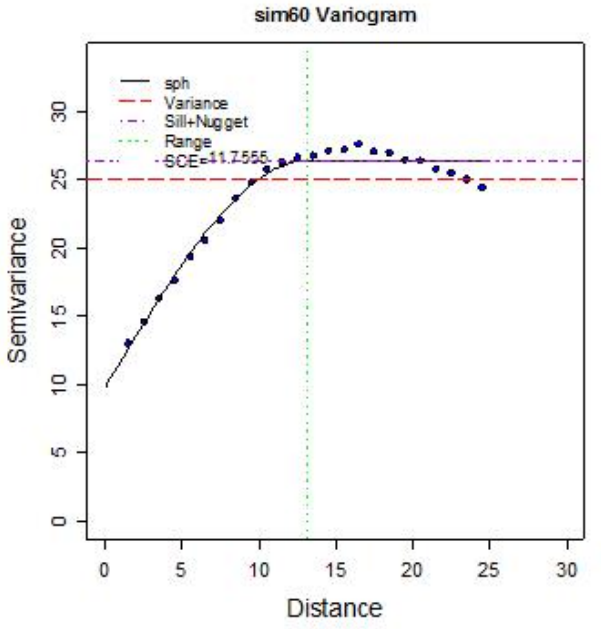

Figura 5.20: Modelo de variograma elegido sin tendencia D2c1480

Una vez elegido el modelo y ajuste del variograma se realiza la validación cruzada (Cuadro 5.11) para verificar si es adecuado. Con la validación cruzada se muestra que los errores tienen media cero y los estimados tienen una distribución parecida a los valores reales. También se realiza el análisis gráfico de residuales para confirmar.

Nombre	**s/tend D2c1480**	**Estimados**	**Error**
Número total	1480	1480	1480
Distancia max	53.07541804	53.07541804	53.07541804
Distancia min	1	1	1
Media	1.43x10⁻16	0.00441	-0.00441
Varianza	25.06070295	12.37894947	13.53733338
Desviación estándar	5.006066614	3.518373129	3.679311536
Coeficiente var	3.5x10¹6	797.8237045	834.3179795
Rango min	-11.68	-10.38	-16.11
1er cuantil	-3.244	-2.34	-1.361
Mediana	-0.7208	-0.1133	-0.1697
3er cuantil	2.189	2.035	0.7286
Máximo rango	18.43	11.19	14.4
Asimetría	0.763274258	0.299298169	0.54393444
Curtosis	3.778914668	3.421619619	5.672551506

Cuadro 5.11: Validación cruzada modelo sin tendencia D2c1480

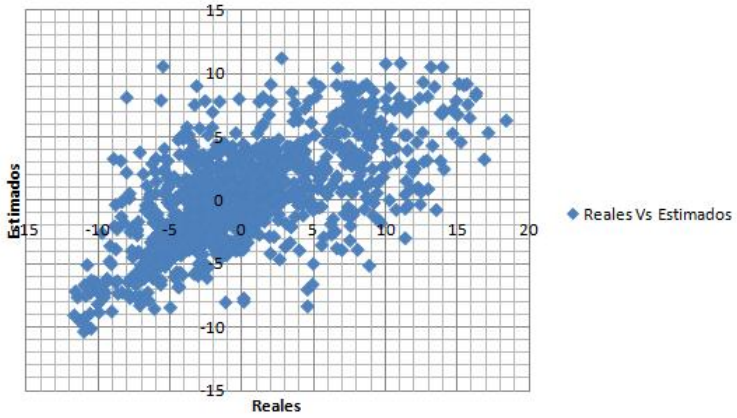

Figura 5.21: Valores reales contra estimados modelo sin tendencia D2c1480

El gráfico de valores reales contra estimados (Figura 5.21) muestra que se aproximan a una línea de 45°. El histograma (Figura 5.22) y q-q plot (Figura 5.23) muestran que aunque no tienen una distribución normal, los residuos se encuentran centrados y simétricos, por lo que el modelo es aceptado y se procede a realizar la estimación con Kriging.

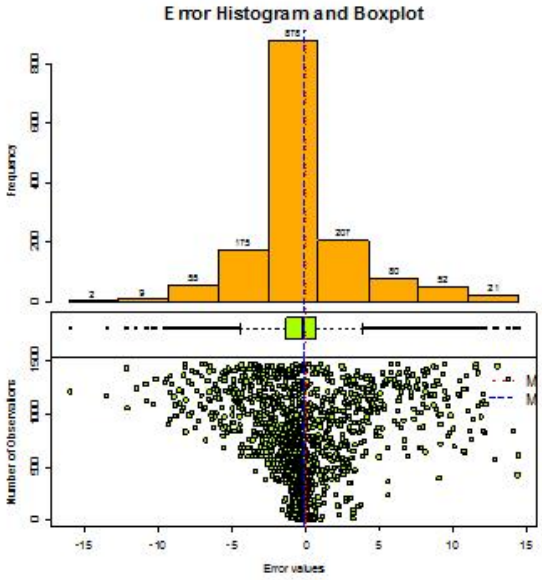

Figura 5.22: Histograma de residuales del modelo sin tendencia D2c1480

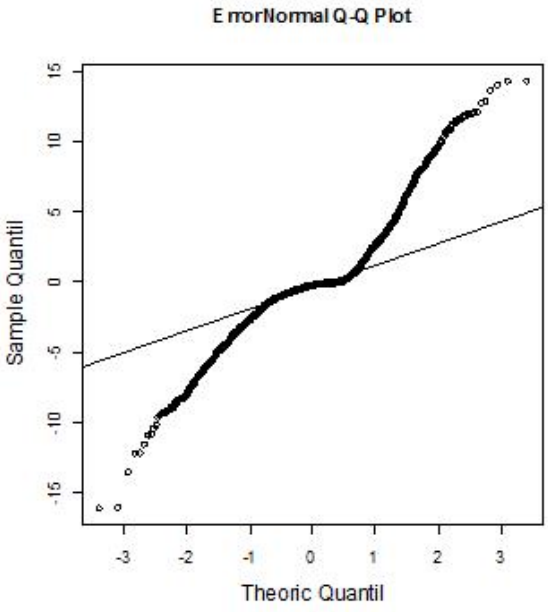

Figura 5.23: Q-Q plot de residuales del modelo sin tendencia D2c1480

Se muestra el mapa de estimaciones con kriging ordinario (Figura 5.24) el cual es semejante a la región mostrada por los datos originales del tipo 2.

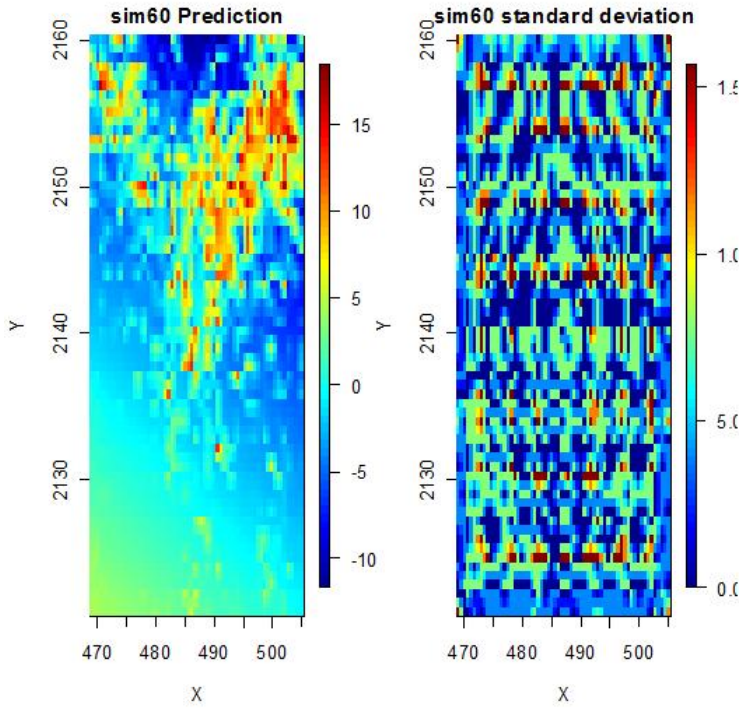

Figura 5.24: Mapa de estimaciones del modelo sin tendencia D2c1480

5.2. Resumen de los modelos de datos 2

Es importante enfatizar las características de los escenarios realizados para los datos del tipo 2. Se muestra un resumen (Figura 5.25) de los modelos ajustados para cada uno de los escenarios y se mencionan las características en común que tienen todos los modelos.

Las siguientes observaciones ocurrieron en todos los escenarios:

1. La distribución de las muestras tiene asimetría positiva.
2. En todos los escenarios de datos 2, la muestra es tratada con una transformación de logaritmo en principio y posteriormente al encontrarse tendencia se regresa a los datos originales del tipo 2 y con ellos se utiliza el modelo de residuos para tratar la tendencia.
3. El caso excluyente es el de D2MAc036 el cual únicamente utilizó la transformación de logaritmo y no el modelo de tendencia.
4. Las bases de datos de cada escenario no cumplen normalidad.

5. Los escenarios presentan indicios de no estacionariedad que se mejoran con la transformación para tendencia.

6. Los escenarios no presentan indicadores de anisotropía.

7. En todos los escenarios los variogramas en 4 direcciones son calculados con el mismo número de intervalos y el mismo lag que en su respectivo variograma adireccional.

8. Todos los modelos pasaron el análisis de residuales, es decir, la línea de 45°, el histograma y q-q plot son muy cercanos a la distribución normal.

9. En todos los modelos se realizó la estimación por medio de Kriging ordinario.

ANÁLISIS ESTRUCTURAL Y VARIOGRÁFICO							
Escenario	Variograma adireccional	Variograma en 4 direcciones	Anisotropía o Tendencia	Modelo	Nugget	Meseta (Sill+Nugget)	Alcance
Datos 2	**inter 25, lag 1**	**lag 1**	**1er grado**	**Esférico**	**9.8**	**26.4**	**13.2**
D2MRc036	inter 10, lag 6	lag 6	1er grado	Esférico	0	11.5	14.5
D2MRc100	inter 12, lag 4	lag 4	1er grado	Esférico	5	27	14
D2MRc400	inter 13, lag 2	lag 2	1er grado	Esférico	9.1	23.5	13.2
D2MAc036	inter 10, lag 3	lag 3	no definida	Exponencial	0	4.5	24
D2MAc100	inter 12, lag 2	lag 2	1er grado	Esférico	8	26.3	15
D2MAc400	inter 13, lag 2	lag 2	1er grado	Esférico	8.3	28	13.5
D2MCc036	inter 10, lag 3	lag 3	1er grado	Esférico	0	28	13.5
D2MCc100	inter 13, lag 2	lag 2	1er grado	Esférico	13.5	32	15.5
D2MCc400	inter 13, lag 2	lag 2	1er grado	Esférico	8.3	28	13.5

Figura 5.25: Información del variograma por escenario de los datos del tipo 2

Información del análisis variográfico de los escenarios de Datos 2:

1. Se observa que los modelos ajustados en casi todos los escenarios son esféricos, excepto en el caso D2MAc036, en el cual es un modelo exponencial.

2. El nugget está en un intervalo de [0, 13.5] para todos los escenarios.

3. La meseta se encuentra en un intervalo de [11.5, 32] en los escenarios con esféricos y 4.5 para el escenario con exponencial.

4. El alcance varía en un intervalo de [13.2, 15.5] para los escenarios esféricos y 24 para el escenario exponencial.

Cabe destacar que el modelo exponencial tiene el máximo alcance, esto es por la forma que tiene el ajuste del modelo con respecto al esférico.

Por otro lado, es importante mencionar que los modelos ajustados se realizaron de manera completamente independiente, lo cual destaca la notoria similitud que existe entre los modelos que se utilizaron para realizar las estimaciones con Kriging.

5.3. Comparación respecto a los datos originales del tipo 2

La relevancia de los distintos escenarios presentados radica en la identificación de las características que afectan la estimación y su adecuado tratamiento. Algunas características definen la distribución de los datos e influyen de manera significativa en la estimación. En la práctica no se conoce la cantidad de datos, el tipo de muestreo o las características que pueden presentar las bases de datos con las cuales se tiene que trabajar. Por lo tanto, es importante destacar las dificultades que se presentan cuando varían las características mencionadas. Las dificultades más relevantes presentadas en los escenarios fueron las siguientes.

1. *Escenarios con 36 observaciones D2MRc036, D2MAc036 y D2MCc036:*

   Para los escenarios con 36 observaciones el ajuste del variograma fue casi completamente visual debido a que no se cuenta con información suficiente para obtener suficientes intervalos del variograma de manera confiable. Debido a que el variograma está calculado pobremente y por lo tanto no permite definir bien sus parámetros, la experiencia cuenta mucho al realizar el ajuste visual. El escenario del muestreo aleatorio tuvo un ajuste completamente visual y no se utilizó un modelo de tendencia debido a que no se encontró ningún modelo que cumpliera con los criterios en la validación cruzada y el análisis de residuales, por lo que se convirtió en el ajuste más complejo. A pesar de la pobre información, los mapas resultan confiables hasta cierto punto. Dentro de los escenarios de 36 observaciones, el muestreo combinado resultó ser el menos adecuado en el mapa de estimación, probablemente debido a que el modelo adaptado no era el más adecuado. El muestreo regular no fue tan efectivo e incluso fue complejo el proceso de adecuarle un modelo, esto debido a que las muestras se encontraban a grandes distancias de separación. Por lo tanto, cuando se tiene una característica con tanta influencia en la estimación, es importante tener mayor número de datos para obtener un modelo adecuado y una estimación exitosa.

2. *Escenarios con 100 obervaciones D2MRc100, D2MAc100 y D2MCc100:*

   Los escenarios con 100 observaciones son los más representativos dentro de la base de datos del tipo 2. Las bases de este tipo permiten realizar los cálculos con suficiente rapidez y confiabilidad. Sin embargo, las características no siempre son fácilmente identificadas, e incluso cuando son identificadas no son fácilmente tratables. Para estos escenarios, se comienzan a observar en los gráficos las similitudes que existen con los datos originales del tipo 2. Desde el análisis exploratorio de datos se observan características en común como asimetría y falta de estacionariedad o tendencia sobre el eje y. El escenario de D2MRc100 es el más parecido a los datos originales del tipo 2. El escenario de D2MCc100 tiene una estimación cercana a los datos originales en las regiones de valores elevados, sin embargo no cumple en las de menores. El escenario de D2MAc100 es el más complejo dentro de los escenarios con 100 observaciones y la estimación no es tan adecuada, probablemente debido a las coordenadas utilizadas ya que no son regulares y tienen regiones donde no se cuenta con información suficiente.

3. *Escenarios con 400 observaciones D2MRc400, D2MAc400 y D2MCc400:*

   En los escenarios con 400 observaciones la característica de tendencia se nota más acentuada y permite identificarla con mayor confiabilidad. Conforme aumentó la cantidad de información, el tipo de muestreo resultó menos importante, debido a que se cuenta con mucha más información sobre el área total. Incluso el mapa que se encuentra mejor estimado es el de D2MAc400. Los muestreos de malla regular y combinado también resultaron en mapas muy parecidos a los de los datos originales del tipo 2. Cuando se tienen tantas muestras también se observa que los variogramas son muy parecidos entre los distintos tipos de muestreo y se asemejan al variograma de los datos originales del tipo 2, por lo que el ajuste visual del modelo se vuelve muy similar y por consiguiente la estimación también lo es. Sin embargo, cuando se cuenta con mayor número de datos el tiempo para realizar la estimación aumenta considerablemente respecto de los demás. Cabe destacar, que la información obtenida en los mapas de estimación resultantes es más acertada pero menos continua.

A continuación se muestra el procedimiento a seguir con tres escenarios representativos. Se utilizan los escenarios de cien observaciones y se realiza el análisis exploratorio de datos, el análisis estructural y la estimación con Kriging para los tres tipos de muestreo.

Los escenarios representativos son:

- D2MRc100
- D2MAc100
- D2MCc100

5.4. Base de datos del tipo 2 con muestreo de malla regular y 100 observaciones (D2MRc100)

Se inicia el escenario de D2MRc100 con el análisis exploratorio. Se muestran la dispersión de los datos (Figura 5.26) y las estadísticas básicas (Cuadro 5.12).

En la dispersión de los datos se muestra que en la región norte es donde la variable tiene los valores más altos. La media es mayor que la mediana por lo que indica una asimetría positiva lo cual se confirma en las estadísticas básicas con el coeficiente de asimetría. El 50 % de la información se encuentra en el intervalo de [0.25, 1.3) y el restante 50 % se encuentra en el intervalo (1.3, 27.5].

Luego se realiza el histograma (Figura 5.27) y q-q plot (Figura 5.28). En el histograma se observa una asimetría notable y 17 datos atípicos. El q-q plot confirma que la distribución no es normal.

Nombre	Estadísticas
Número total	100
Distancia max	50.91168825
Distancia min	4
Media	5.08
Varianza	49.07285551
Desviación estándar	7.005202032
Coeficiente var	1.378862822
Rango min	0.25
1er cuantil	0.25
Mediana	1.282
3er cuantil	6.187
Máximo rango	27.5
Asimetría	1.459465339
Curtosis	3.877884348

Cuadro 5.12: Estadísticas básicas D2MRc100

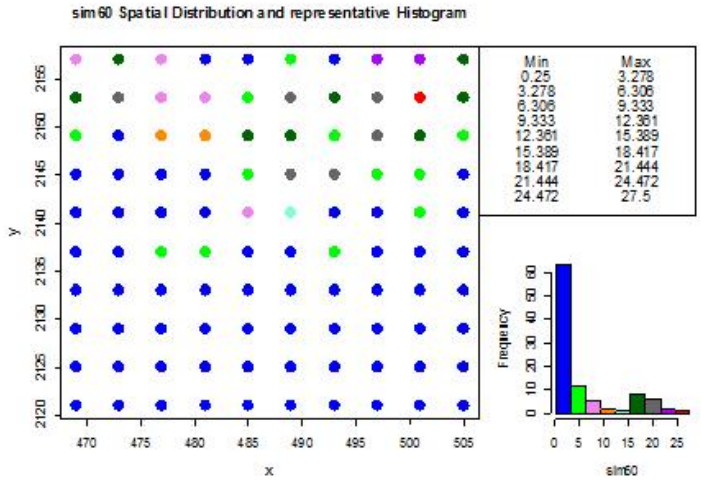

Figura 5.26: Distribución de D2MRc100

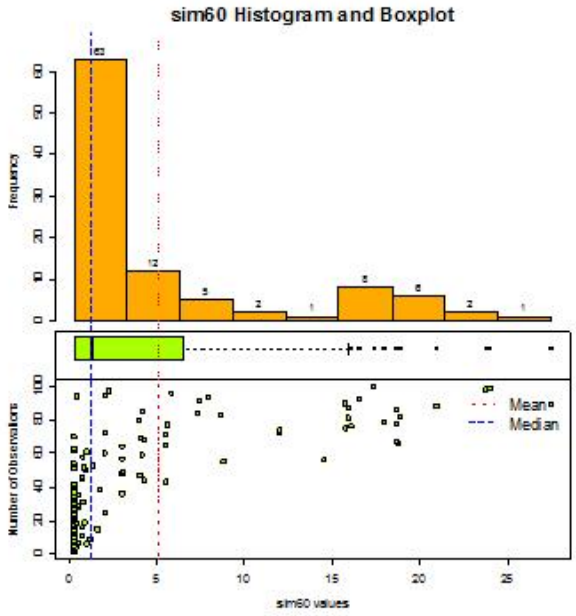

Figura 5.27: Histograma de D2MRc100

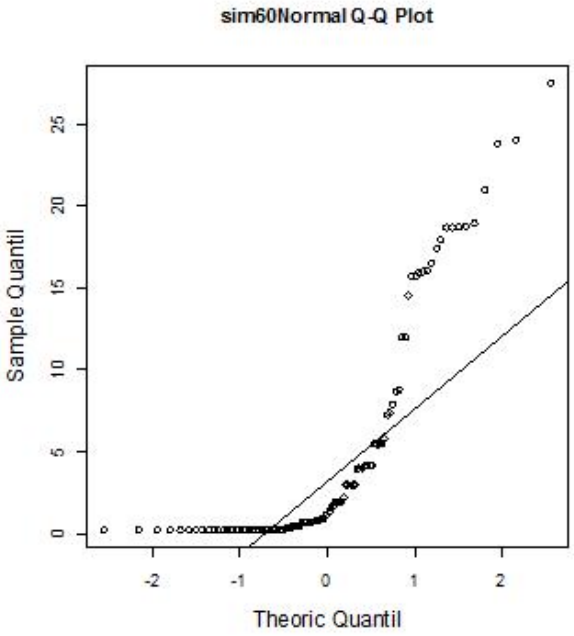

Figura 5.28: Q-Q plot de D2MRc100

Como no cumple los supuestos se aplica una transformación de logaritmo para mejorar la distribución de los datos reduciendo la escala, y se realiza nuevamente el análisis exploratorio.

Dentro de las estadísticas básicas (Cuadro 5.13), la media sigue siendo mayor que la mediana lo que sigue indicando una asimetría positiva, sin embargo se corrobora con el coeficiente de asimetría que los datos se encuentran más centrados y que la asimetría disminuyó considerablemente.

Nombre	**Estadísticas**
Número total	100
Distancia max	50.91168825
Distancia min	4
Media	0.4634
Varianza	2.710505911
Desviación estándar	1.646361416
Coeficiente var	3.552672252
Rango min	-1.386
1er cuantil	-1.386
Mediana	0.2453
3er cuantil	1.818
Máximo rango	3.314
Asimetría	0.248515318
Curtosis	1.578014535

Cuadro 5.13: Estadísticas básicas de logarítmo D2MRc100

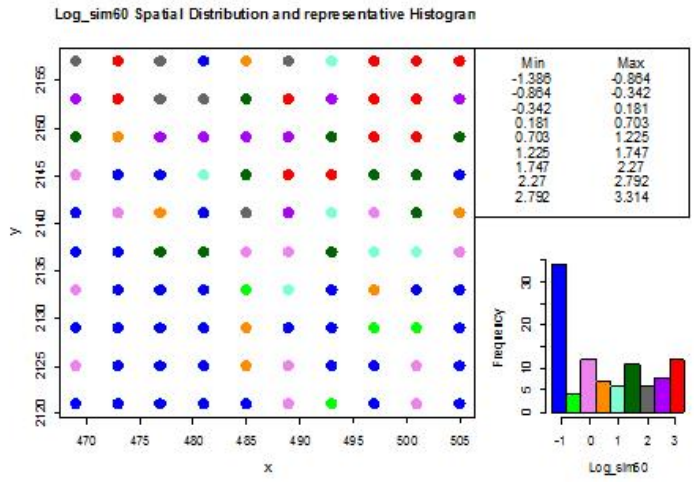

Figura 5.29: Distribución de logD2MRc100

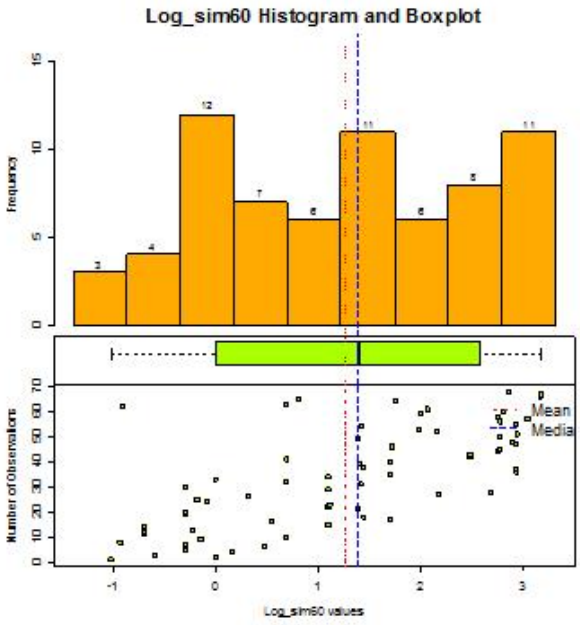

Figura 5.30: Histograma de logD2MRc100

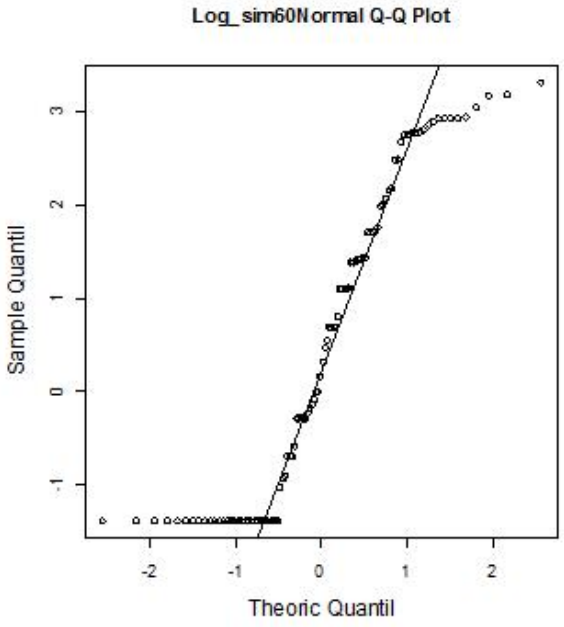

Figura 5.31: Q-Q plot de logD2MRc100

El histograma (Figura 5.30) muestra que la distribución es más centrada, la media y mediana son más cercanas y ya no existen datos atípicos. El q-q plot (Figura 5.31) confirma que no se tiene una distribución normal, sin embargo los datos son suficientemente simétricos para continuar.

Se verifica la estacionariedad (Figura 5.33) y se realiza el gráfico con respecto a las coordenadas (Figura 5.32) para buscar indicios de tendencia.

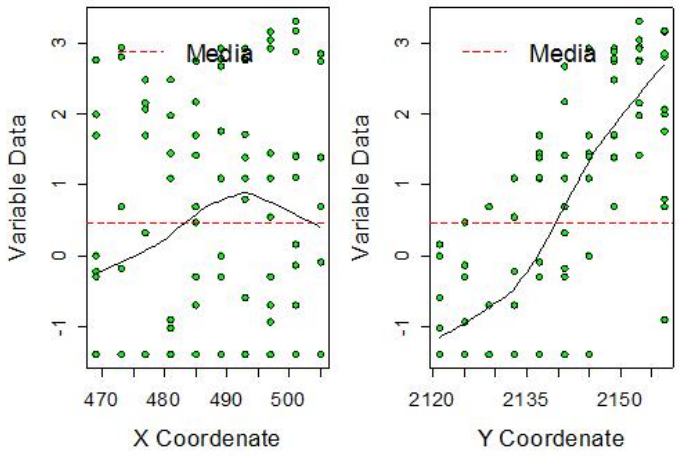

Figura 5.32: Gráfico con respecto a las coordenadas de logD2MRc100

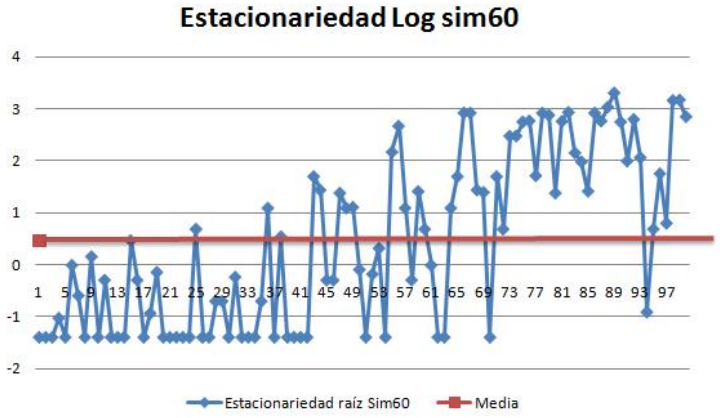

Figura 5.33: Gráfico de estacionariedad de logD2MRc100

No se observa estacionariedad y se tiene una tendencia notable en el gráfico de coordenadas. Se realizan los variogramas para corroborar la sospecha de tendencia. El variograma adireccional se realiza con un lag de 4 debido a que es una malla regular y es la distancia de separación entre observaciones.

Disctancia max	50.91168825
Distancia min	4
Dirección	0°
Tolerancia	90°
Intervalos	12
Distancia Lag	4

Cuadro 5.14: Variograma adireccional de logD2MRc100

Disctancia max	50.91168825
Distancia min	4
Dirección	0°,45°,90°,135°
Tolerancia	22.5°
Intervalos	12
Distancia Lag	4

Cuadro 5.15: Variograma 4 direcciones raíz de Datos 1

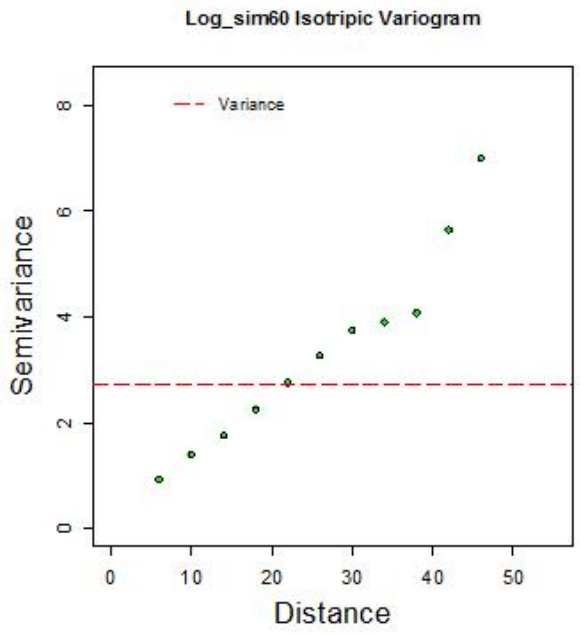

Figura 5.34: Variograma adireccional de logD2MRc100

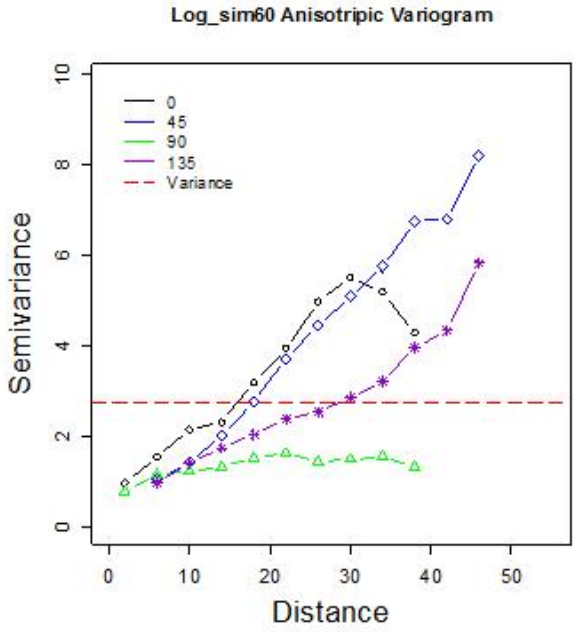

Figura 5.35: Variograma en 4 direcciones de logD2MRc100

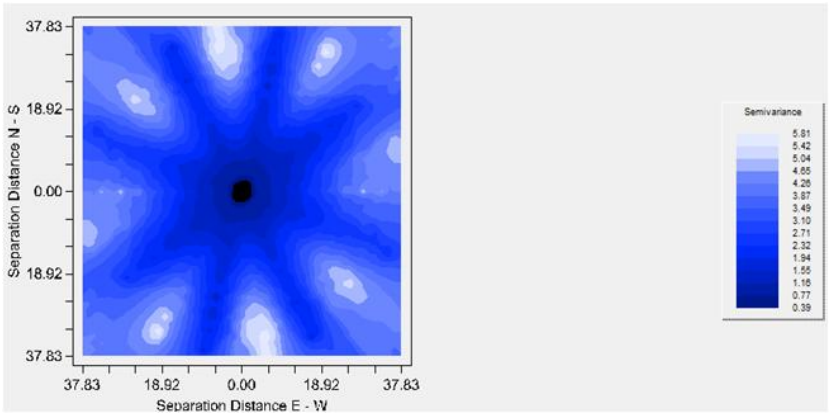

Figura 5.36: Mapa de anisotropía de logD2MRc100

En el variograma adireccional (Cuadro 5.14) Figura 5.34 muestra un comportamiento lineal ascendente, lo cual corrobora la tendencia. El variograma en cuatro direcciones (Cuadro 5.15 Figura5.35) y el mapa de anisotropía (Figura 5.36) muestran una ligera anisotropía, sin embargo puede ser ocasionada por la tendencia que se observa.

Por lo tanto se genera el modelo de tendencia (Cuadro 5.16) a partir de los datos iniciales sin transformación logarítmica y se realiza nuevamente el análisis exploratorio.

La nueva transformación en base a la tendencia de primer grado muestra en las estadísticas básicas (Cuadro 5.17) que también redujo la escala en comparación a los datos iniciales y el histograma (Figura 5.37) muestra una distribución más centrada sin datos atípicos y con un q-q plot (Figura 5.38) más cercano a una distribución normal.

Coeficiente	**Valor**
Intercepto	-902.691983
Coeficiente de x	0.102958383
Coeficiente de y	0.400949822

Cuadro 5.16: Modelo de Tendencia de 1er grado modificada D2MRc100

Nombre	**Estadísticas**
Número total	100
Distancia max	50.91168825
Distancia min	4
Media	1.03x10^{-}16
Varianza	26.22469701
Desviación estándar	5.121005469
Coeficiente var	4.98x10^{1}6
Rango min	-11.27
1er cuantil	-3.476
Mediana	-0.9369
3er cuantil	3.12
Máximo rango	15.36
Asimetría	0.474700904
Curtosis	3.169100435

Cuadro 5.17: Estadísticas básicas sin tendencia D2MRc100

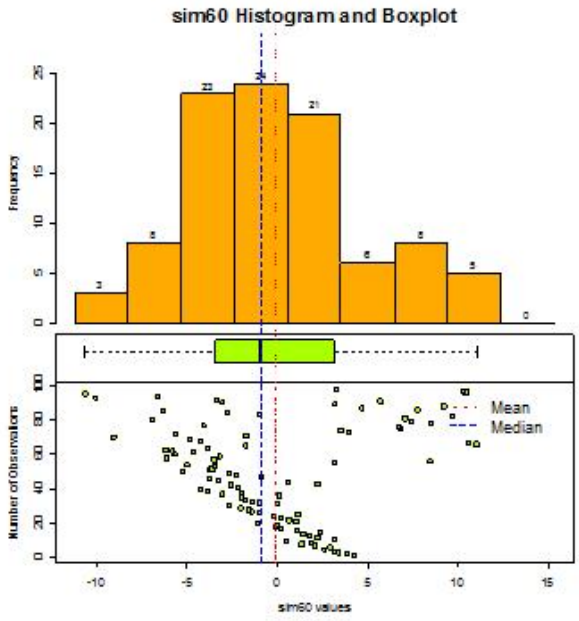

Figura 5.37: Histograma del modelo sin tendencia D2MRc100

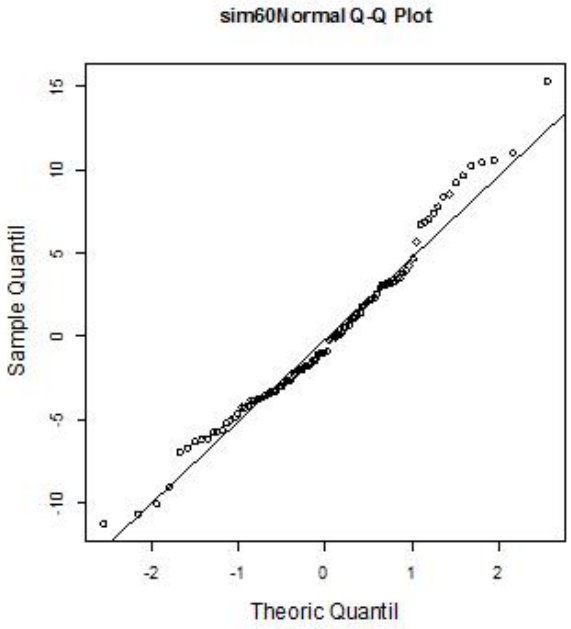

Figura 5.38: Q-Q plot del modelo sin tendencia de D2MRc100

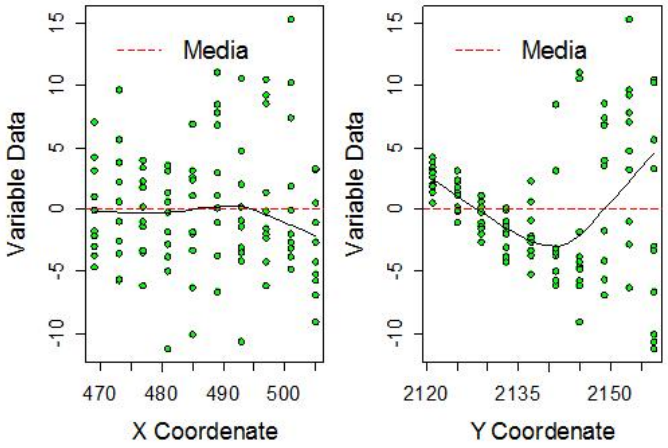

Figura 5.39: Gráfico respecto a las coordenadas modelo s/tendD2MRc100

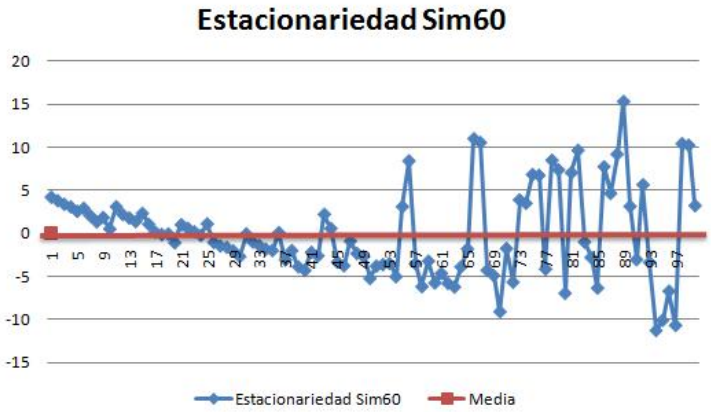

Figura 5.40: Estacionariedad del modelo sin tendencia de D2MRc100

Aunque la estacionariedad (Figura 5.40) no se observa tan adecuada, se confirma que ya no se cuenta con tendencia en el gráfico con respecto a las coordenadas (Figura 5.39). Por lo que se procede a realizar el análisis variográfico.

Disctancia max	50.91168825
Distancia min	4
Dirección	0°
Tolerancia	90°
Intervalos	12
Distancia Lag	4

Cuadro 5.18: Variograma adireccional del modelo sin tendencia de D2MRc100

Disctancia max	50.91168825
Distancia min	4
Dirección	0°,45°,90°,135°
Tolerancia	22.5°
Intervalos	12
Distancia Lag	4

Cuadro 5.19: Variograma 4 direcciones modelo sin tendencia D2MRc100

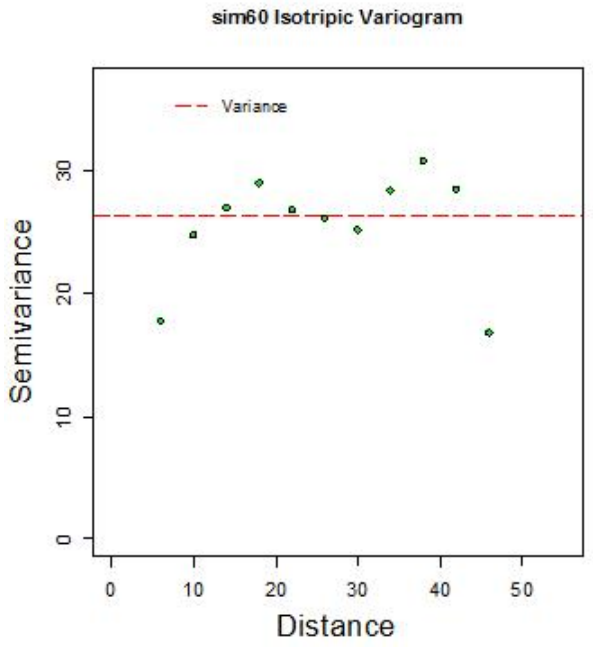

Figura 5.41: Variograma adireccional modelo sin tendencia de D2MRc100

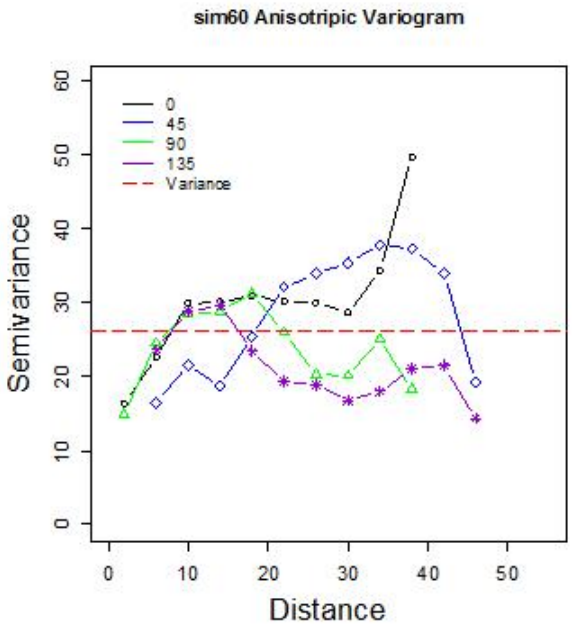

Figura 5.42: Variograma en 4 direcciones del modelo sin tendencia D2MRc100

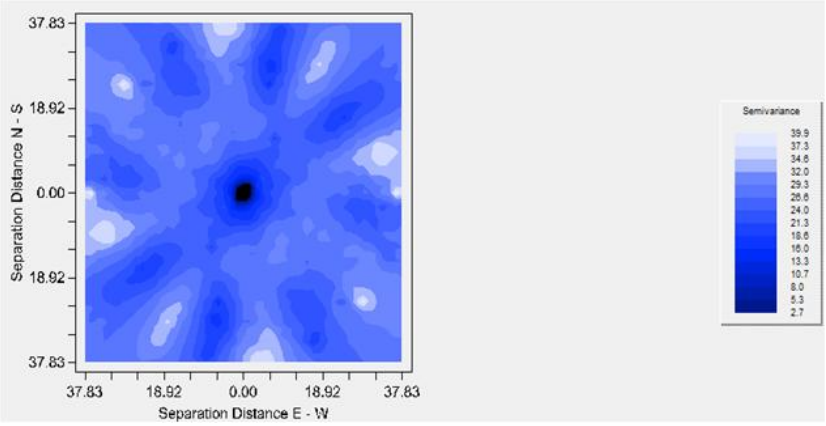

Figura 5.43: Mapa de anisotropía del modelo sin tendencia D2MRc100

El variograma adireccional (Cuadro 5.18 Figura 5.41) no muestra tendencia. El variograma en 4 direcciones (Cuadro 5.19 Figura 5.42) no tiene indicios de anisotropía, por lo que se procede con las propuestas del ajuste del modelo (Cuadro 5.20 Figura 5.44). El modelo que mejor se adapta al variograma es el esférico, por lo que se realiza el ajuste visual para llegar al modelo elegido (Cuadro 5.21 Figura 5.45).

Modelo	Nugget	Sill+Nugget	Rango	SCE
Exponencial	0	27.61	4.83	27.7
Esférico	2.61	27.23	13.74	22.7
Gaussiano	6.64	27.28	6.85	147

Cuadro 5.20: Propuestas para modelos de variograma s/tendencia D2MRc100

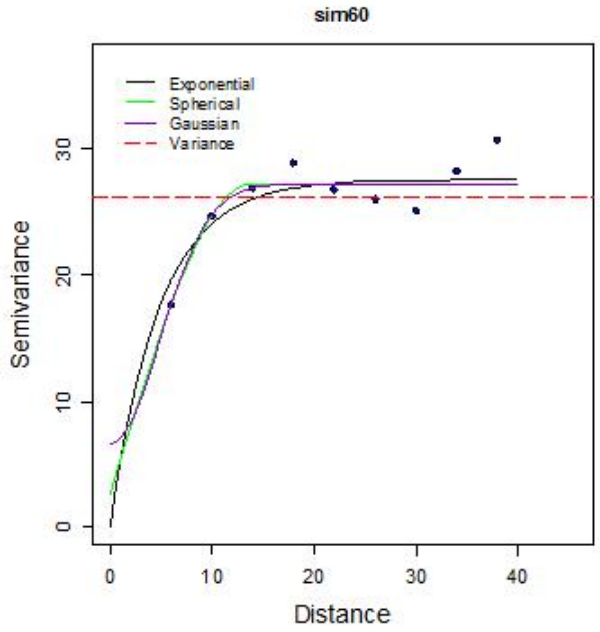

Figura 5.44: Propuestas de modelos de variograma sin tendencia D2MRc100

Modelo	Nugget	Sill+Nugget	Rango	SCE
Esférico	5	27	14	24.5

Cuadro 5.21: Modelo de variograma elegido sin tendencia D2MRc100

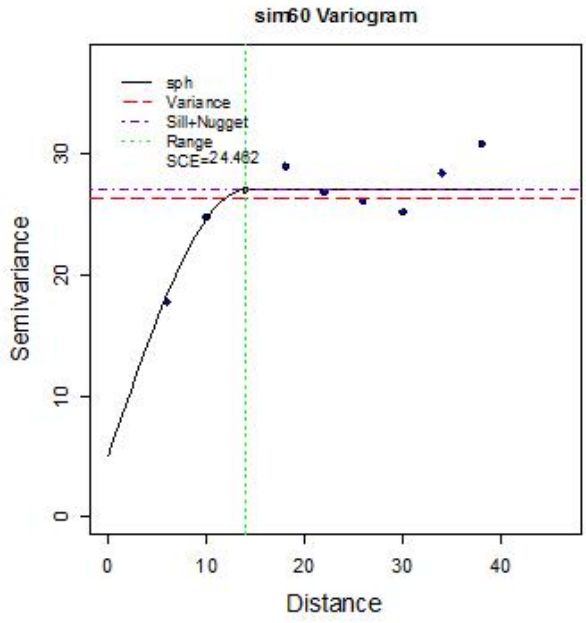

Figura 5.45: Modelo de variograma elegido sin tendencia D2MRc100

Una vez elegido el modelo esférico con su respectivo ajuste visual es necesario verificar si el modelo es adecuado, por lo que se realiza la validación cruzada (Cuadro 5.22). Se verifica que los residuales tienen media muy cercana a cero y los estimados tienen una distribución muy cercana a los datos sin tendencia. Se realiza también el análisis gráfico de residuales.

Nombre	**D2MRc100 s/tend**	**Estimados**	**Error**
Número total	100	100	100
Distancia max	50.91168825	50.91168825	50.91168825
Distancia min	4	4	4
Media	1.03x10^{-}16	0.04162	-0.04162
Varianza	26.22469701	11.241411	15.46704808
Desviación estándar	5.121005469	3.352821349	3.932816812
Coeficiente var	4.98x10^{1}6	80.5657073	94.5025503
Rango min	-11.27	-7.856	-14.04
1er cuantil	-3.476	-2.26	-1.17
Mediana	-0.9369	-0.06498	-0.1154
3er cuantil	3.12	1.97	1.315
Máximo rango	15.36	9.025	9.688
Asimetría	0.474700904	0.324479392	-0.48040858
Curtosis	3.169100435	3.243461909	5.071655581

Cuadro 5.22: Validación cruzada del modelo sin tendencia D2MRc100

Se realiza el gráfico de valores reales contra estimados (Figura 5.46), el histograma (Figura 5.47) y el q-q plot (Figura 5.48).

Aunque los residuales no muestran una distribución normal están centrados y se aproximan a una línea de 45°. Por lo tanto, es suficiente para realizar las estimaciones con Kriging. Se obtiene un mapa del área completa (Figura 5.49) a una distancia de 1km por 1km para cada estimación.

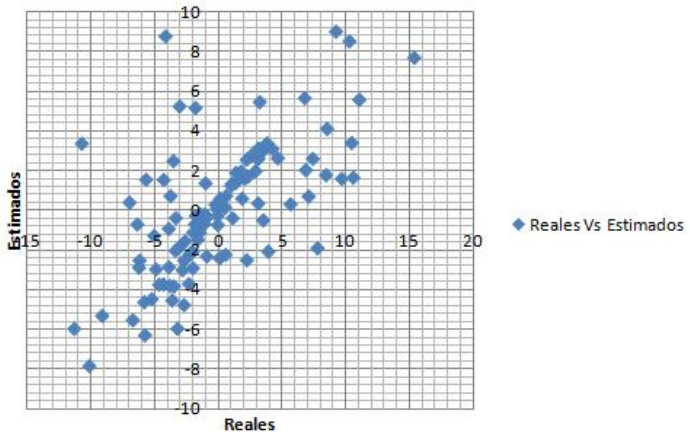

Figura 5.46: Valores reales contra estimados modelo sin tendencia D2MRc100

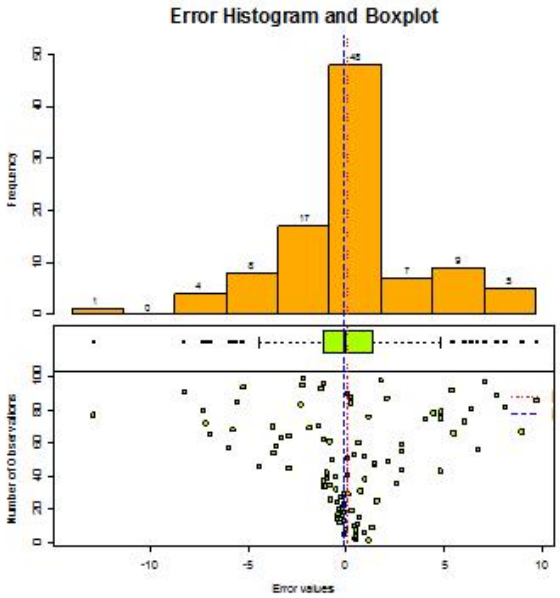

Figura 5.47: Histograma de residuales modelo sin tendencia D2MRc100

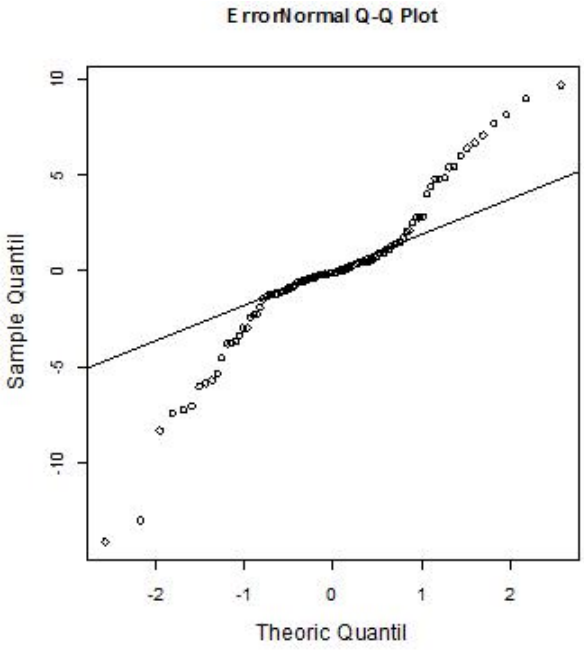

Figura 5.48: Q-Q plot de residuales del modelo sin tendencia D2MRc100

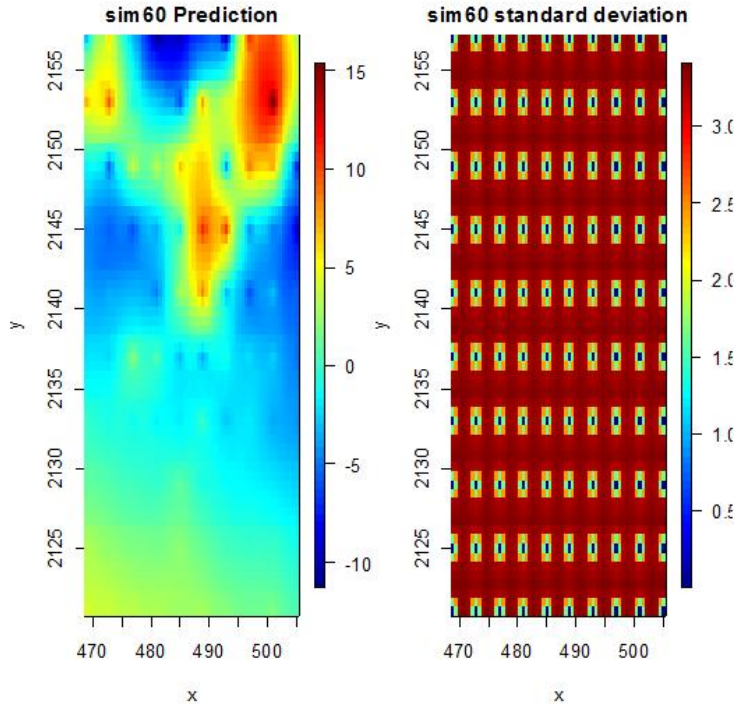

Figura 5.49: Mapa de estimaciones del modelo sin tendencia de D2MRc100

5.5. Base de datos del tipo 2 con muestreo aleatorio y 100 observaciones (D2MAc100)

Se inicia el escenario de D2MAc100 con el análisis exploratorio. Se muestran la dispersión de los datos (Figura 5.50) y las estadísticas básicas (Cuadro 5.23). Dentro de las estadísticas básicas la media es mayor que la mediana lo cual indica una asimetría positiva y se confirma con el coeficiente de asimetría. El 50 % de la información se encuentra en el intervalo de [0.25, 1.06] y el restante 50 % se encuentra en el intervalo (1.6, 27.5].

Nombre	Estadísticas
Número total	100
Distancia max	48.7954916
Distancia min	1
Media	4.589
Varianza	42.66389891
Desviación estándar	6.531760781
Coeficiente var	1.423410789
Rango min	0.25
1er cuantil	0.25
Mediana	1.056
3er cuantil	4.801
Máximo rango	27.5
Asimetría	1.775639315
Curtosis	5.191128447

Cuadro 5.23: Estadísticas básicas D2MAc100

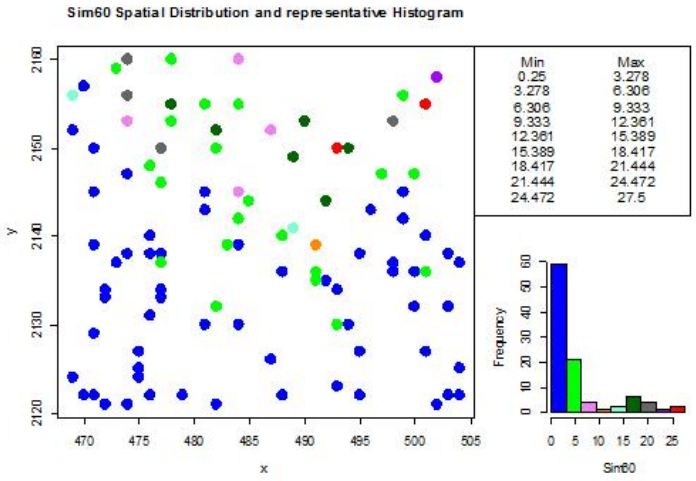

Figura 5.50: Distribución de D2MAc100

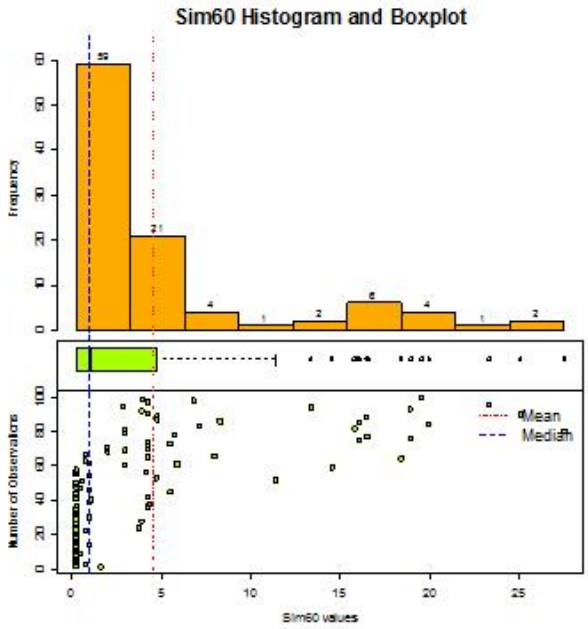

Figura 5.51: Histograma de D2MAc100

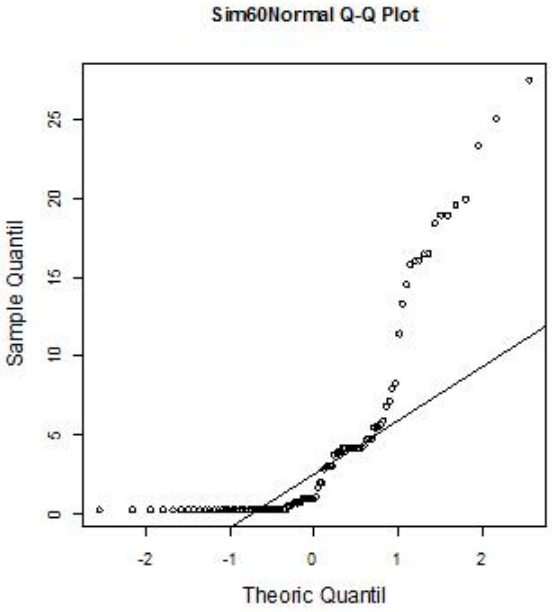

Figura 5.52: Q-Q plot de D2MAc100

El histograma (Figura 5.51) y q-q plot (Figura 5.52) muestran que no se tiene una distribución normal y de hecho es notablemente asimétrica, además de que se tienen 15 datos atípicos.

Por lo cual se realiza una transformación logarítmica para reducir la escala y centrar la distribución de los datos. Se realiza nuevamente el análisis exploratorio de datos.

Nombre	**Estadísticas**
Número total	100
Distancia max	48.7954916
Distancia min	1
Media	0.3919
Varianza	2.638098206
Desviación estándar	1.624222339
Coeficiente var	4.144648493
Rango min	-1.386
1er cuantil	-1.386
Mediana	0.05348
3er cuantil	1.569
Máximo rango	3.314
Asimetría	0.215124364
Curtosis	1.575017334

Cuadro 5.24: Estadísticas básicas logaritmo de D2MAc100

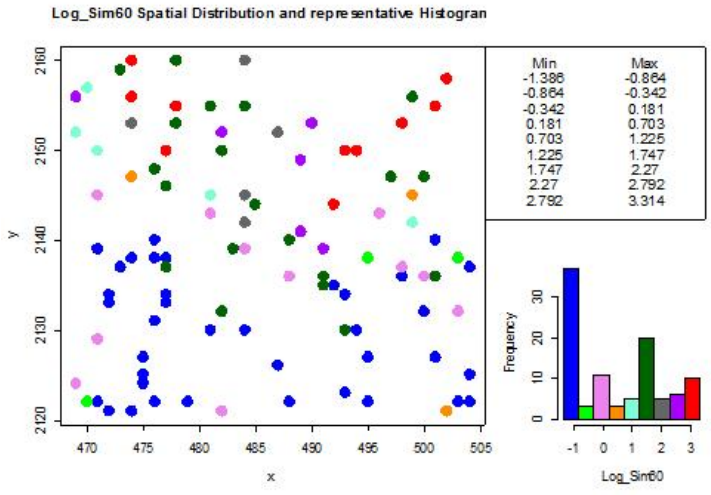

Figura 5.53: Distribución de logaritmo de D2MAc100

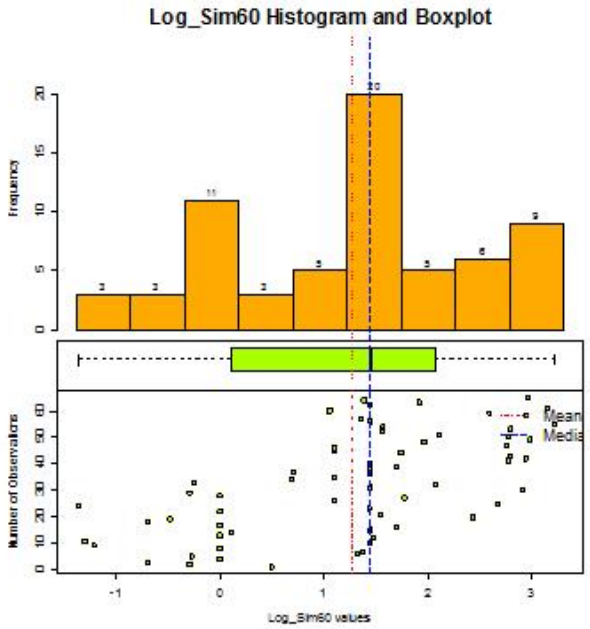

Figura 5.54: Histograma de logarítmo de D2MAc100

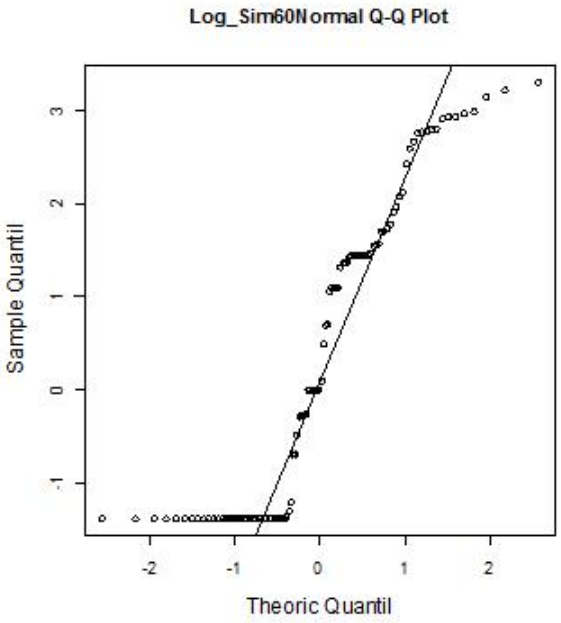

Figura 5.55: Q-Q plot de logarítmo de D2MAc100

Dentro de las estadísticas básicas (Cuadro 5.24) se observa que la media sigue siendo mayor que la mediana y hay una disminución del coeficiente de asimetría. El 75 % de los datos se encuentran en el intervalo de [-1.38, 1.57] y el restante 25 % se encuentra en el intervalo de (1.57, 3.31] por lo tanto se reduce la escala considerablemente.

El histograma (Figura 5.54) muestra los datos de manera más simétrica aunque el q-q plot (Figura 5.55) no confirma normalidad. Además, ya no se tienen datos atípicos. Por lo tanto, se continúa con la metodología y se generan el gráfico respecto a las coordenadas (Figura 5.56), donde se observa una posible tendencia con respecto al eje y. Se realiza el gráfico de estacionariedad (Figura 5.57) donde la muestra no se observa estacionaria y se continúa observando una tendencia.

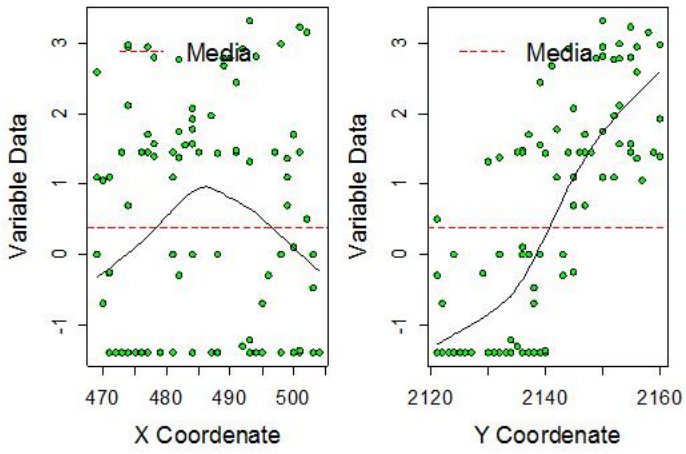

Figura 5.56: Gráfico respecto a las coordenadas de logaritmo de D2MAc100

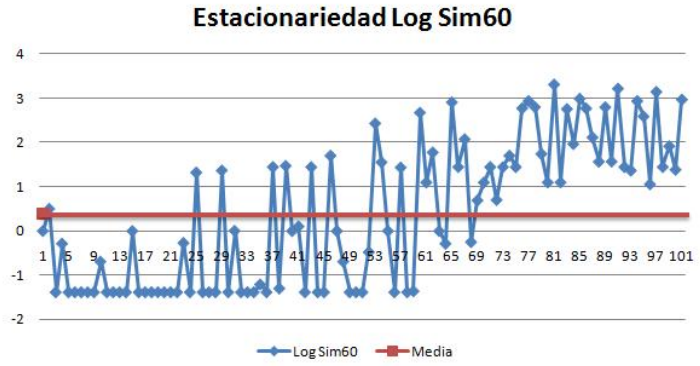

Figura 5.57: Estacionariedad de logaritmo de D2MAc100

Posteriormente se realiza el variograma adireccional (Cuadro 5.25 Figura 5.58) y el variograma en 4 direcciones (Cuadro 5.26 Figura 5.59) para verificar tendencia y anisotropía, así como el mapa de anisotropía (Figura 5.60). El variograma adireccional confirma el comportamiento de h^2 y es visible una tendencia lineal.

Disctancia max	48.7954916
Distancia min	1
Dirección	0°
Tolerancia	90°
Intervalos	12
Distancia Lag	2

Cuadro 5.25: Variograma adireccional de logaritmo de D2MAc100

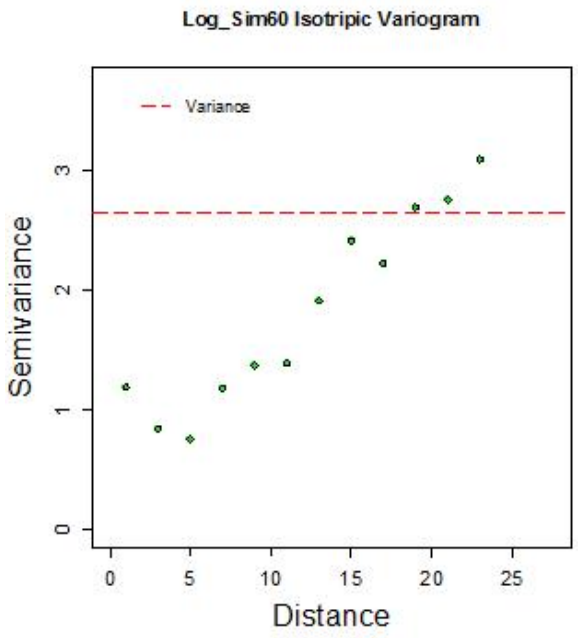

Figura 5.58: Variograma adireccional de logaritmo de D2MAc100

Disctancia max	48.7954916
Distancia min	1
Dirección	0°,45°,90°,135°
Tolerancia	22.5°
Intervalos	12
Distancia Lag	2

Cuadro 5.26: Variograma 4 direcciones de logaritmo de D2MAc100

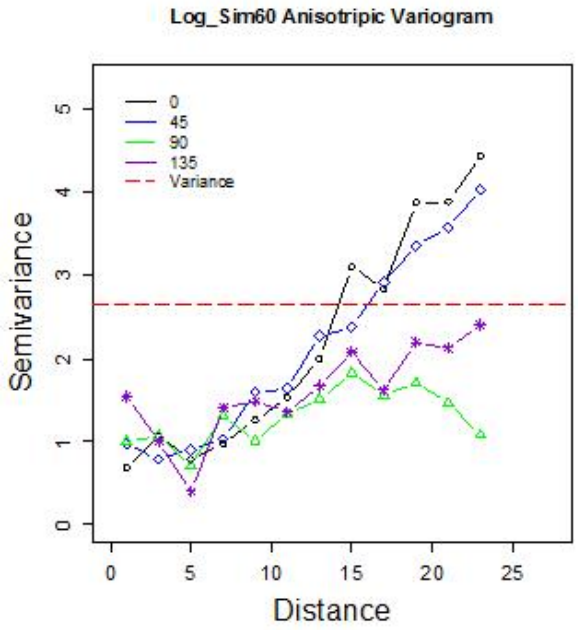

Figura 5.59: Variograma en 4 direcciones de logaritmo de D2MAc100

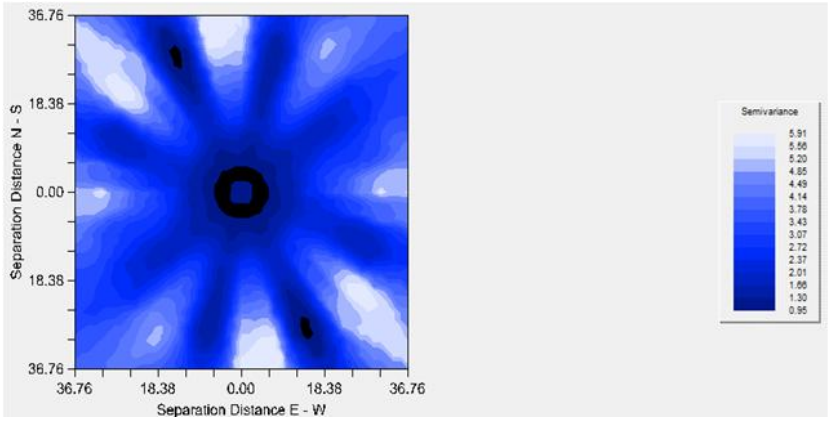

Figura 5.60: Mapa de anisotropía de logaritmo de D2MAc100

Debido a la tendencia, se procede a realizar una transformación para generar un modelo sin tendencia de 1er grado (Cuadro 5.27) y se hace sobre los datos sin transformación de logaritmo. Se realiza nuevamente el análisis exploratorio de datos.

Dentro de las estadísticas básicas (Cuadro 5.27) se observa que la asimetría se redujo aunque no considerablemente. Sin embargo, el histograma (Figura 5.61) y q-q plot (Figura 5.62) muestran que el comportamiento es más cercano a la distribución normal.

Coeficiente	**Valor**
Intercepto	-817.765018
Coeficiente de x	0.087824995
Coeficiente de y	0.364497648

Cuadro 5.27: Modelo de Tendencia de 1er grado modificada D2MAc100

Nombre	**Estadísticas**
Número total	100
Distancia max	48.7954916
Distancia min	1
Media	-4.41x10^{-}17
Varianza	24.93538769
Desviación estándar	4.993534589
Coeficiente var	1.13x10^{1}7
Rango min	-7.995
1er cuantil	-3.454
Mediana	-0.8921
3er cuantil	2.303
Máximo rango	18.3
Asimetría	1.083294986
Curtosis	4.159882579

Cuadro 5.28: Estadísticas básicas sin tendencia de D2MAc100

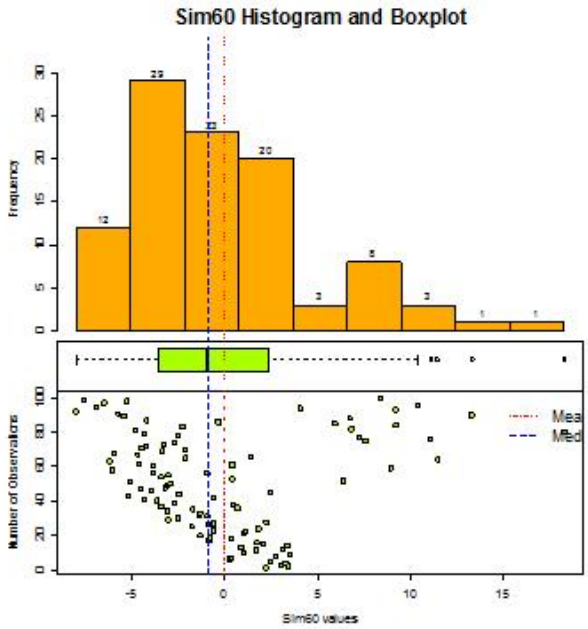

Figura 5.61: Histograma sin tendencia de D2MAc100

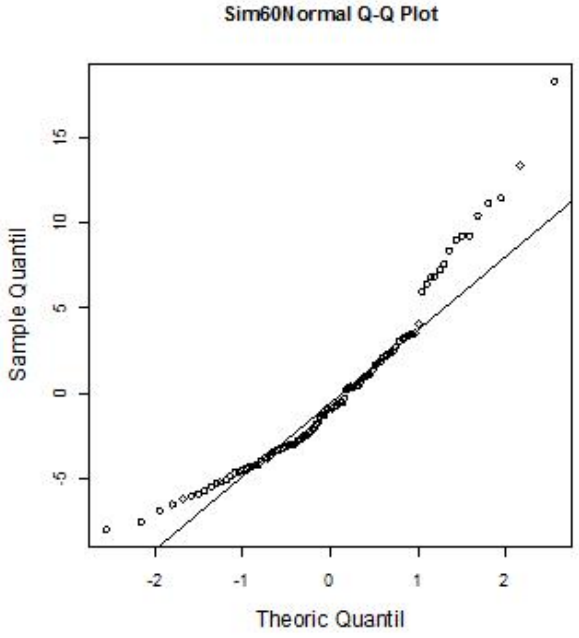

Figura 5.62: Q-Q plot sin tendencia de D2MAc100

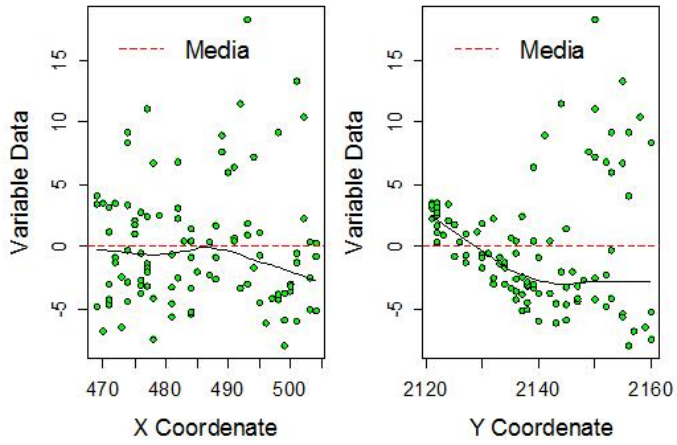

Figura 5.63: Gráfico respecto a las coordenadas sin tendencia de D2MAc100

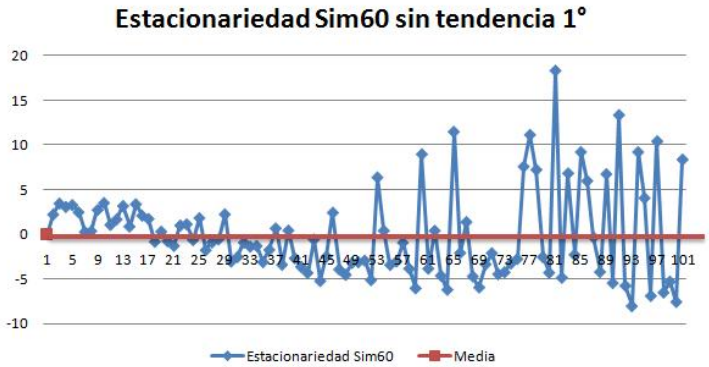

Figura 5.64: Gráfico de estacionariedad sin tendencia de D2MAc100

La estacionariedad (Figura 5.64) se observa con mejor comportamiento. En el gráfico con respecto a las coordenadas (Figura 5.63) se observa la corrección de la tendencia. Por lo tanto, se procede a realizar el análisis variográfico.

El variograma adireccional (Cuadro 5.29 Figura 5.65) ya no muestra tendencia y debido a la cantidad de información se utiliza un lag de 2 y 12 intervalos. En el variograma en 4 direcciones (Cuadro 5.30 Figura 5.66) no se observan diferencias considerables en los alcances. Aunque el mapa de anisotropía (Figura5.67) se observa distinto, no tiene elipses visibles que se consideren como indicios de anisotropía, por lo que no se consideran y se procede a realizar el ajuste del modelo de variograma.

Disctancia max	48.7954916
Distancia min	1
Dirección	0°
Tolerancia	90°
Intervalos	12
Distancia Lag	2

Cuadro 5.29: Variograma adireccional sin tendencia de D2MAc100

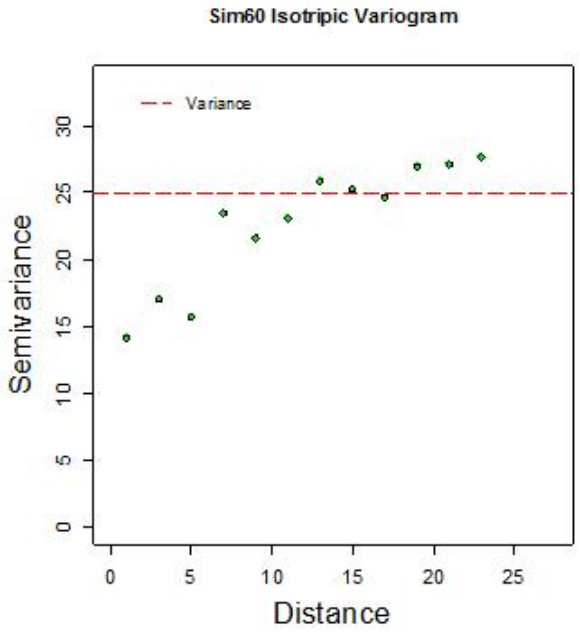

Figura 5.65: Variograma adireccional sin tendencia de D2MAc100

Disctancia max	48.7954916
Distancia min	1
Dirección	0°,45°,90°,135°
Tolerancia	22.5°
Intervalos	12
Distancia Lag	2

Cuadro 5.30: Variograma 4 direcciones sin tendencia de D2MAc100

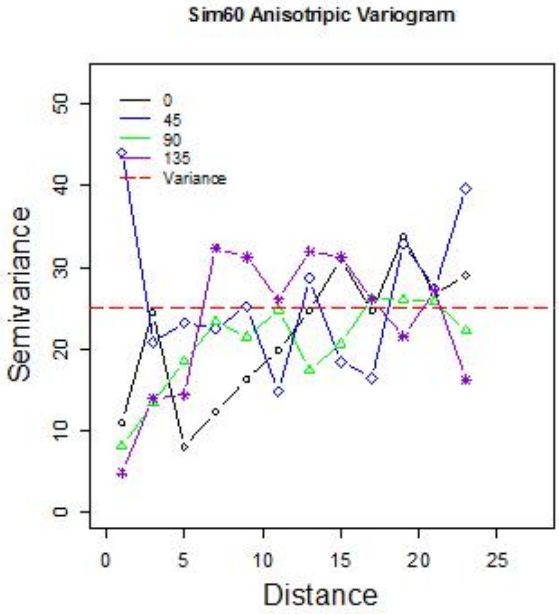

Figura 5.66: Variograma en 4 direcciones sin tendencia de D2MAc100

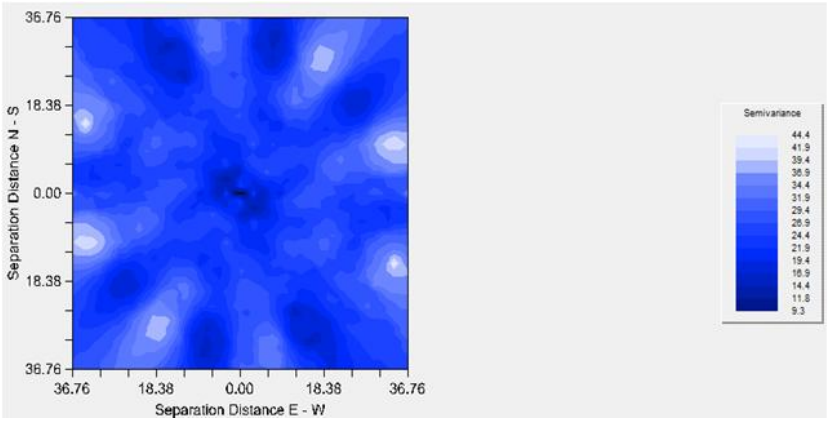

Figura 5.67: Mapa de anisotropia sin tendencia de D2MAc100

Para realizar el ajuste del modelo de variograma, se generan las propuestas de modelos de variograma (Cuadro 5.31), donde se tiene que la menor suma de cuadrados del error está en el modelo exponencial, sin embargo en la Figura 5.68 de los modelos propuestos, se observa que también se le puede ajustar un modelo esférico en el cual el nugget sería menor. Es importante disminuir el nugget ya que tiene una notoria afectación sobre las estimaciones. Por lo tanto, se elige el modelo esférico y se realiza un significativo ajuste visual para llegar al modelo elegido (Cuadro 5.32 Figura 5.69).

Modelo	**Nugget**	**Sill+Nugget**	**Rango**	**SCE**
Exponencial	10.92	28.79	8.97	24.1
Esférico	13.92	27.07	21	24.7
Gaussiano	15.3	26.87	9.59	57.1

Cuadro 5.31: Propuestas para modelos de variograma sin tendencia D2MAc100

Modelo	**Nugget**	**Sill+Nugget**	**Rango**	**SCE**
Esférico	8	26.3	15	55.8

Cuadro 5.32: Modelo de variograma elegido sin tendencia D2MAc100

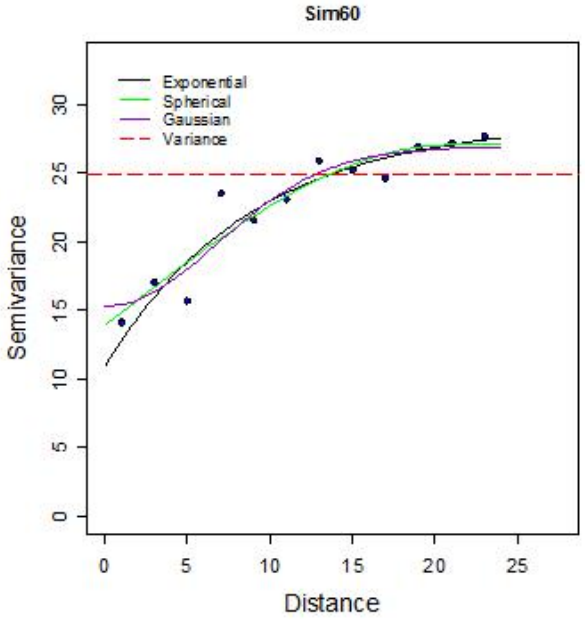

Figura 5.68: Propuestas de modelos de variograma s/tendencia de D2MAc100

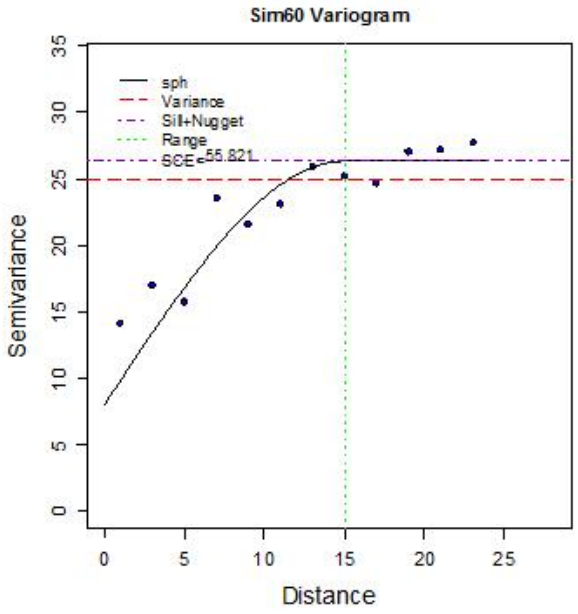

Figura 5.69: Modelo de variograma elegido sin tendencia de D2MAc100

Como el ajuste visual fue significativo, verificar que éste sea adecuado también resulta más importante. Se utiliza la validación cruzada (Cuadro 5.33) y el análisis de residuales para verificarlo. La validación cruzada muestra que la media de los errores es cercana a cero. Además, la distribución de los estimados es también cercana a la distribución de los datos sin tendencia.

Nombre	**s/tend D2MAc100**	**Estimados**	**Error**
Número total	100	100	100
Distancia max	48.7954916	48.7954916	48.7954916
Distancia min	1	1	1
Media	4.41x10^{-}17	-0.00404	0.00404
Varianza	24.93538769	8.398011385	22.51487296
Desviación estándar	4.993534589	2.89793226	4.744983979
Coeficiente var	1.13x10^{1}7	717.3672749	1174.594822
Rango min	-7.995	-5.498	-18.46
1er cuantil	-3.454	-2.13	-1.956
Mediana	-0.8921	-0.2939	-0.2415
3er cuantil	2.303	2.104	1.215
Máximo rango	18.3	10.47	12.88
Asimetría	1.083294986	0.808962674	0.138074643
Curtosis	4.159882579	4.396537846	5.577061036

Cuadro 5.33: Validación cruzada sin tendencia D2MAc100

También se genera el gráfico de valores reales contra estimados (Figura 5.70), el cual muestra que la información se aproxima a la línea de 45°. El histograma (Figura 5.71) y q-q plot (Figura 5.72) muestran una distribución simétrica y similar a la normal. Por lo tanto, el modelo es adecuado y aceptado. Con lo cual se procede a realizar la estimación con Kriging y se muestra el mapa de valores estimados (Figura 5.73) a una distancia de 1km por 1km.

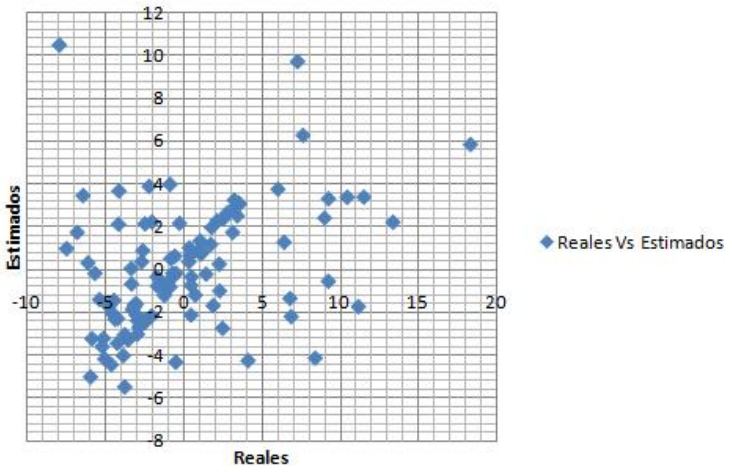

Figura 5.70: Valores reales contra estimados sin tendencia D2MAc100

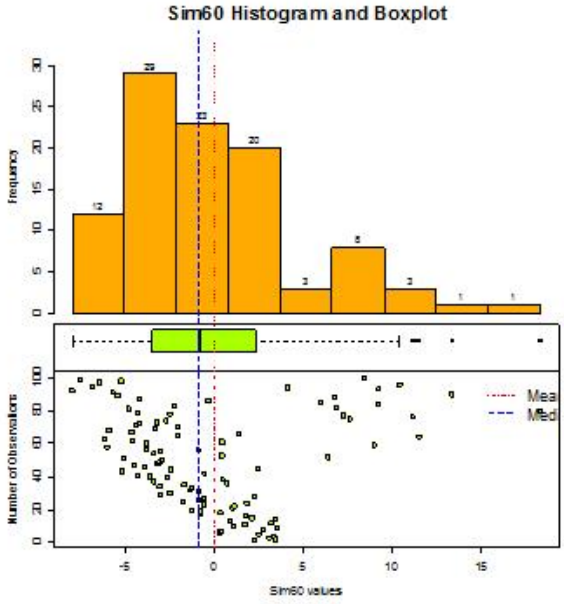

Figura 5.71: Histograma de residuales sin tendencia de D2MAc100

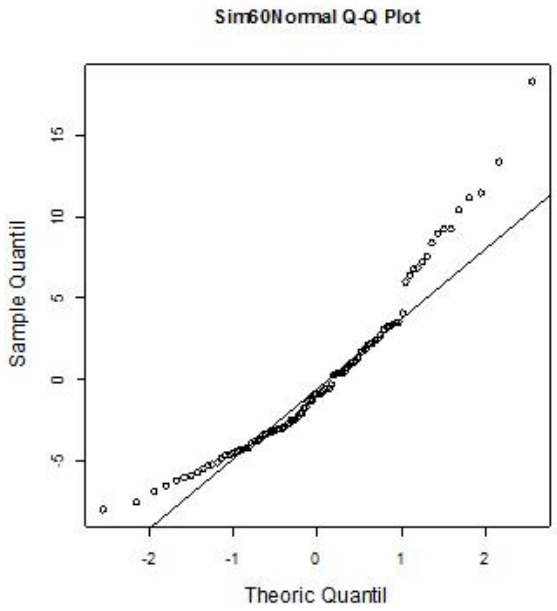

Figura 5.72: Q-Q plot de residuales sin tendencia de D2MAc100

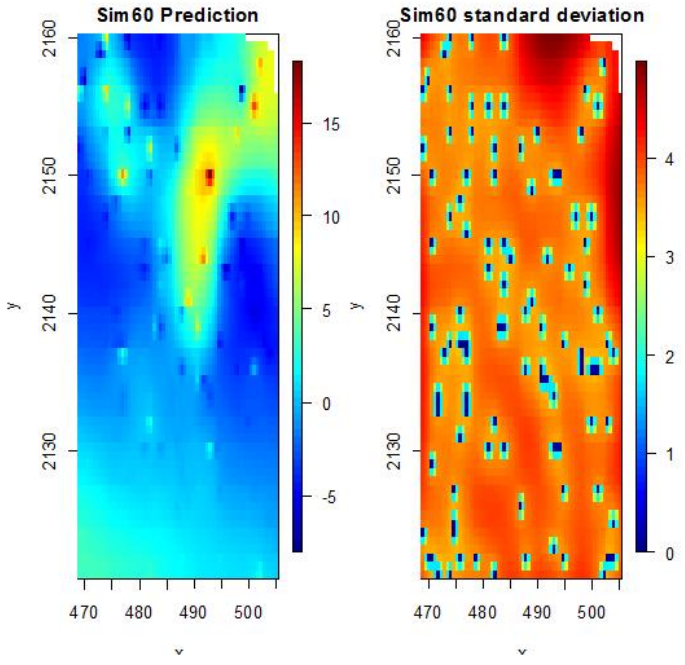

Figura 5.73: Mapa de estimaciones sin tendencia D2MAc100

5.6. Base de datos del tipo 2 con muestreo combinado y 100 observaciones (D2MCc100)

Se inicia el escenario de D2MCc100 con el análisis exploratorio. Se muestra la dispersión de los datos (Figura 5.74) y sus estadísticas básicas (Cuadro 5.34). Dentro de las estadísticas básicas se muestra que la media es mayor que la mediana lo cual sugiere una asimetría positiva la cual es confirmada en el coeficiente de asimetría. El 50 % de la información se encuentra dentro del intervalo [0.25, 1.15], mientras que el restante 50 % se encuentra en el intervalo de (1.15, 24.89].

Nombre	**Estadísticas**
Número total	100
Distancia max	52.400229003
Distancia min	1
Media	5.189
Varianza	44.96320696
Desviación estándar	6.705460981
Coeficiente var	1.292280553
Rango min	0.25
1er cuantil	0.25
Mediana	1.15
3er cuantil	7.819
Máximo rango	24.89
Asimetría	1.268263985
Curtosis	3.338629504

Cuadro 5.34: Estadísticas básicas DCMAc100

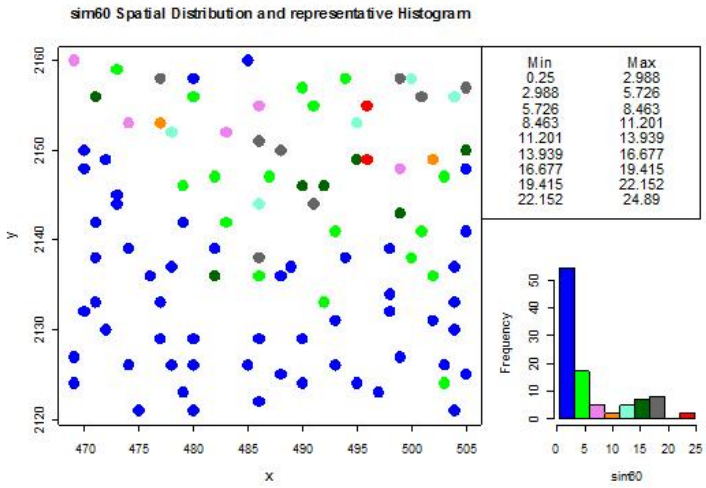

Figura 5.74: Distribución de D2MCc100

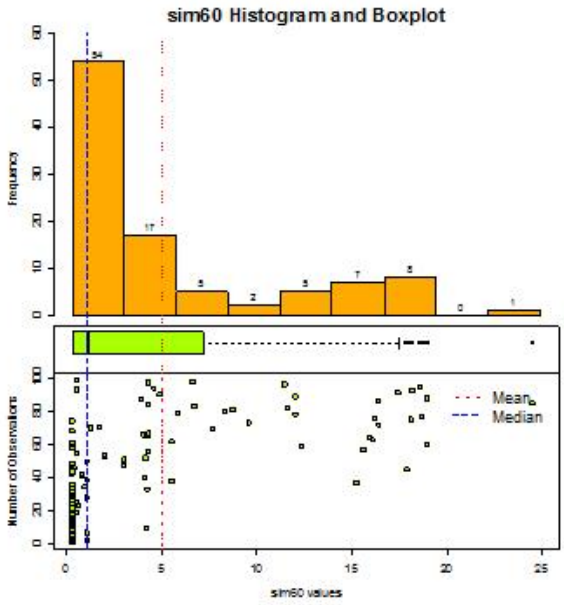

Figura 5.75: Histograma de D2MCc100

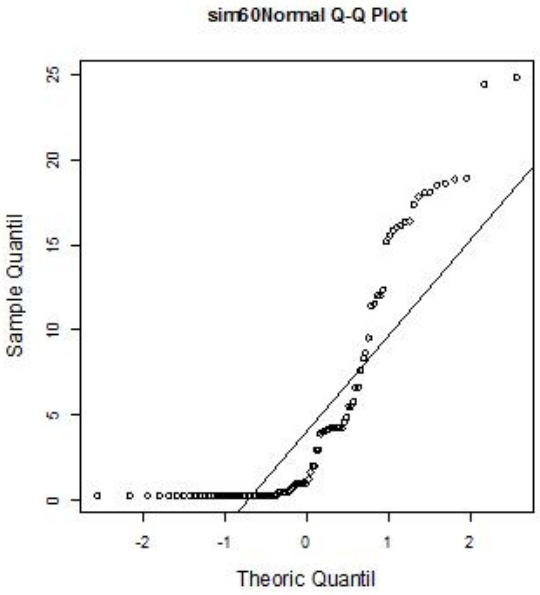

Figura 5.76: Q-Q plot de D2MCc100

El histograma (Figura 5.75) confirma la asimetría mencionada y muestra que se cuenta con varios datos atípicos de los cuales 2 son influyentes. El q-q plot (Figura 5.76) muestra que la distribución no es normal. Por lo tanto se procede a realizar una transformación de logaritmo con la cual se reduce la escala, se reduce la asimetría y se eliminan los datos atípicos. Una vez realizada la transformación se muestra la distribución de los datos (Figura 5.77) y se repite el análisis exploratorio.

Las estadísticas básicas (Cuadro 5.35) muestran la disminución en la escala, ahora el 75 % de la información se encuentra en el intervalo [-1.38, 2.05] y el restante 25 % se encuentra en el intervalo (2.05, 3.22]. La media sigue siendo mayor que la mediana pero la asimetría se ve considerablemente menor.

Nombre	**Estadísticas**
Número total	100
Distancia max	52.40229003
Distancia min	1
Media	0.478
Varianza	2.891212069
Desviación estándar	1.700356454
Coeficiente var	3.557141394
Rango min	-1.386
1er cuantil	-1.386
Mediana	0.1351
3er cuantil	2.056
Máximo rango	3.214
Asimetría	0.165967866
Curtosis	1.416521349

Cuadro 5.35: Estadísticas básicas logD2MCc100

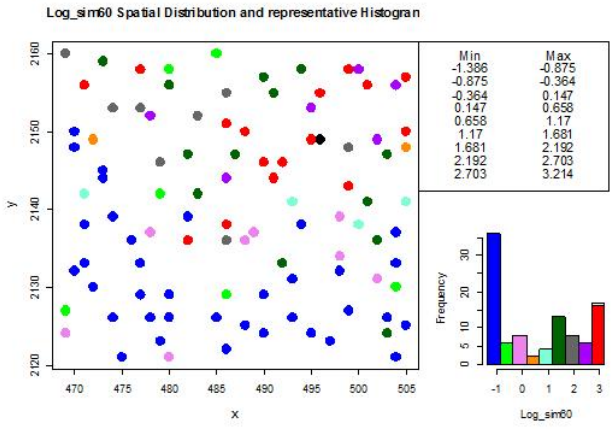

Figura 5.77: Distribución de logD2MCc100

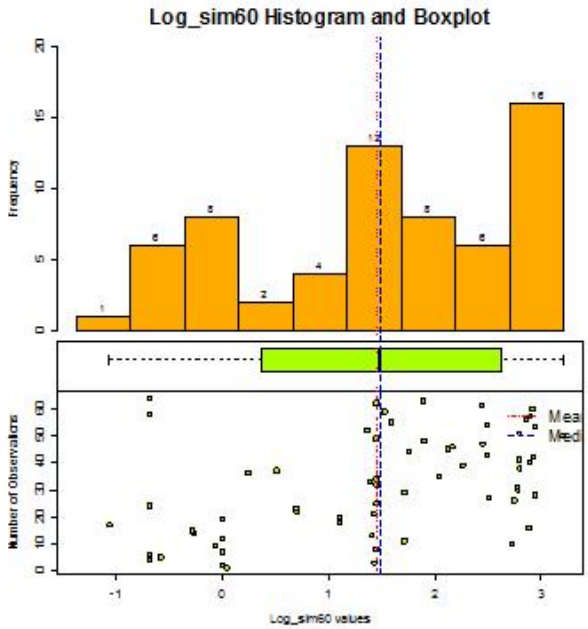

Figura 5.78: Histograma de logD2MCc100

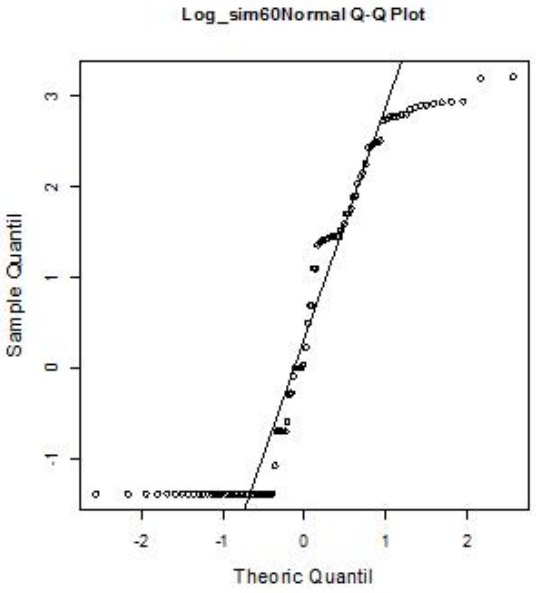

Figura 5.79: Q-Q plot de logD2MCc100

El histograma (Figura 5.78) muestra una distribución más simétrica, pero el q-q plot (Figura 5.79) confirma que no tiene distribución normal. Sin embargo, se procede a realizar el gráfico con respecto a las coordenadas (Figura5.80) y el gráfico de estacionariedad (Figura 5.81) para verificar tendencia, en los cuales sí se observa una posible tendencia con respecto al eje y. Por lo tanto, se verifica con el análisis variográfico.

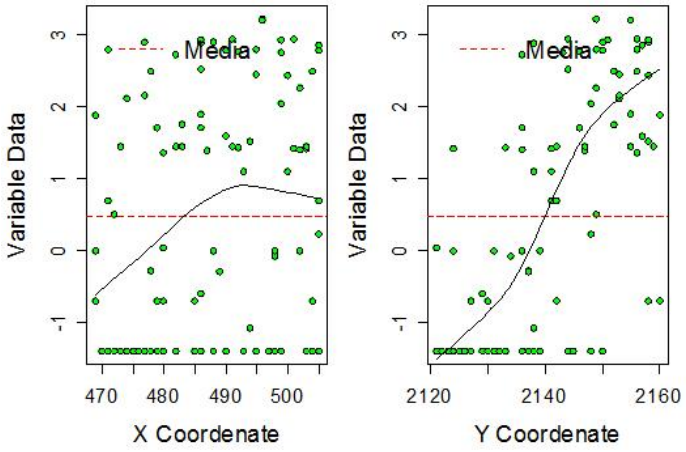

Figura 5.80: Gráfico con respecto a las coordenadas logD2MCc100

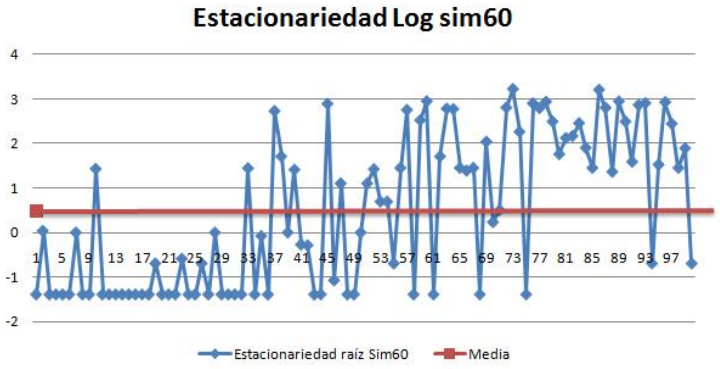

Figura 5.81: Gráfico de estacionariedad de logD2MCc100

Al realizar el variograma adireccional (Cuadro 5.36 Figura 5.82) se confirma la tendencia lineal. De igual manera, se realiza el variograma en cuatro direcciones (Cuadro 5.37 Figura 5.83), así como el mapa de anisotropía (Figura 5.84) para verificar si se presentan diferentes alcances o elipses notorios los cuales son indicios de anisotropía.

Disctancia max	50.91168825
Distancia min	4
Dirección	0°
Tolerancia	90°
Intervalos	13
Distancia Lag	2

Cuadro 5.36: Variograma adireccional de logD2MCc100

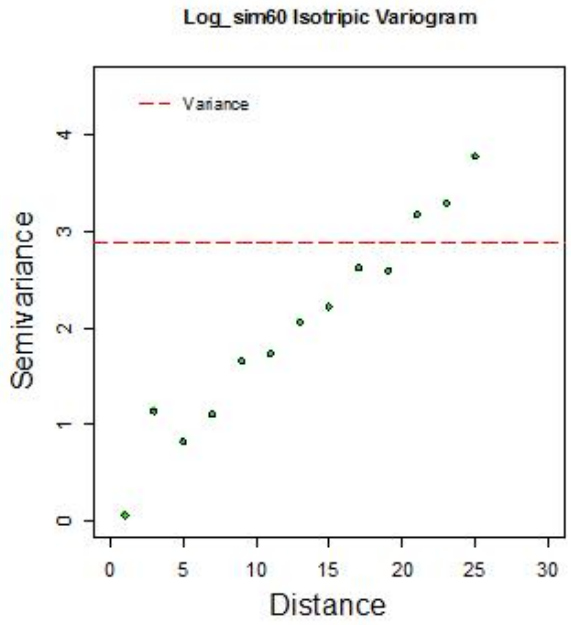

Figura 5.82: Variograma adireccional de logD2MCc100

Disctancia max	50.91168825
Distancia min	4
Dirección	0°,45°,90°,135°
Tolerancia	22.5°
Intervalos	13
Distancia Lag	2

Cuadro 5.37: Variograma 4 direcciones de logD2MCc100

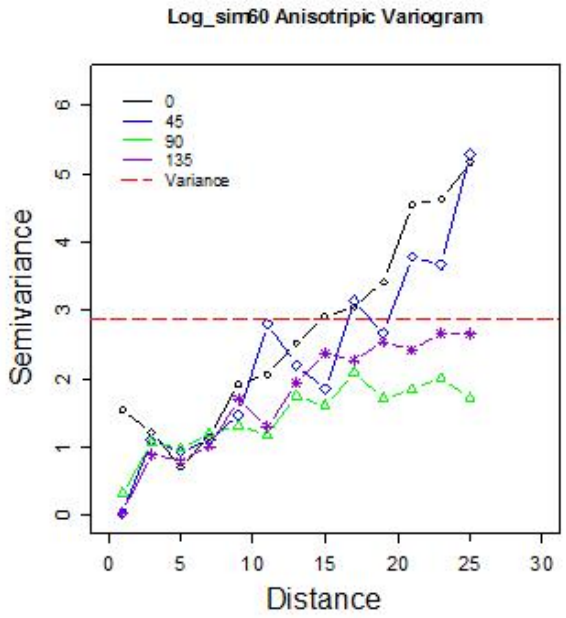

Figura 5.83: Variograma en 4 direcciones de logD2MCc100

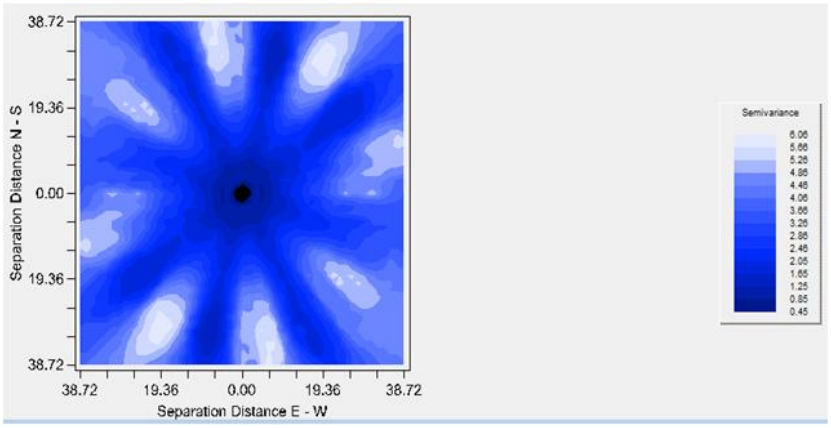

Figura 5.84: Mapa de anisotropía de logD2MCc100

Debido a que no se observa anisotropía y por el contrario se continúa observando una tendencia lineal, se procede a realizar la transformación para un modelo sin tendencia de 1er grado (Cuadro 5.38). Cabe destacar que el modelo sin tendencia se genera en base a los datos sin la transformación de logaritmo.

Las estadísticas básicas (Cuadro 5.39) y el histograma (Figura 5.85) muestran una asimetría reducida, y una distribución más centrada. La muestra se observa más simétrica y el q-q plot (Figura 5.86) muestra que su distribución es más cercana a la normal.

Coeficiente	**Valor**
Intercepto	-777.03762
Coeficiente de x	0.130500156
Coeficiente de y	0.335731296

Cuadro 5.38: Modelo de Tendencia de 1er grado modificada D2MCc100

Nombre	**Estadísticas**
Número total	100
Distancia max	52.40229003
Distancia min	1
Media	3.01x10^{-}17
Varianza	27.57452021
Desviación estándar	5.251144657
Coeficiente var	1.74x10^{1}7
Rango min	-10.93
1er cuantil	-3.334
Mediana	-0.847
3er cuantil	1.99
Máximo rango	15.71
Asimetría	0.8834086
Curtosis	3.613571082

Cuadro 5.39: Estadísticas básicas sin tendencia de D2MCc100

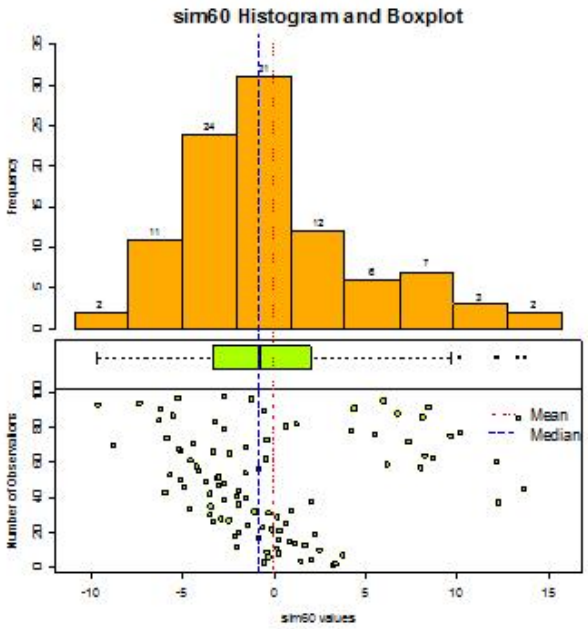

Figura 5.85: Histograma sin tendencia de D2MCc100

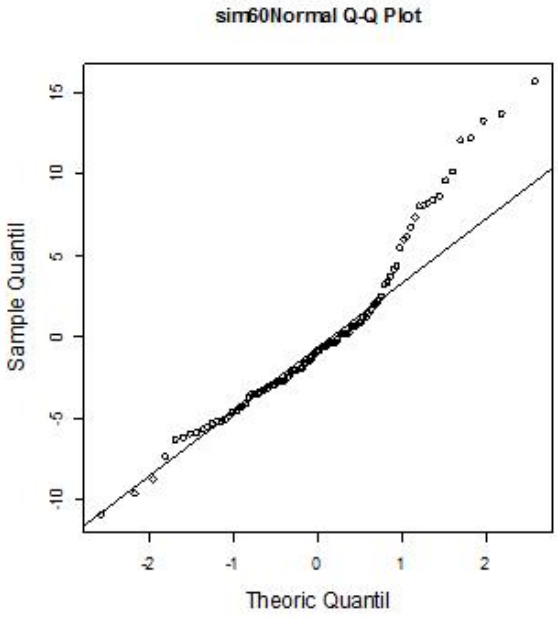

Figura 5.86: Q-Q plot sin tendencia de D2MCc100

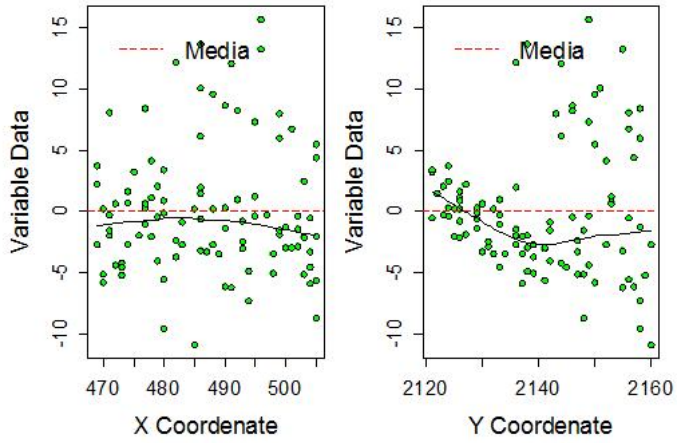

Figura 5.87: Gráfico respecto a las coordenadas s/tend D2MCc100

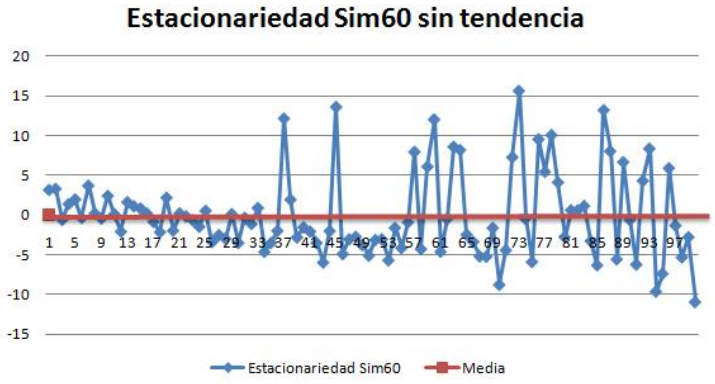

Figura 5.88: Gráfico de estacionariedad sin tendencia de D2MCc100

Se observa un mejor comportamiento en la estacionariedad (Figura 5.88) y se elimina la tendencia observada con respecto al eje y en el gráfico con respecto a las coordenadas (Figura 5.87). Por lo tanto, se procede a realizar el análisis variográfico.

El variograma adireccional (Cuadro 5.40 Figura 5.89) muestra que se eliminó la tendencia lineal y el variograma en cuatro direcciones (Cuadro 5.41 Figura 5.90) no presenta indicios de anisotropía, de igual manera que en el mapa de anisotropía (Figura 5.91). Por lo cual, se procede a realizar el ajuste del modelo de variograma.

Disctancia max	50.91168825
Distancia min	4
Dirección	0°
Tolerancia	90°
Intervalos	13
Distancia Lag	2

Cuadro 5.40: Variograma adireccional sin tendencia de D2MCc100

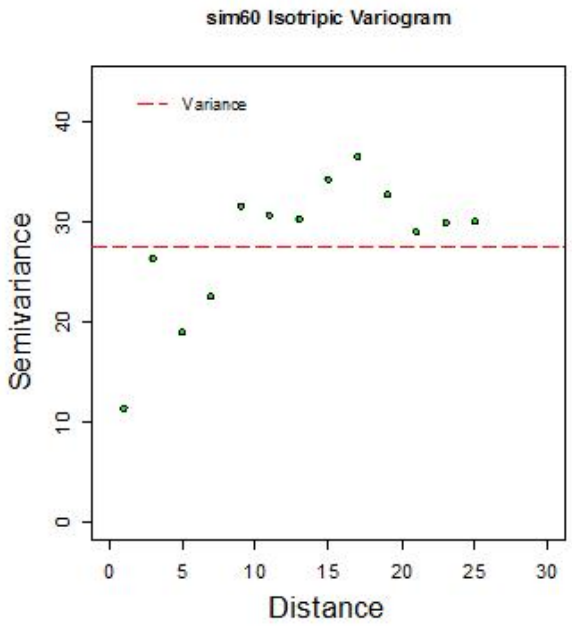

Figura 5.89: Variograma adireccional sin tendencia D2MCc100

Disctancia max	50.91168825
Distancia min	4
Dirección	0°,45°,90°,135°
Tolerancia	22.5°
Intervalos	13
Distancia Lag	2

Cuadro 5.41: Variograma 4 direcciones sin tendencia de D2MCc100

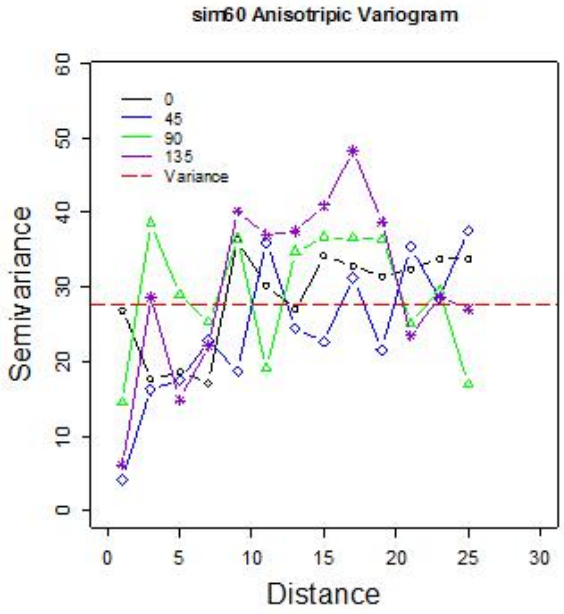

Figura 5.90: Variograma en 4 direcciones sin tendencia D2MCc100

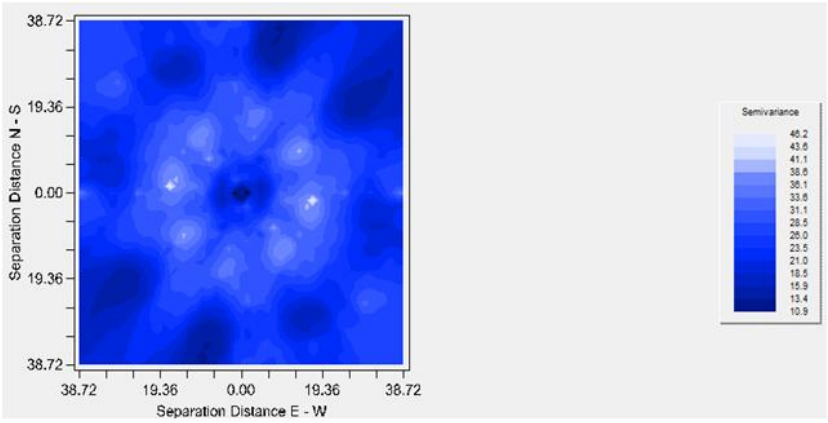

Figura 5.91: Mapa de anisotropía sin tendencia D2MCc100

Se presentan las propuestas de modelos de variograma sin tendencia (Cuadro 5.42 Figura 5.92). Se observa que el modelo esférico es el que mejor se ajusta debido a la menor suma de cuadrados del error, por lo que se procede a realizar el ajuste visual para mejorar la estimación.

Modelo	**Nugget**	**Sill+Nugget**	**Rango**	**SCE**
Exponencial	6.93	32.41	4.56	153
Esférico	14.34	32.14	15.29	148
Gaussiano	17.02	32.17	7.48	270

Cuadro 5.42: Propuestas para modelos de variograma sin tendencia D2MCc100

Modelo	**Nugget**	**Sill+Nugget**	**Rango**	**SCE**
Esférico	13.5	32	15.5	148

Cuadro 5.43: Modelo de variograma elegido sin tendencia D2MCc100

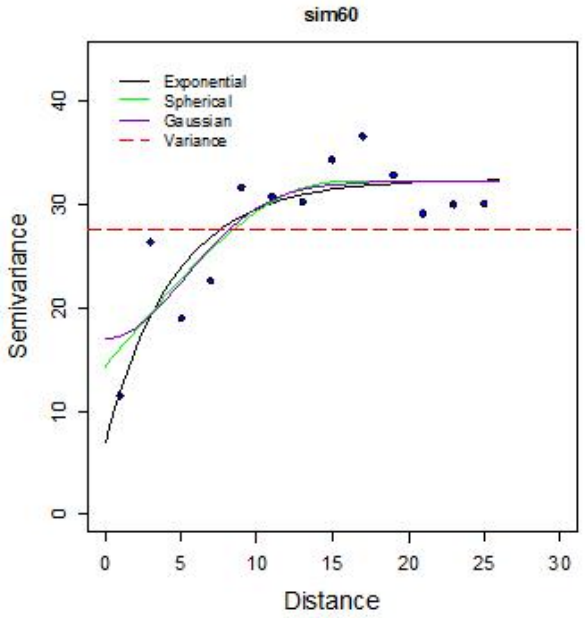

Figura 5.92: Propuestas de variograma sin tendencia de D2MCc100

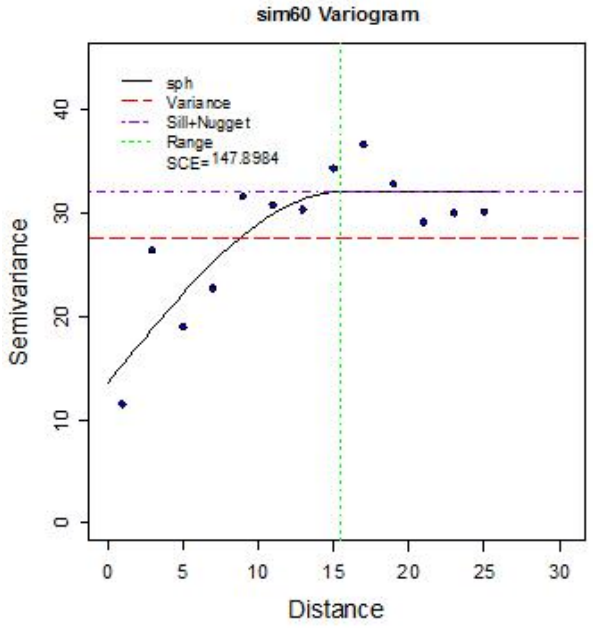

Figura 5.93: Modelo de variograma elegido sin tendencia D2MCc100

Una vez elegido el modelo de variograma (Cuadro 5.43 Figura 5.93), se verifica si es adecuado mediante validación cruzada (Cuadro 5.44). Dentro de las estadísticas de la validación cruzada, la media de los errores es muy cercana a cero, con lo cual muestra que es adecuado. Sin embargo, también se utiliza el análisis de residuales para aceptar el modelo.

Nombre	**s/tend D2MCc100**	**Estimados**	**Error**
Número total	100	100	100
Distancia max	52.40229003	52.40229003	52.40229003
Distancia min	1	1	1
Media	3.01x10^{-}17	0.06226	-0.06226
Varianza	27.57452021	7.765048576	22.09326323
Desviación estándar	5.251144657	2.786583675	4.700347139
Coeficiente var	1.74x10^{1}7	44.75626336	75.49386596
Rango min	-10.93	-4.552	-9.958
1er cuantil	-3.334	-1.854	-2.454
Mediana	-0.847	-0.04907	-0.3777
3er cuantil	1.999	1.542	0.7998
Máximo rango	15.71	8.361	12.9
Asimetría	0.8834086	0.64906393	0.860769605
Curtosis	3.613571082	3.322791255	4.084687604

Cuadro 5.44: Validación cruzada sin tendencia D2MCc100

El gráfico de valores reales contra estimados (Figura 5.94) muestra que los datos se aproximan a una línea de 45°. El histograma (Figura 5.95) y q-q plot (Figura 5.96) muestran que los residuos tienen una distribución relativamente simétrica y cercana a la normal. Por lo tanto, se acepta el modelo y se procede a realizar la estimación con Kriging. Se genera un mapa de estimaciones con kriging (Figura 5.97) a 1km por 1km.

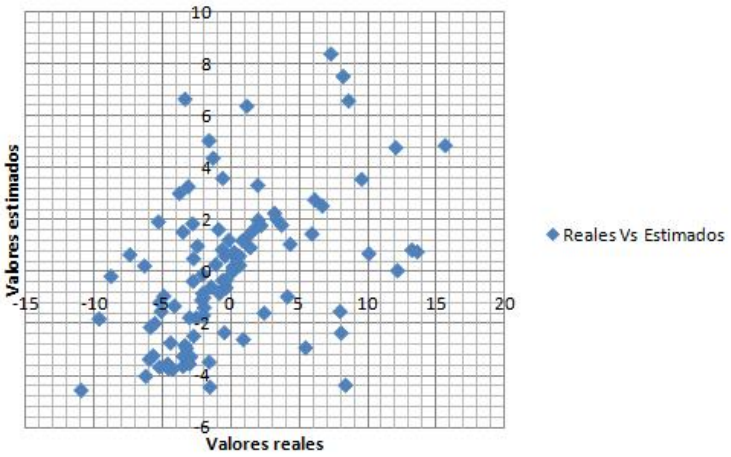

Figura 5.94: Valores reales contra estimados sin tendencia D2MCc100

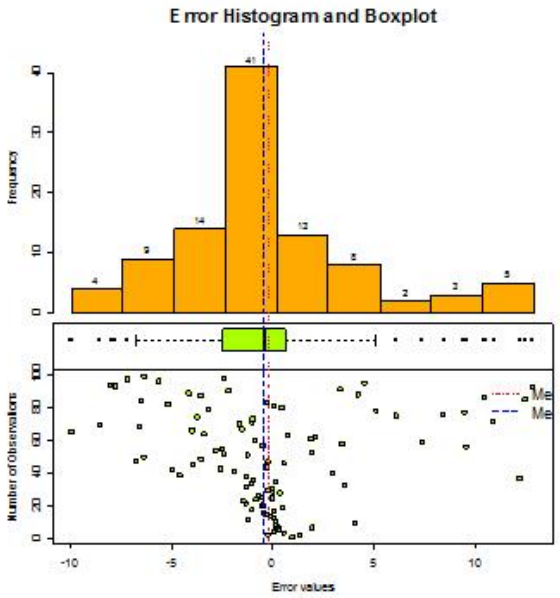

Figura 5.95: Histograma de residuales sin tendencia D2MCc100

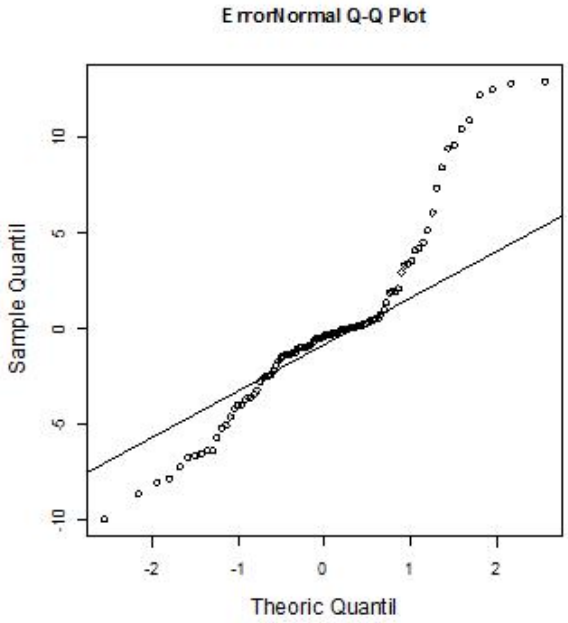

Figura 5.96: Q-Q plot de residuales sin tendencia de D2MCc100

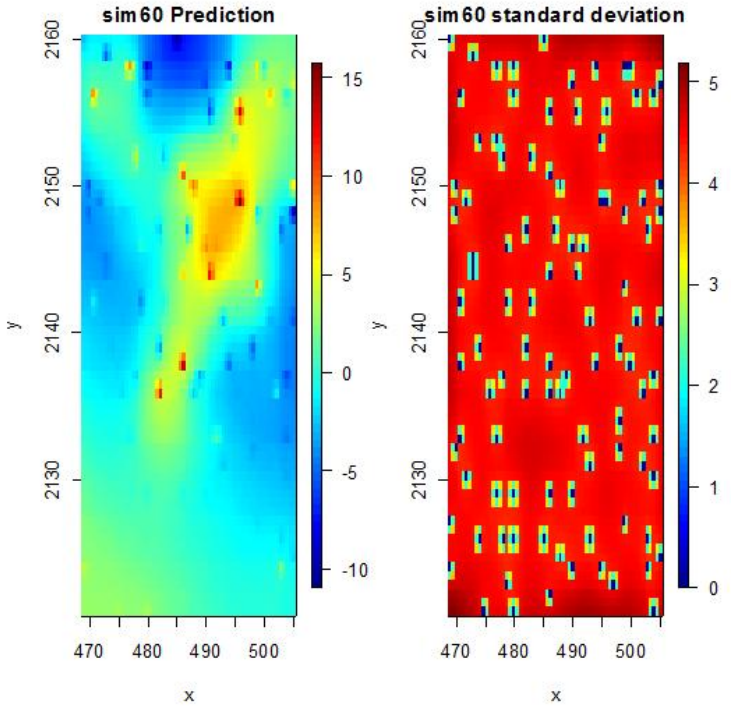

Figura 5.97: Mapa de estimaciones sin tendencia de D2MCc100

5.7. Conclusiones de los modelos de datos 2

- Todos los modelos presentaron asimetría positiva.
- Ningún modelo presenta anisotropía significativa, por lo que en todos los casos se utilizaron modelos de variograma isotrópicos.
- Inicialmente se utiliza una transformación de logaritmo para reducir la asimetría que presentan los datos y cumplir con los supuestos del modelo. Sin embargo, la base de datos del tipo 2 también presenta tendencia lineal, por lo que la transformación final es un modelo de tendencia de 1er grado.
- No fue necesario eliminar los datos atípicos debido a que la transformación reduce la escala y permite que ya no existan datos atípicos.
- El lag de distancia para calcular los variogramas varía dependiendo del espaciamiento que existe entre muestras.
- Casi todos los modelos son ajustados con un modelo esférico y parámetros del variograma muy cercanos.
- Se realizó un ajuste visual con criterio propio para los modelos de variograma presentados.
- Se realizó validación cruzada en todos los casos y los modelos se aceptaron mediante el criterio de media de los errores cercana a cero, y se verificó que los residuales se asemejen una distribución normal o por lo menos simétrica.
- En todos los casos se estimó con Kriging ordinario.
- Cuando se tienen 400 observaciones, el muestreo utilizado pasa a 2do grado de importancia debido a que la información cubre una gran parte de la región.
- Los escenarios con 36 observaciones son mucho más complejos para ajustarles un modelo adecuado y en ocasiones se logra únicamente con ajuste visual.
- Los escenarios con 100 observaciones son los más representativos, ya que tienen una cantidad suficiente de información para estimar de manera adecuada y no tienen tantos datos que la estimación requiera de mucho tiempo.
- Los modelos son adecuados dependiendo de la precisión con la que se realice el ajuste del modelo, así como la información que se tenga a la mano.

Capítulo 6

Aplicación de la metodología a los datos del tipo 3

6.1. Proceso de estimación de los datos originales del tipo 3

Siendo consistente con el proceso de recomendaciones para la aplicación de los procesos de la metodología de geoestadística, es importante mostrar el proceso del cual proviene cada uno de los escenarios propuestos dentro del grupo de datos del tipo 3. Por lo tanto, se comienza el capítulo mostrando dicho proceso. En principio, se muestra la dispersión de los datos del tipo 3 (Figura 6.1) y sus estadísticas básicas (Cuadro 6.1).

Dentro de las estadísticas básicas se observa que la media es mayor que la mediana, lo cual indica una asimetría positiva la cual se confirma con el coeficiente de asimetría. Además, el 50 % de la información se encuentra dentro del intervalo [0.25, 0.38] y el restante 50 % se encuentra en el intervalo (0.38, 27.5].

Con el histograma (Figura 6.2) resalta la asimetría y los 128 datos atípicos. El q-q plot (Figura 6.3) muestra que no se asemeja a una distribución normal. Por lo tanto, es necesario aplicar una transformación logarítmica para reducir la escala y mejorar la simetría. Se inicia nuevamente el análisis exploratorio de datos con la variable ya transformada.

Nombre	Estadísticas
Número total	1480
Distancia max	53.07541804
Distancia min	1
Media	5.413
Varianza	41.6459877
Desviación estándar	6.453370259
Coeficiente var	1.192243327
Rango min	0.25
1er cuantil	0.3801
Mediana	3.56
3er cuantil	7.04
Máximo rango	27.5
Asimetría	1.428529561
Curtosis	4.188130993

Cuadro 6.1: Estadísticas básicas D3c1480

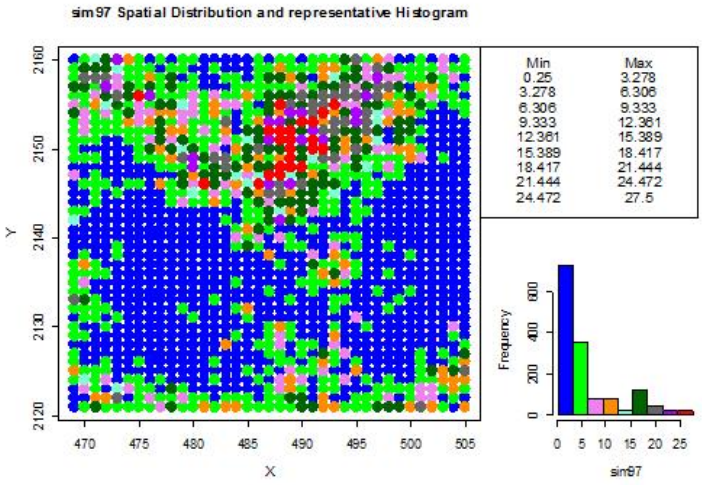

Figura 6.1: Distribución de D3c1480

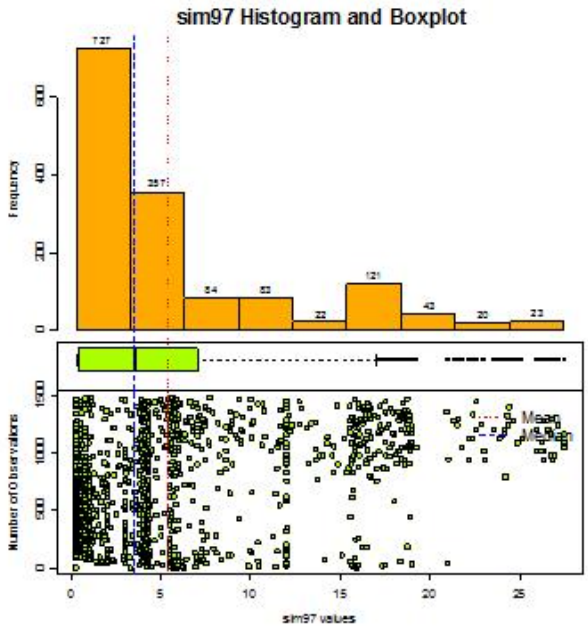

Figura 6.2: Histograma de D3c1480

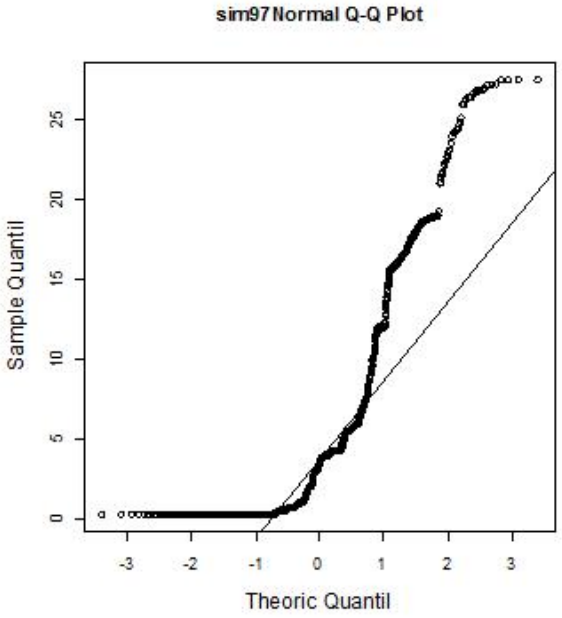

Figura 6.3: Q-Q plot de D3c1480

Una vez realizada la transformación logarítmica, las estadísticas básicas muestran una reducción en la escala, ahora el 75 % de la información se encuentra en el intervalo [-1.38, 1.95] y el restante 25 % se encuentra en el intervalo de (1.95, 3.314]. Además, la media ahora es menor que la mediana, pero son mucho más cercanas, lo cual indica una asimetría negativa y se confirma con el coeficiente de asimetría el cual es muy cercano a cero.

El histograma (Figura 6.5) y q-q plot (Figura 6.6) confirman que la distribución es más simétrica y ya no existen datos atípicos. Se procede a verificar tendencia y estacionariedad.

Nombre	**Estadísticas**
Número total	1480
Distancia max	53.07541804
Distancia min	1
Media	0.7364
Varianza	2.45770813
Desviación estándar	1.567707923
Coeficiente var	2.128800933
Rango min	-1.386
1er cuantil	-0.9673
Mediana	1.27
3er cuantil	1.952
Máximo rango	3.314
Asimetría	-0.123910243
Curtosis	1.582558317

Cuadro 6.2: Estadísticas básicas logD3c1480

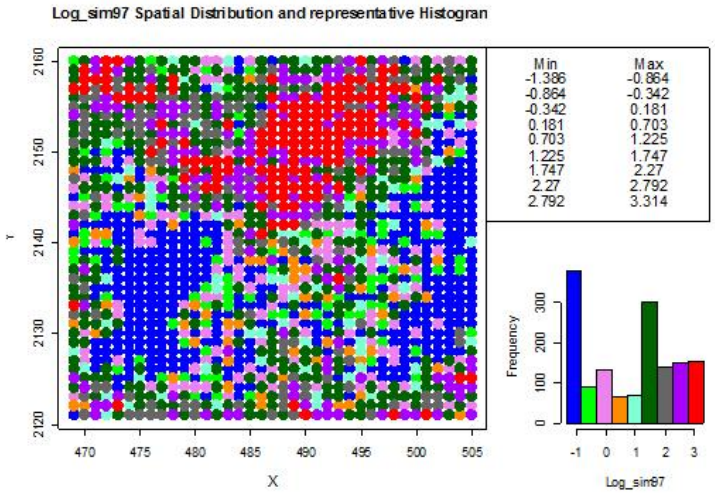

Figura 6.4: Distribución de logD3c1480

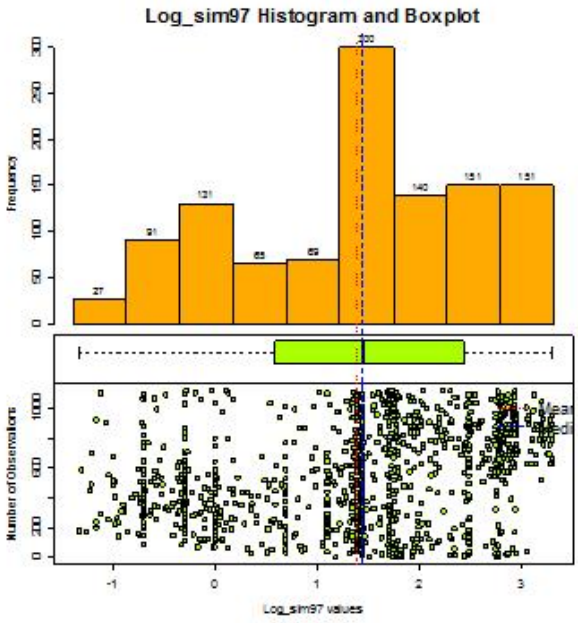

Figura 6.5: Histograma de logD3c1480

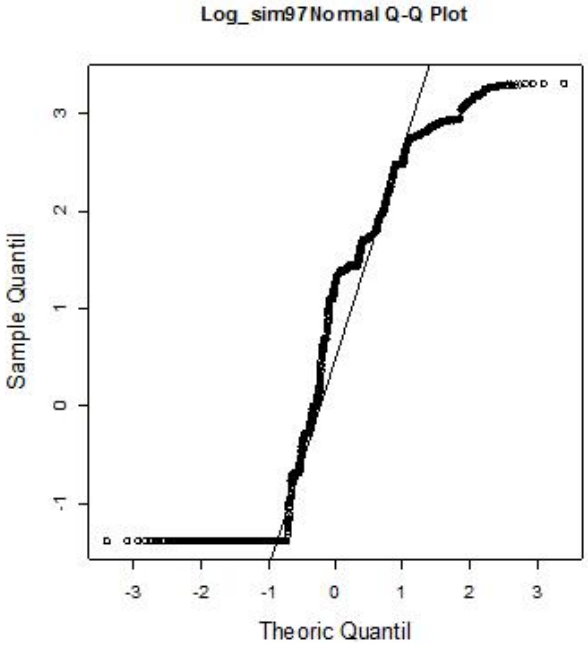

Figura 6.6: Q-Q plot de logD3c1480

En el gráfico con respecto a las coordenadas (Figura 6.6) no se observa tendencia. En el gráfico de estacionariedad (Figura 6.8) se observa una muestra que si se mantiene estacionaria.

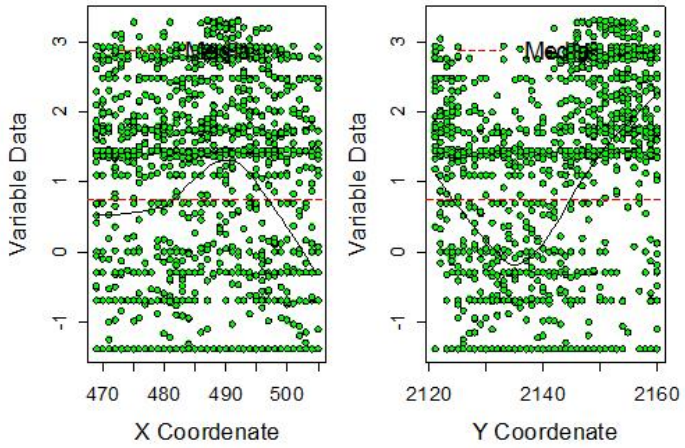

Figura 6.7: Gráfico respecto a las coordenadas de logD3c1480

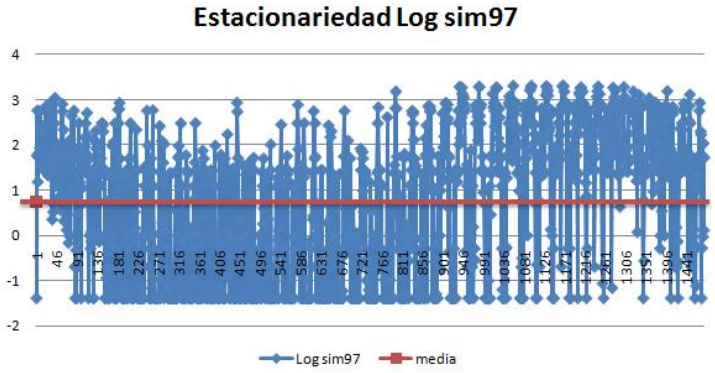

Figura 6.8: Gráfico de estacionariedad de logD3c1480

Ahora se procede a realizar el análisis variográfico. Dentro del variograma adireccional (Cuadro 6.3 Figura 6.9) no se observa comportamiento de h^2, por lo que se confirma que no hay tendencia. Sin embargo, en el variograma en cuatro direcciones (Cuadro 6.4 Figura 6.10) se observa una diferencia considerable en los alcances de los variogramas para cada una de las direcciones, por lo tanto se verificará si se tiene la característica de anisotropía.

Disctancia max	53.07541804
Distancia min	1
Dirección	0°
Tolerancia	90°
Intervalos	25
Distancia Lag	1

Cuadro 6.3: Variograma adireccional de logD3c1480

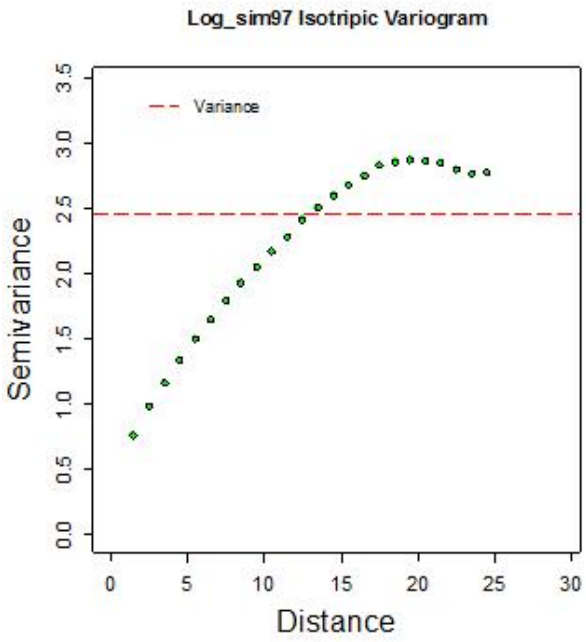

Figura 6.9: Variograma adireccional de logD3c1480

Disctancia max	53.07541804
Distancia min	1
Dirección	0°,45°,90°,135°
Tolerancia	22.5°
Intervalos	25
Distancia Lag	1

Cuadro 6.4: Variograma 4 direcciones logD3c1480

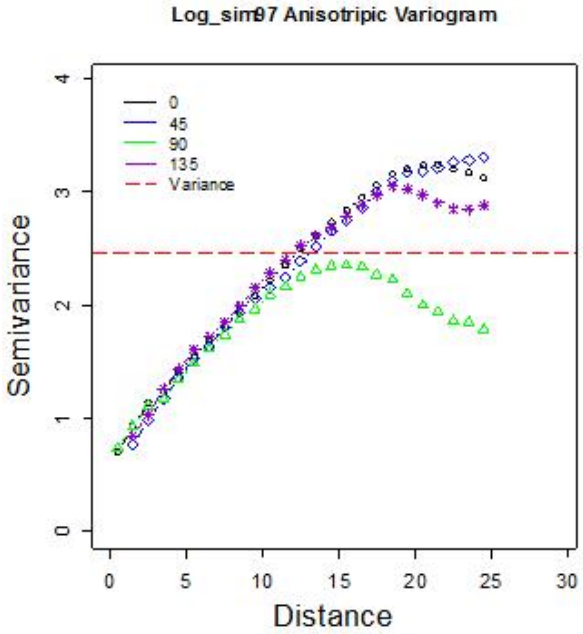

Figura 6.10: Variograma en 4 direcciones de logD3c1480

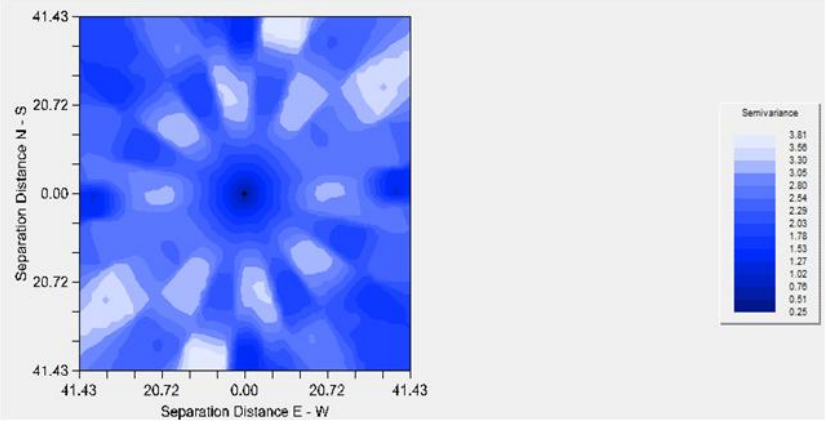

Figura 6.11: Mapa de anisotropía de logD3c1480

A pesar de que en el mapa de anisotropía (Figura 6.11) no se observan elipses, el variograma en 4 direcciones mostró un indicio. Debido a ésto, se verifica la anisotropía realizando numerosos variogramas en diferentes direcciones hasta encontrar las direcciones de mínimo y máximo alcance, las cuales deben estar con una diferencia de 90°.

Después de realizadas las pruebas, se encuentra que sí existe una ligera anisotropía en direcciones de 110° y 20°, por lo tanto se ajusta el variograma correspondiente a la dirección de mínimo (Cuadro 6.5 Figura 6.12) y máximo alcance (Cuadro 6.6 Figura 6.13).

Modelo	**Nugget**	**Sill+Nugget**	**Rango**	**SCE**
Esférico	0.6	2.35	14.7	0.32

Cuadro 6.5: Variograma anisotrópico con eje menor en dirección 110°

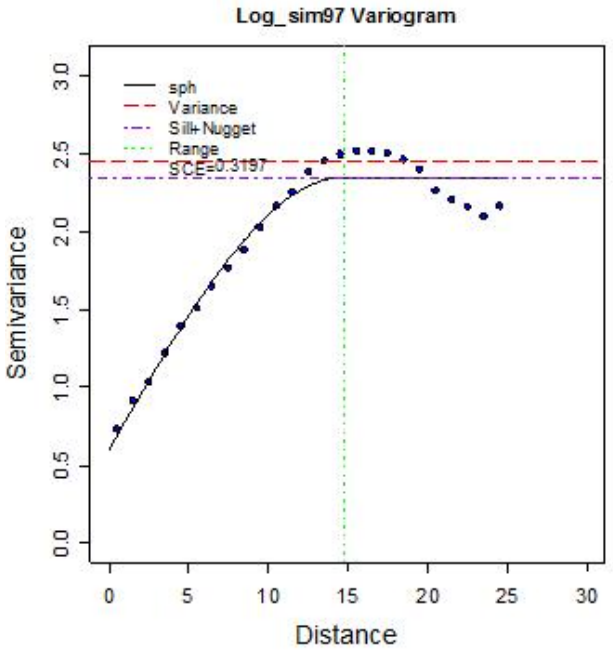

Figura 6.12: Variograma anisotrópico con eje menor a 110°

Modelo	**Nugget**	**Sill+Nugget**	**Rango**	**SCE**
Esférico	0.6	3.3	24.5	0.0552

Cuadro 6.6: Variograma anisotrópico con eje mayor en dirección 20°

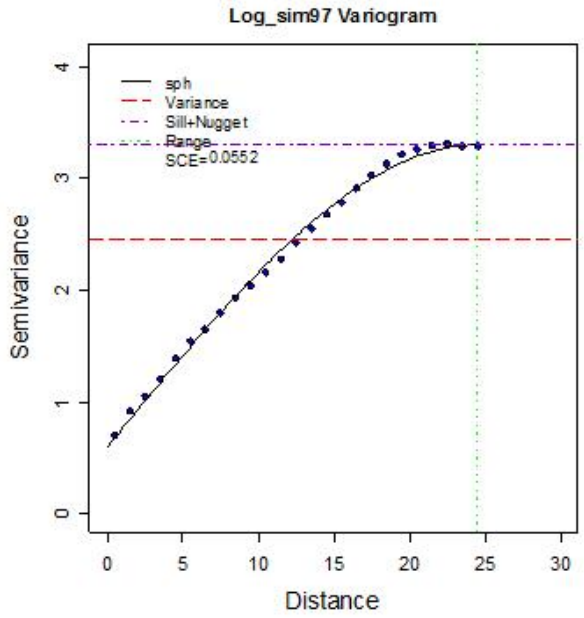

Figura 6.13: Variograma anisotrópico con eje Mayor a 20°

Una vez definidos los alcances y ajustes para cada una de las direcciones de anisotropía, se realiza la validación cruzada (Cuadro 6.7) sobre el variograma anisotrópico de mayor alcance.

Se observa que la media es cercana a cero y la distribución de los estimados es similar a la distribución de los datos originales.

Nombre	**logD3c1480**	**Estimados**	**Error**
Número total	1480	1480	1480
Distancia max	53.07541804	53.07541804	53.07541804
Distancia min	1	1	1
Media	0.7364	0.7359	0.0004782
Varianza	2.45770813	1.668655937	0.761039226
Desviación estándar	1.567707923	1.29176466	0.872375622
Coeficiente var	2.128800933	1.755235635	1824.106207
Rango min	-1.386	-1.47	-3.236
1er cuantil	-0.9673	-0.3565	-0.486
Mediana	1.27	0.7571	0.01127
3er cuantil	1.952	1.822	0.5251
Máximo rango	3.314	3.229	2.797
Asimetría	-0.123910243	0.006705647	-0.1974349
Curtosis	1.582558317	1.825719933	3.671528802

Cuadro 6.7: Validación cruzada logD3c1480

El gráfico de valores reales contra estimados (Figura 6.14) muestra que los datos se aproximan a una línea de 45°.

El histograma (Figura 6.15) muestra que los residuales son simétricos y el q-q plot (Figura 6.16) muestra que se asemejan a una distribución normal. Por lo tanto, se procede a realizar la estimación con Kriging.

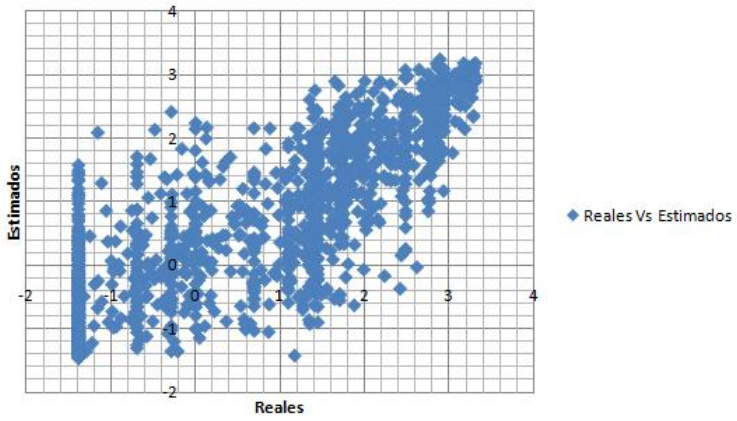

Figura 6.14: Gráfico de valores reales contra estimados de logD3c1480

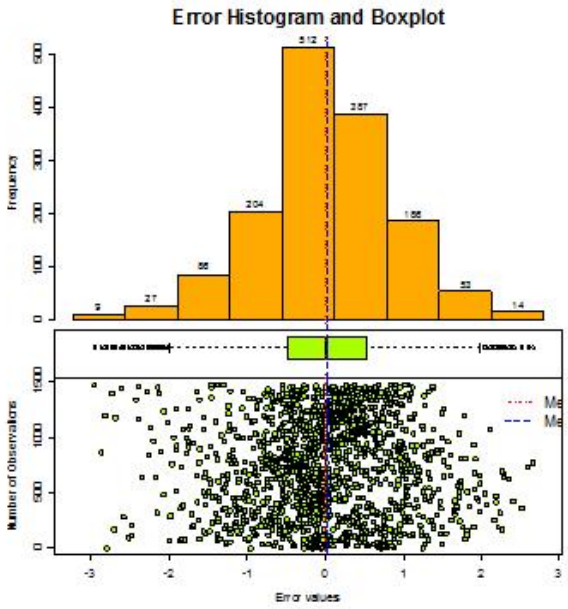

Figura 6.15: Histograma de errores de logD3c1480

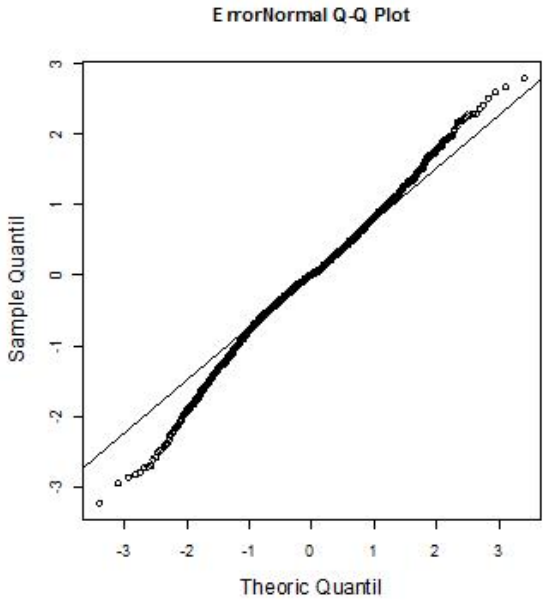

Figura 6.16: Q-Q plot de errores de logD3c1480

Se obtiene un mapa de valores estimados a 1km por 1km (Figura 6.17), con un índice de anisotropía (alcance menor/alcance mayor) de 0.6.

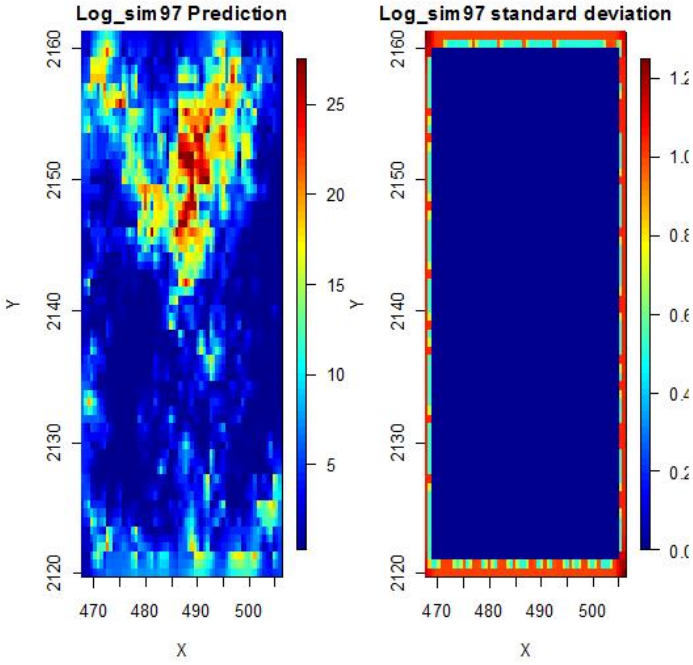

Figura 6.17: Mapa de estimaciones con kriging de logD3c1480

6.2. Resumen de los modelos de datos 3

Para resumir los modelos de datos del tipo 3 es importante destacar las características de los escenarios realizados para los datos del tipo 3. Se presenta una tabla con el resumen de los modelos ajustados para cada uno de los escenarios y se mencionan las características en común que tienen todos los modelos. Las siguientes observaciones ocurrieron en todos los escenarios:

1. La distribución de las muestras tiene asimetría positiva y posterior a la transformación se tiene una ligera asimetría negativa.
2. En todos los escenarios de datos 3, la muestra es tratada con una transformación de logaritmo para reducir la escala y asimetría.
3. En todos los escenarios los variogramas en 4 direcciones son calculados con el mismo número de intervalos y tolerancia de 22.5°, así como la misma distancia del lag que en el caso del variograma adireccional.
4. En particular, el caso excluyente es el D3MCc036 ya que se utilizó una tolerancia de 30° para mejorar la información calculada en los variogramas.
5. Las bases de datos de los escenarios del tipo 3 presentan indicadores de anisotropía.
6. Las bases de datos de los escenarios del tipo 3 no cumplen normalidad.
7. Todos los modelos de los escenarios del tipo 3 pasaron el análisis de residuales, es decir, la línea de 45°, el histograma y q-q plot son muy cercanos a la distribución normal.
8. En todos los modelos de los escenarios del tipo 3 se realizó la estimación por medio de Kriging ordinario.

En la figura 6.18 se muestra un resumen de la información de los variogramas de todos los escenarios de Datos 3. También, se anexa información del análisis variográfico de los escenarios:

1. Se observa que los modelos ajustados en todos los escenarios son esféricos.
2. El nugget va de un rango de entre [0.3, 1] para todos los escenarios.
3. La meseta se encuentra en un intervalo de [2.9, 3.4] para todos los escenarios.
4. Y el alcance varía en un intervalo de [13.4, 30] para todos los escenarios.

Por otro lado, es importante mencionar que los modelos ajustados se realizaron de manera completamente independiente y sin conocer el modelo de los datos originales del tipo 3, lo cual destaca la notoria similitud que existe entre los modelos que se utilizaron para realizar las estimaciones con Kriging.

ANÁLISIS VARIOGRÁFICO

Escenario	Variograma adireccional (No. intervalos, distancia de lag en Km)	Variog en 4 direcc (Distancia del lag en Km)	Ejes de anisotropía	Modelo	Nugget	Meseta (Sill+Nugget)	Alcance
Datos 3	**inter 25, lag 1Km**	**lag 1 Km**	**Mayor 20, menor 110**	**Esférico**	**0.6**	**3.3**	**24.5**
D3MRc036	inter 10, lag 6 Km	lag 6 Km	Mayor 45, menor 135	Esférico	0.4	3	27
D3MRc100	inter 11, lag 4 Km	lag 4 Km	Mayor 45, menor 135	Esférico	0.3	3.2	19
D3MRc400	inter 13, lag 2 Km	lag 2 Km	Mayor 20, menor 110	Esférico	0.7	3.2	24.4
D3MAc036	inter 10, lag 3 Km	lag 3 Km	mayor 130, menor 40	Esférico	0.6	3.4	20
D3MAc100	inter 12, lag 3 Km	lag 3 Km	Mayor 45, menor 135	Esférico	1	2.9	30
D3MAc400	inter 13, lag 2 Km	lag 2 Km	Mayor 20, menor 110	Esférico	0.5	3.28	24
D3MCc036	inter 10, lag 3 Km	lag 3 Km, Tolerancia 30°	mayor 68, menor 158	Esférico	0	3.1	13.4
D3MCc100	inter 11, lag 2 Km	lag 2 Km	mayor 120, menor 30	Esférico	0.4	3.17	22
D3MCc400	inter 13, lag 2 Km	lag 2 Km	Mayor 20, menor 110	Esférico	0.5	3.3	24

Figura 6.18: Resumen de los modelos de variograma por escenario Datos 3

6.3. Comparación respecto a los datos originales del tipo 3

Dentro de los diferentes escenarios presentados destaca la importancia en la identificación de características que afectan la estimación y su adecuado tratamiento dentro de la metodología. Características como la anisotropía influyen de manera directa en la estimación ya que introducen los modelos anisotrópicos para el ajuste del variograma, así como el radio de búsqueda contra el cual se realizan las estimaciones. En la práctica no se conoce la cantidad de datos, el tipo de muestreo o las características que pueden presentar las bases de datos con las cuales se tiene que trabajar. Cada uno de los escenarios muestra diversas dificultades para llegar a la estimación. Algunas de las dificultades más relevantes presentadas en los escenarios fueron las siguientes:

1. *Escenarios con 36 observaciones D3MRc036, D3MAc036 y D3MCc036:*

   Para los escenarios con 36 observaciones el ajuste del variograma fue casi completamente visual debido a que no se cuenta con información suficiente para obtener suficientes intervalos del variograma de manera confiable. El variograma está calculado pobremente y por lo tanto no permite definir bien sus parámetros. En los escenarios con 36 observaciones, el muestreo aleatorio es uno de los dos escenarios donde el muestreo influye directamente en la estimación, debido a que la ubicación de las observaciones genera un problema al encontrar los ejes de anisotropía ya que se muestran en una dirección que no es la de los datos originales del tipo 3 y por lo tanto el ajuste del modelo de variograma es afectado y por consiguiente

la estimación también. El caso del muestreo combinado es un caso donde al encontrar los ejes de anisotropía, el alcance no es el óptimo y por lo tanto al realizar las estimaciones no permite estimar adecuadamente sobre toda la región. Por último, el caso del muestreo regular es el mejor ajustado ya que es el más cercano a los datos originales del tipo 3 y la estimación es cercana. Por lo tanto, cuando se tiene una característica con tanta influencia en la estimación, es importante tener mayor número de datos para obtener un modelo adecuado y una estimación exitosa.

2. *Escenarios con 100 observaciones D3MRc100, D3MAc100 y D3MCc100:*

   Los escenarios con 100 observaciones son los más representativos dentro de la base de datos del tipo 3. Las bases con esta cantidad de observaciones permiten realizar los cálculos con suficiente rapidez. Sin embargo, las características no siempre son fácilmente identificadas, e incluso cuando son identificadas no son fácilmente tratables. En el caso de la anisotropía es importante encontrar acertadamente los ejes de anisotropía para realmente mejorar la estimación. El escenario D3MRc100 es el más parecido a los datos originales del tipo 3, debido a que el muestreo regular permite identificar de manera más acertada la dirección de anisotropía, mejorar el ajuste del modelo de variograma y la estimación. El escenario de D3MCc100 es el otro escenario que no coincide con la dirección de anisotropía de los datos originales del tipo 3, sin embargo al realizar la estimación el modelo sí mapea de manera cercana, lo cual es la diferencia notable contra el escenario de D3MAc036, ya que aún cuando la dirección no es la más cercana a la original, es posible estimar más acertadamente debido a la cantidad de observaciones con las que se cuenta. El escenario de D3MAc100 resalta que aunque se tenga mayor número de observaciones, no necesariamente son las ubicaciones más significativas, en este escenario las observaciones permiten encontrar los ejes de anisotropía de manera cercana a los datos originales del tipo 3 pero las estimaciones no son acertadas ya que no se cuenta con ubicaciones que permitan encontrar las regiones dentro del área con valores altos y por lo tanto no es una estimación exitosa.

3. *Escenarios con 400 observaciones D3MRc400, D3MAc400 y D3MCc400:*

   En los escenarios con 400 observaciones los ejes de anisotropía son encontrados adecuadamente. La dirección de anisotropía se vuelve visible y correcta respecto a los datos originales del tipo 3. Conforme aumentó la cantidad de información, el tipo de muestreo resultó menos importante, debido a que se cuenta con más información sobre toda la región. El mapa de estimación del escenario de D3MCc400 resultó el más cercano a los datos originales del tipo 3. Sin embargo, los mapas de los escenarios D3MRc400 y D3MAc400 son también muy cercanos a los datos originales del tipo 3. Cuando se tienen tantas muestras también se observa que los variogramas son muy parecidos entre los distintos tipos de muestreo y se asemejan al variograma de los datos originales del tipo 3, por lo que el ajuste visual del modelo se vuelve muy similar y por consiguiente la estimación también lo es. Sin embargo, cuando se cuenta con mayor número de datos el tiempo para realizar la estimación aumenta considerablemente respecto de los demás.

A continuación se muestra el procedimiento a seguir con tres escenarios representativos. Se utilizan los escenarios de cien observaciones para cada uno de los tipos de muestreo dentro del grupo de datos originales del tipo 3 y se realiza el análisis exploratorio de datos, el análisis estructural y la estimación con Kriging.

Por lo tanto, los escenarios representativos son:

- D3MRc100
- D3MAc100
- D3MCc100

6.4. Base de datos del tipo 3 con muestreo de malla regular y 100 observaciones (D3MRc100)

Se inicia el escenario representativo de D3MRc100 mostrando la distribución de los datos (Figura 6.19) y sus estadísticas básicas (Cuadro 6.8).

En las estadísticas básicas se observa que la media es mayor que la mediana, lo que indica una asimetría positiva que se confirma con el coeficiente de asimetría. El 50 % de los datos se encuentran en el intervalo [0.25, 3.6] mientras que el restante 50 % se encuentra en el intervalo (3.6, 27.5].

El histograma (Figura 6.20) muestra la asimetría que tienen los datos y destaca un dato atípico. El q-q plot (Figura 6.21) muestra que no se distribuyen normal.

Nombre	**Estadísticas**
Número total	100
Distancia max	50.91168825
Distancia min	4
Media	5.554
Varianza	42.39974255
Desviación estándar	6.51150847
Coeficiente var	1.172471963
Rango min	0.25
1er cuantil	0.25
Mediana	3.609
3er cuantil	8.077
Máximo rango	27.5
Asimetría	1.203558485
Curtosis	3.427444797

Cuadro 6.8: Estadísticas básicas D3MRc100

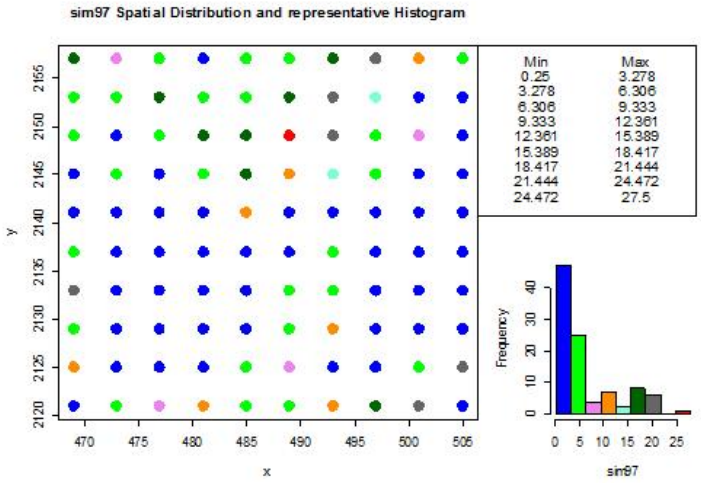

Figura 6.19: Distribución de D3MRc100

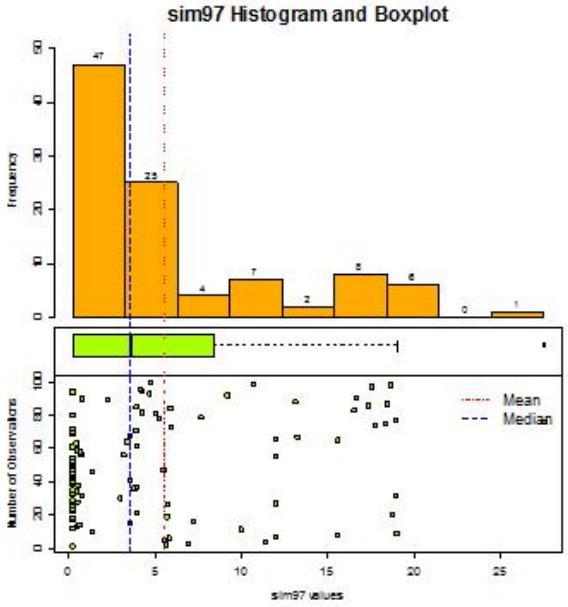

Figura 6.20: Histograma de D3MRc100

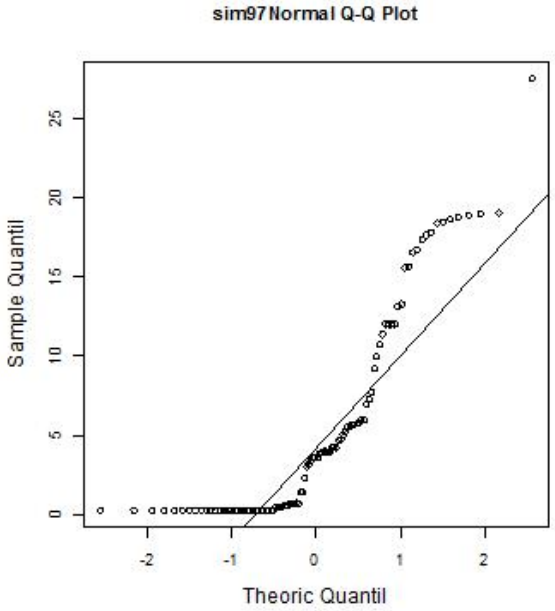

Figura 6.21: Q-Q plot de D3MRc100

Por lo tanto, se propone realizar una transformación logarítmica para reducir la escala y mejorar la asimetría de los datos.

Se inicia nuevamente el análisis exploratorio de datos y una vez realizada la transformación, se muestra la distribución de los datos (Figura 6.22) y sus estadísticas básicas (Cuadro 6.9).

Nombre	**Estadísticas**
Número total	100
Distancia max	50.91168825
Distancia min	4
Media	0.6711
Varianza	2.809894191
Desviación estándar	1.676273901
Coeficiente var	2.497644695
Rango min	-1.386
1er cuantil	-1.386
Mediana	1.284
3er cuantil	2.086
Máximo rango	3.314
Asimetría	-0.087622173
Curtosis	1.40402425

Cuadro 6.9: Estadísticas básicas logD3MRc100

En las estadísticas básicas se observa que la media ahora es menor que la mediana, pero son más cercanas. Se muestra una asimetría negativa pero muy cercana a cero. Ahora el 75 % de la información se encuentra en el intervalo [-1.38, 2] y el restante 25 % se encuentra en el intervalo (2, 3.314].

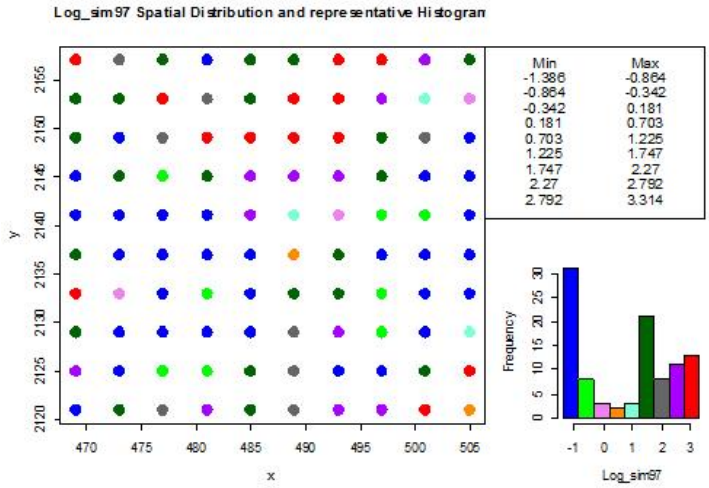

Figura 6.22: Distribución de logD3MRc100

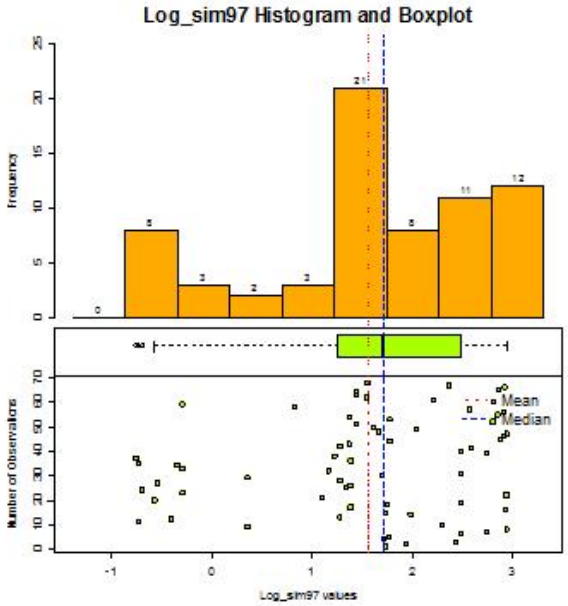

Figura 6.23: Histograma de logD3MRc100

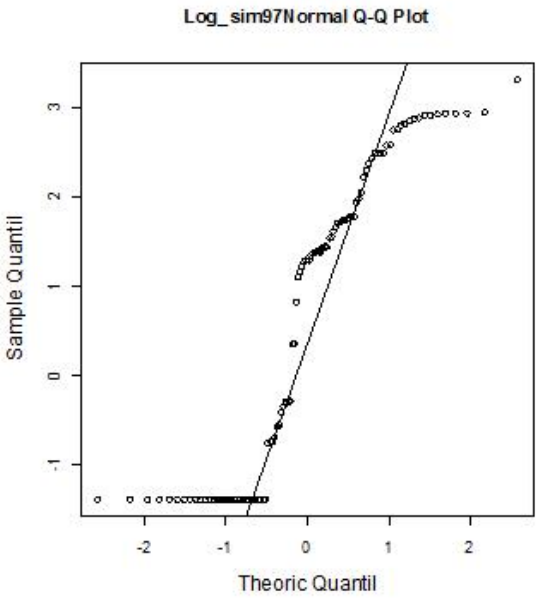

Figura 6.24: Q-Q plot de logD3MRc100

El histograma (Figura 6.23) muestra la corrección en la asimetría aunque el q-q plot (Figura 6.24) no muestra que la distribución sea normal. Sin embargo, es suficiente mejorar la asimetría para continuar con la metodología. Por lo que se procede a realizar el gráfico con respecto a las coordenadas (Figura 6.25) y el gráfico de estacionariedad (Figura 6.26).

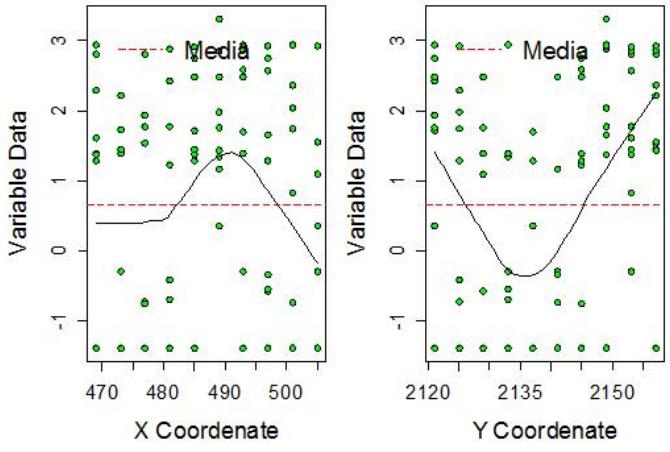

Figura 6.25: Gráfico con respecto a las coordenadas de logD3MRc100

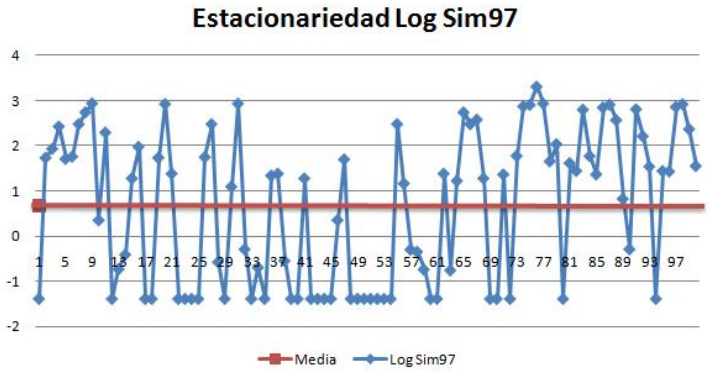

Figura 6.26: Estacionariedad de logD3MRc100

En el gráfico de estacionariedad no se observan indicios significativos de no estacionariedad. Y en el gráfico respecto a las coordenadas no se observa tendencia. Por lo que se continúa con el análisis variográfico.

El variograma adireccional (Cuadro 6.10 Figura 6.27) no muestra comportamiento de h^2, lo cual confirma que no se tienen indicadores contundentes de tendencia. Sin embargo, el variograma en cuatro direcciones (Cuadro 6.11 Figura 6.28) muestra alcances a diferentes distancias de manera significativa, lo cual es indicio de anisotropía. Entonces se realiza el mapa de anisotropía (Figura 6.29).

Disctancia max	50.91168825
Distancia min	4
Dirección	0°
Tolerancia	90°
Intervalos	11
Distancia Lag	4

Cuadro 6.10: Variograma adireccional de logD3MRc100

Disctancia max	50.91168825
Distancia min	4
Dirección	0°,45°,90°,135°
Tolerancia	22.5°
Intervalos	11
Distancia Lag	4

Cuadro 6.11: Variograma 4 direcciones logD3MRc100

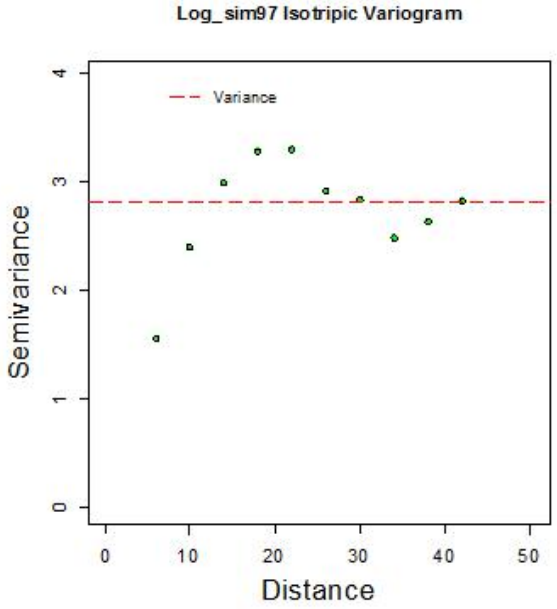

Figura 6.27: Variograma de logD3MRc100

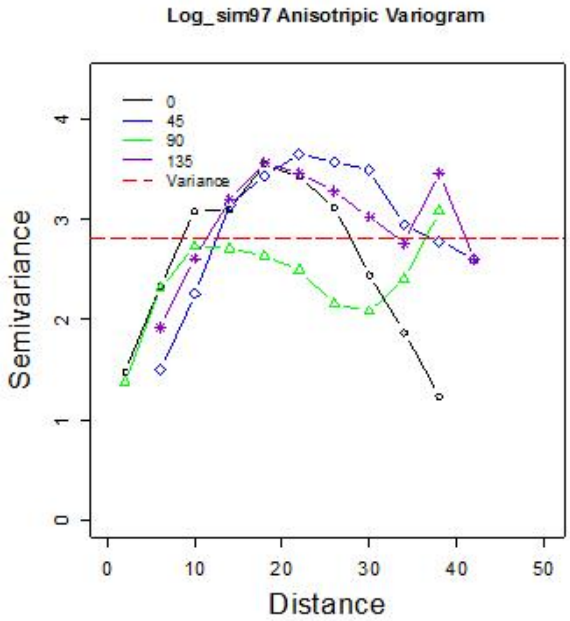

Figura 6.28: Variograma en 4 direcciones logD3MRc100

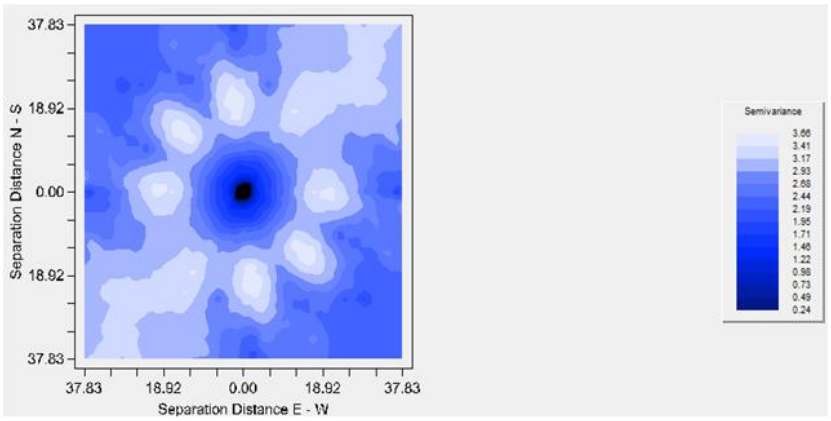

Figura 6.29: Mapa de anisotropía logD3MRc100

El mapa de anisotropía (Figura 6.29) muestra elipses con un eje mayor en las direcciones cercanas a 45°. Por lo tanto se confirma que es necesario encontrar el eje de mayor alcance y el eje de menor alcance para posteriormente utilizar un variograma anisotrópico en la estimación. Se realizan numerosos variogramas en distintas direcciones para encontrar las direcciones de mínimo y máximo alcance.

Se concluye que el eje menor se encuentra en la dirección de 135° y se realiza un ajuste visual (Cuadro 6.12 Figura 6.30). De igual manera, después de diversos variogramas se elige el variograma del eje mayor en dirección de 45°(Cuadro 6.13 Figura 6.31) como se observa en el mapa de anisotropía. Cabe destacar que las direcciones de mínimo y máximo alcance deben ser perpendiculares.

Modelo	**Nugget**	**Sill+Nugget**	**Rango**	**SCE**
Esférico	0.55	3.16	16.5	0.878

Cuadro 6.12: Variograma eje menor 135° logD3MRc100

Modelo	**Nugget**	**Sill+Nugget**	**Rango**	**SCE**
Esférico	0.3	3.2	19	1.16

Cuadro 6.13: Variograma eje Mayor 45° logD3MRc100

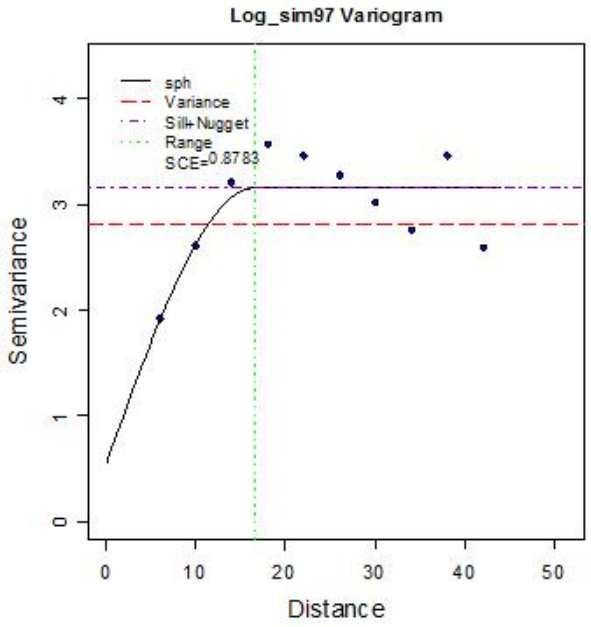

Figura 6.30: Variograma anisotrópico con eje menor a 135° logD3MRc100

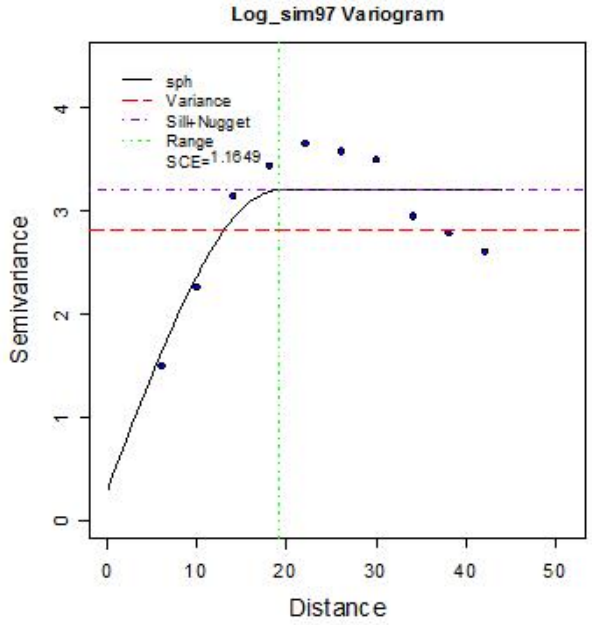

Figura 6.31: Variograma anisotrópico con eje Mayor a 45° logD3MRc100

Una vez realizados los ajustes para las direcciones de anisotropía, se elige el modelo de mayor alcance como modelo de variograma anisotrópico y se realiza la validación cruzada (Cuadro 6.14) para verificar si es adecuado.

Se observa que la media de los errores es cercana a cero y la distribución de los estimados es similar a la de los datos transformados. También se realiza el análisis de residuales con el cual se muestra en el gráfico de valores reales contra estimados (Figura 6.32) que los datos parecen aproximarse a una línea de 45°.

Nombre	**logD3MRc100**	**Estimados**	**Error**
Número total	100	100	100
Distancia max	50.91168825	50.91168825	50.91168825
Distancia min	4	4	4
Media	0.6711	0.661	0.0101
Varianza	2.809894191	1.588401575	1.452753194
Desviación estándar	1.676273901	1.260318045	1.205302117
Coeficiente var	2.497644695	1.906560509	119.347653
Rango min	-1.386	-1.716	-3.291
1er cuantil	-1.386	-0.2079	-0.6826
Mediana	1.284	0.7287	0.03838
3er cuantil	2.086	1.605	0.8612
Máximo rango	3.314	3.105	2.55
Asimetría	-0.087622173	-0.041830911	-0.38992478
Curtosis	1.40402425	2.136050232	3.284185391

Cuadro 6.14: Validación cruzada logD3MRc100

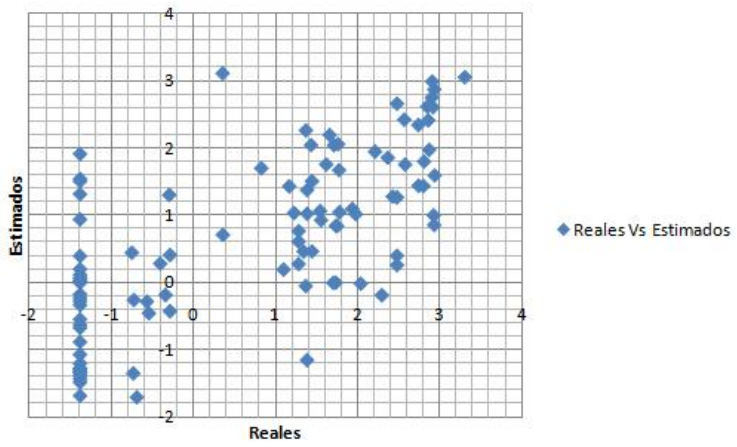

Figura 6.32: Valores reales contra estimados de logD3MRc100

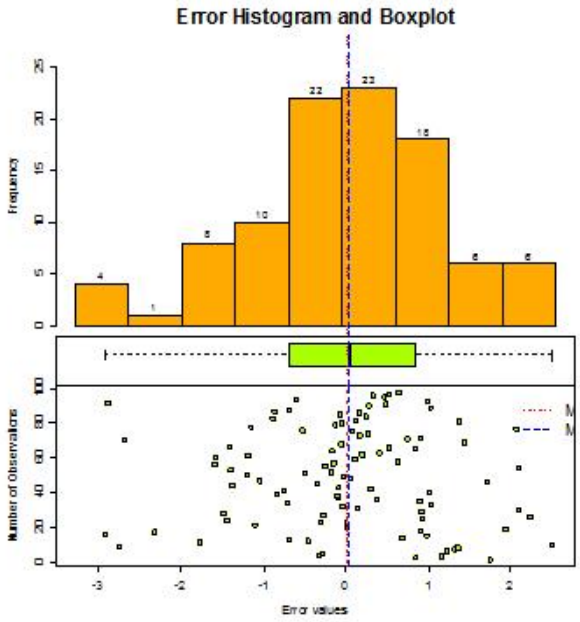

Figura 6.33: Histograma de errores de logD3MRc100

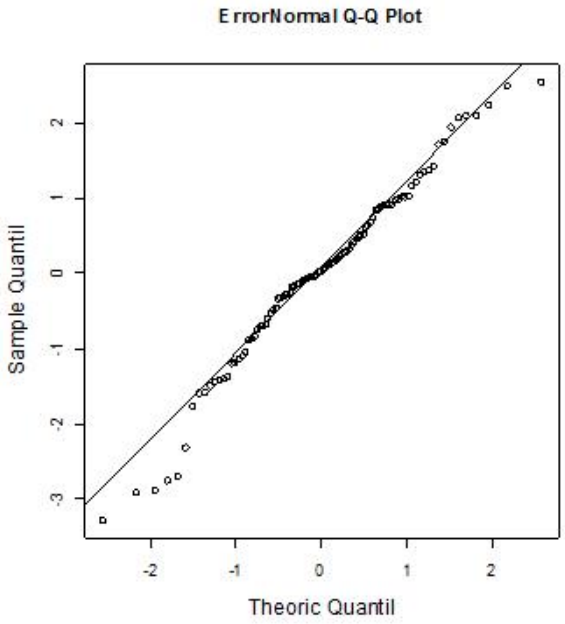

Figura 6.34: Q-Q plot de errores de logD3MRc100

Con el histograma (Figura 6.33) y q-q plot (Figura 6.34) se muestra que los errores tienen una distribución suficientemente simétrica y cercana a la normal. Por lo tanto, se procede a realizar la estimación con kriging.

Se obtiene un mapa de estimaciones a 1km por 1km (Figura 6.35), con un índice de anisotropía de 0,8684.

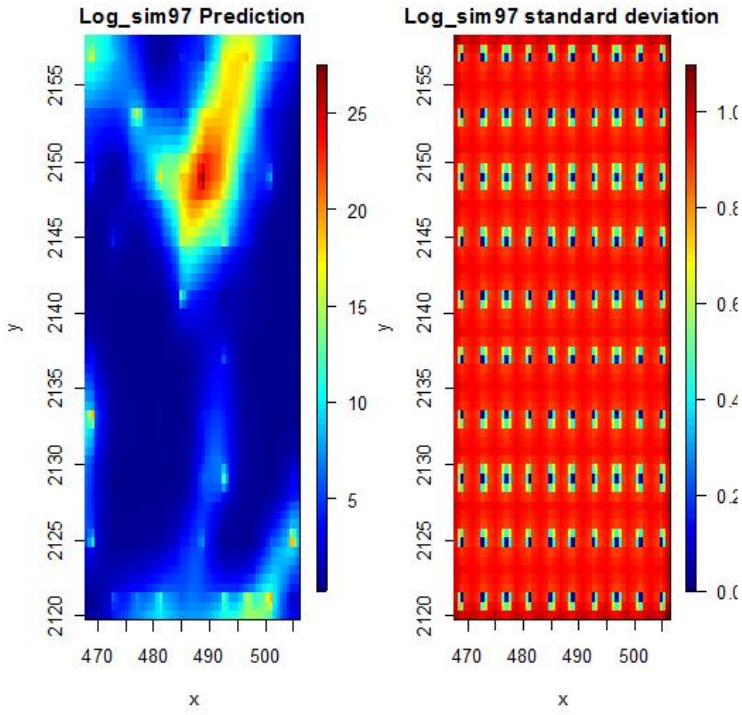

Figura 6.35: Mapa de estimaciones con kriging de logD3MRc100

6.5. Base de datos del tipo 3 con muestreo aleatorio y 100 observaciones (D3MAc100)

Se inicia el escenario representativo de D3MAc100 con la distribución de los datos (Figura 6.36)y sus estadísticas básicas (Cuadro 6.15).

En las estadísticas básicas se muestra que la media es mayor que la mediana, lo cual indica una asimetría positiva, que se confirma con el coeficiente de asimetría. Además, el 50 % de la información se encuentra en el intervalo [0.25, 3.9] y el restante 50 % se encuentra en el intervalo (3.9, 27.5].

Nombre	Estadísticas
Número total	100
Distancia max	50.44799302
Distancia min	1
Media	5.403
Varianza	40.34758408
Desviación estándar	6.351974818
Coeficiente var	1.175729125
Rango min	0.25
1er cuantil	0.4854
Mediana	3.897
3er cuantil	7.11
Máximo rango	27.5
Asimetría	1.452355463
Curtosis	4.462655909

Cuadro 6.15: Estadísticas básicas D3MAc100

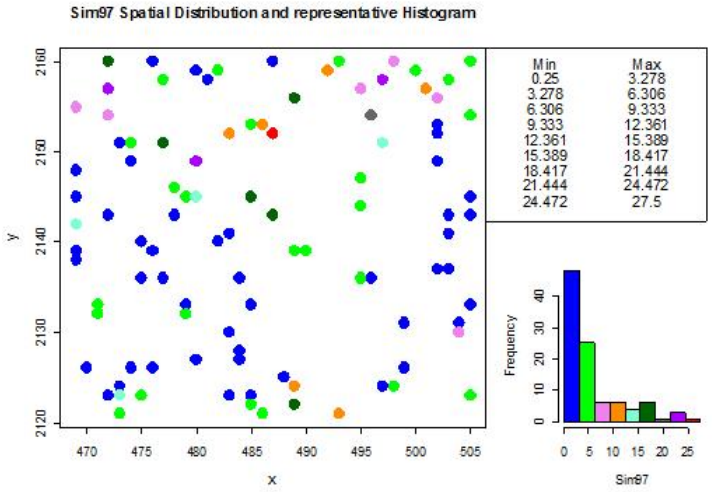

Figura 6.36: Distribución de D3MAc100

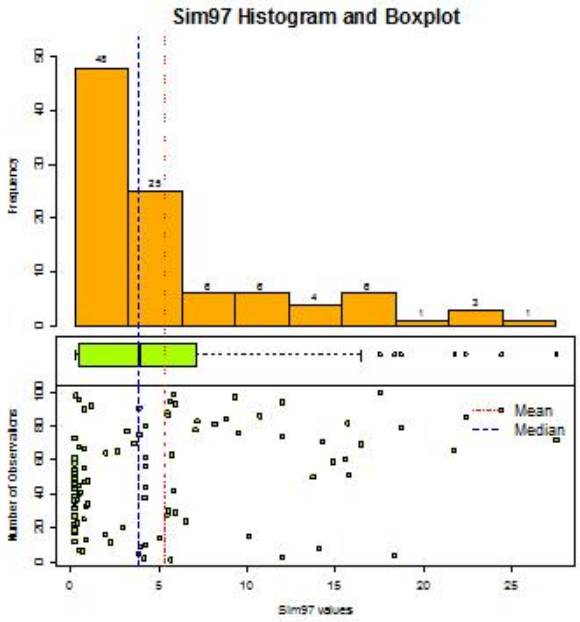

Figura 6.37: Histograma de D3MAc100

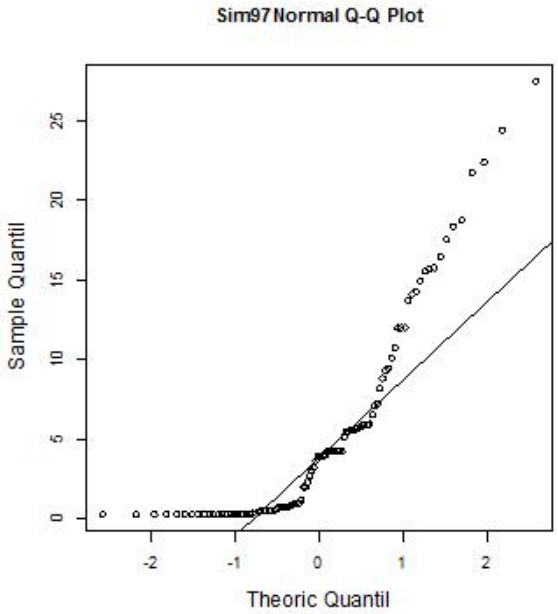

Figura 6.38: Q-Q plot de D3MAc100

El histograma (Figura 6.37) confirma la asimetría y muestra siete datos atípicos. El q-q plot (Figura 6.38) muestra que no se tiene distribución normal. Se propone una transformación logarítmica para reducir la escala y mejorar la asimetría. Se inicia nuevamente el análisis exploratorio de datos.

Se muestra la distribución de los datos (Figura 6.39) y se observa en las estadísticas básicas (Cuadro 6.16) que la media ahora es menor que la mediana, pero son más cercanas, lo cual indica una ligera asimetría negativa. Además, el 75 % de la información se encuentra en el intervalo [-1.38, 1.36] mientras que el restante 25 % se encuentra en el intervalo (1.36, 3.31].

Nombre	**Estadísticas**
Número total	100
Distancia max	50.44799302
Distancia min	1
Media	0.7544
Varianza	2.447299398
Desviación estándar	1.564384671
Coeficiente var	2.073733655
Rango min	-1.386
1er cuantil	-0.7239
Mediana	1.36
3er cuantil	1.961
Máximo rango	3.314
Asimetría	-0.139546019
Curtosis	1.552064161

Cuadro 6.16: Estadísticas básicas logD3MAc100

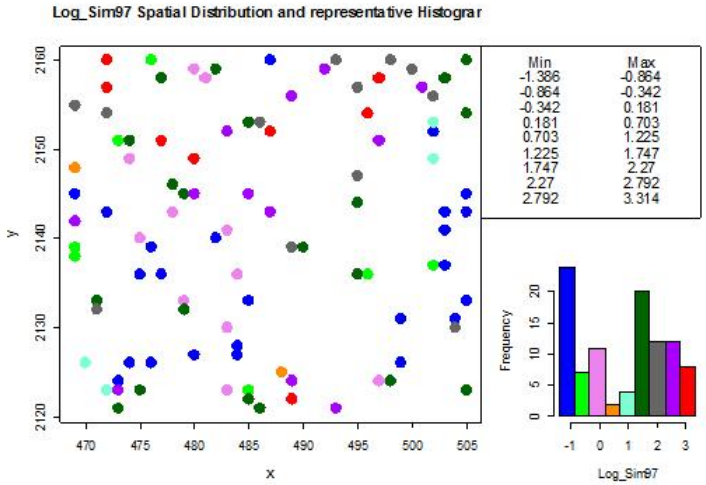

Figura 6.39: Distribución de logD3MAc100

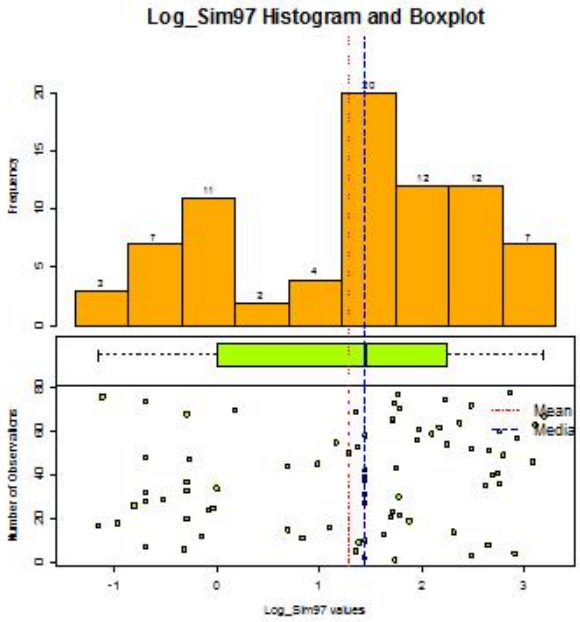

Figura 6.40: Histograma de logD3MAc100

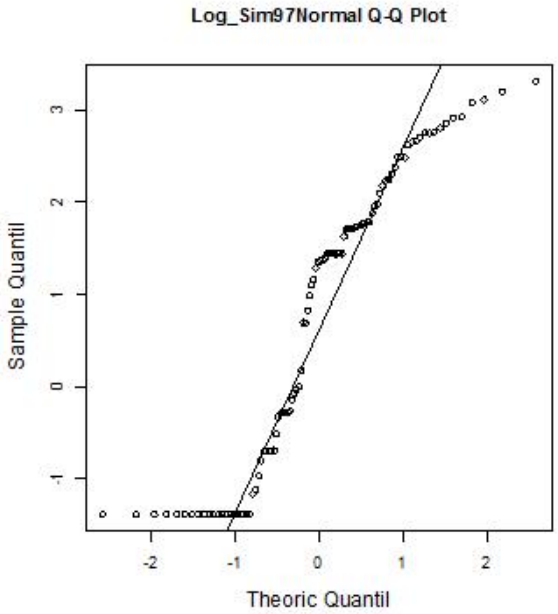

Figura 6.41: Q-Q plot de logD3MAc100

El histograma (Figura 6.40) y q-q plot (Figura 6.41) muestran una mejora en la asimetría de la muestra y ya no presentan datos atípicos.

Se realiza también el gráfico de estacionariedad (Figura 6.43) el cual no muestra indicios significativos de no se cumpla con estacionariedad y el gráfico con respecto a las coordenadas (Figura 6.42) no parece tener tendencia. Por lo que se continúa con el análisis variogáfico.

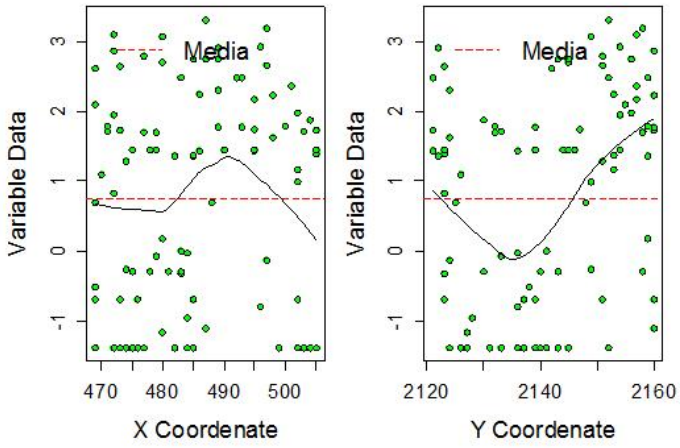

Figura 6.42: Gráfico respecto a las coordenadas de logD3MAc100

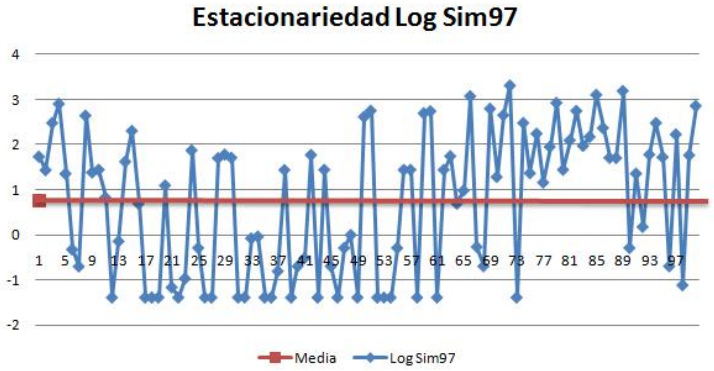

Figura 6.43: Estacionariedad de logD3MAc100

El variograma adireccional (Cuadro 6.17 Figura 6.44) no muestra comportamiento de h^2, lo que confirma que no hay indicios de tendencia. Sin embargo, el variograma en cuatro direcciones (Cuadro 6.18 Figura 6.45) muestra diferentes alcances para las diferentes direcciones lo cual es un indicio de anisotropía.

El mapa de anisotropía (Figura 6.46) confirma que existe anisotropía y se observan dos elipses con eje mayor tanto en la dirección cercana a 45° como en la dirección cercana a 135°. Esto puede ser ocasionado por la distribución de los datos por lo que es importante revisar las distintas direcciones para encontrar el variograma de mayor alcance que sea el más significativo.

Disctancia max	50.44799302
Distancia min	1
Dirección	0°
Tolerancia	90°
Intervalos	12
Distancia Lag	3

Cuadro 6.17: Variograma adireccional de logD3MAc100

Disctancia max	50.44799302
Distancia min	1
Dirección	0°,45°,90°,135°
Tolerancia	22.5°
Intervalos	12
Distancia Lag	3

Cuadro 6.18: Variograma 4 direcciones logD3MAc100

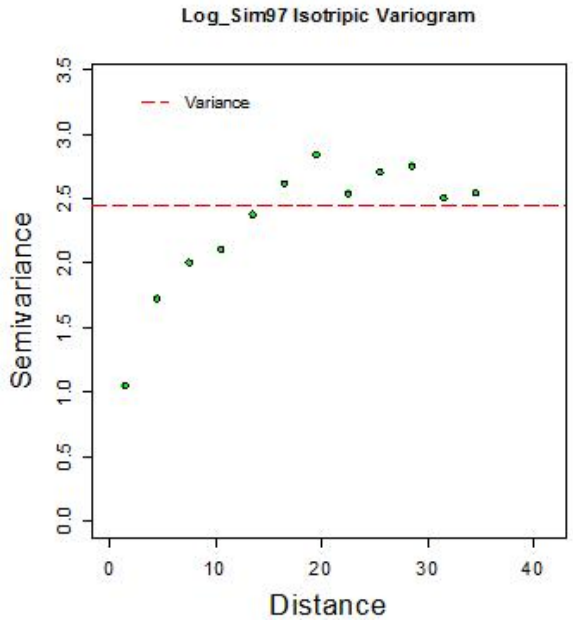

Figura 6.44: Variograma adireccional de logD3MAc100

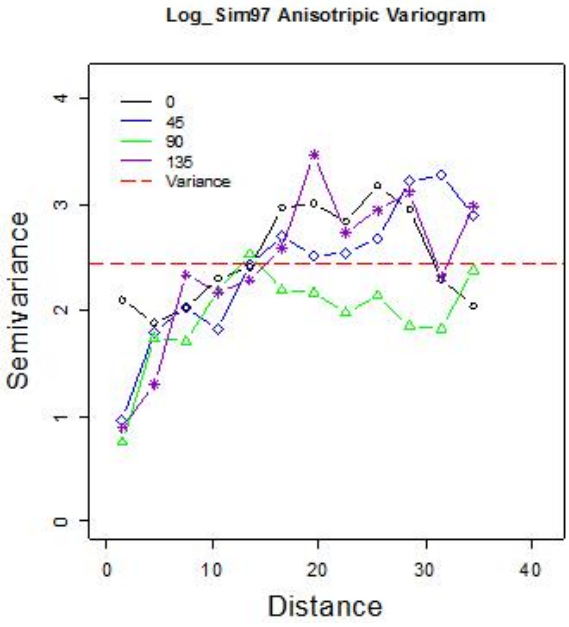

Figura 6.45: Variograma en 4 direcciones de logD3MAc100

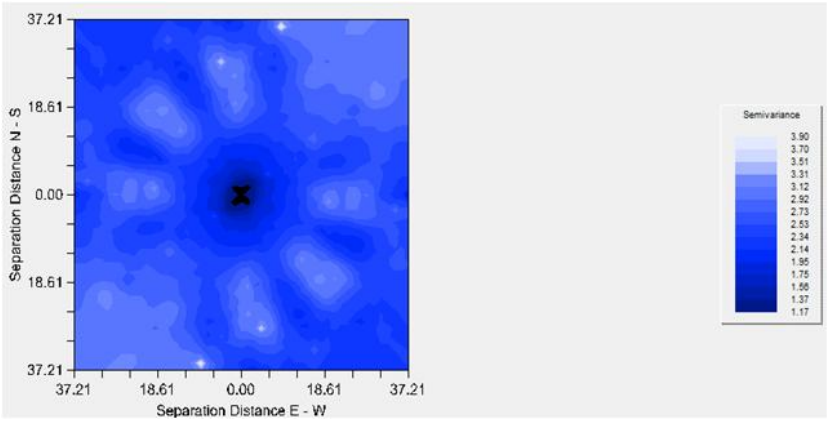

Figura 6.46: Mapa de anisotropía de logD3MAc100

Después de numerosas pruebas con variogramas en diferentes direcciones, se encuentra y se ajusta el variograma en la dirección del eje menor a 135° (Cuadro 6.19 Figura 6.47) y del eje mayor a 45° (Cuadro 6.20 Figura 6.48).

Modelo	**Nugget**	**Sill+Nugget**	**Rango**	**SCE**
Esférico	0.7	2.9	21	1.12

Cuadro 6.19: Variograma anisotrópico eje menor a 135° de logD3MAc100

Modelo	Nugget	Sill+Nugget	Rango	SCE
Esférico	1	2.9	30	0.752

Cuadro 6.20: Variograma anisotrópico eje Mayor a 45° de logD3MAc100

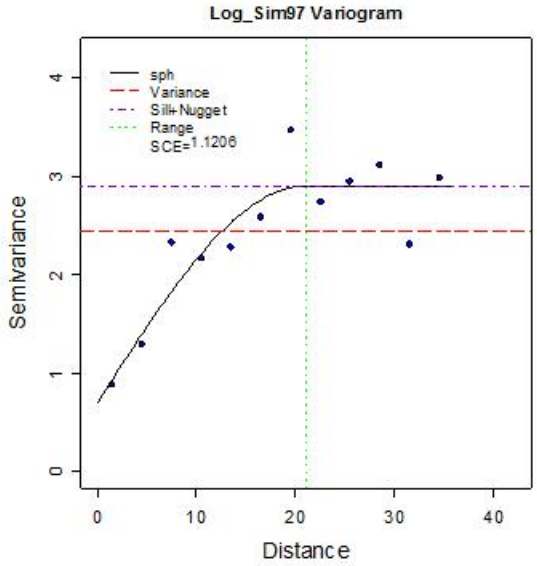

Figura 6.47: Variograma anisotrópico eje menor 135° de logD3MAc100

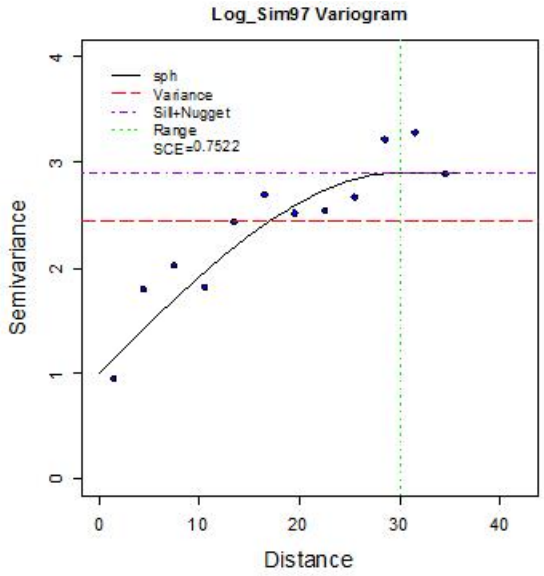

Figura 6.48: Variograma anisotrópico eje Mayor a 45° de logD3MAc100

Una vez elegidos los ajustes para los variogramas, se propone el variograma anisotrópico en dirección de 45° como modelo para la estimación por ser el de mayor alcance. Por lo que se verifica con la validación cruzada (Cuadro 6.21).

En la validación cruzada se observa que la media del error es cercana a cero. Además, la distribución de los estimados no es tan similar a la de los valores reales como se quisiera. También se realiza el análisis de residuales.

En el gráfico de valores reales contra estimados (Figura 6.49) se observa que los datos se aproximan a una línea de 45°

Nombre	**logD3MAc100**	**Estimados**	**Error**
Número total	100	100	100
Distancia max	50.44799302	50.44799302	50.44799302
Distancia min	1	1	1
Media	0.7544	0.7373	0.01712
Varianza	2.447299398	0.827289604	1.449256053
Desviación estándar	1.564384671	0.909554618	1.203850511
Coeficiente var	2.073733655	1.233687697	70.33551625
Rango min	-1.386	-1.027	-2.846
1er cuantil	-0.7239	0.1131	-0.8649
Mediana	1.36	0.8428	0.2263
3er cuantil	1.961	1.431	0.8044
Máximo rango	3.314	2.469	3.176
Asimetría	-0.139546019	-0.110760995	-0.08313444
Curtosis	1.552064161	2.061745177	2.777170247

Cuadro 6.21: Validación cruzada logD3MAc100

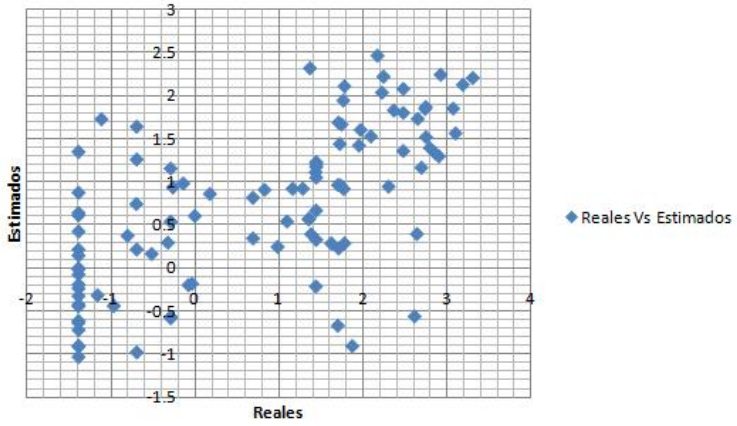

Figura 6.49: Valores reales contra estimados de logD3MAc100

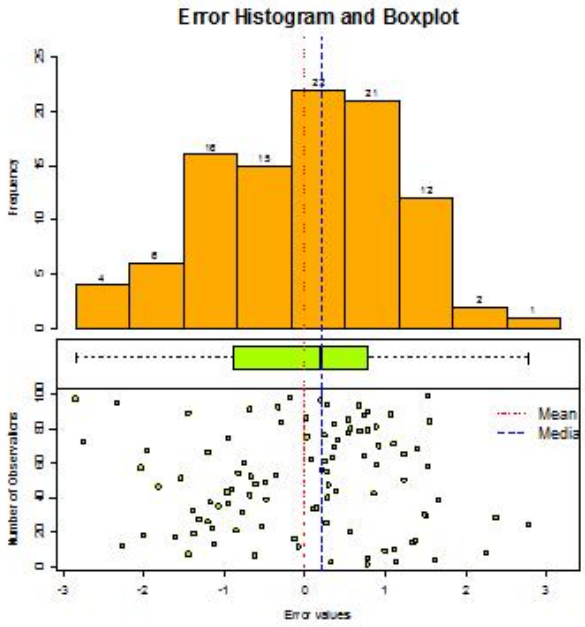

Figura 6.50: Histograma de errores de logD3MAc100

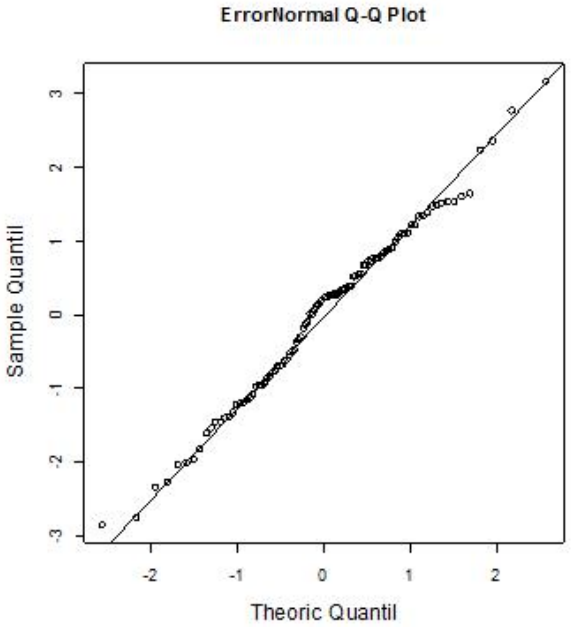

Figura 6.51: Q-Q plot de errores de logD3MAc100

El histograma (Figura 6.50) y q-q plot (Figura 6.51) muestran que los residuales se aproximan a una distribución simétrica y normal. Por lo tanto el modelo de variograma es aceptado y se procede a realizar la estimación con Kriging.

Se obtiene un mapa de valores estimados (Figura 6.69) a distancia de 1km por 1km, con un índice de anisotropía de 0.7.

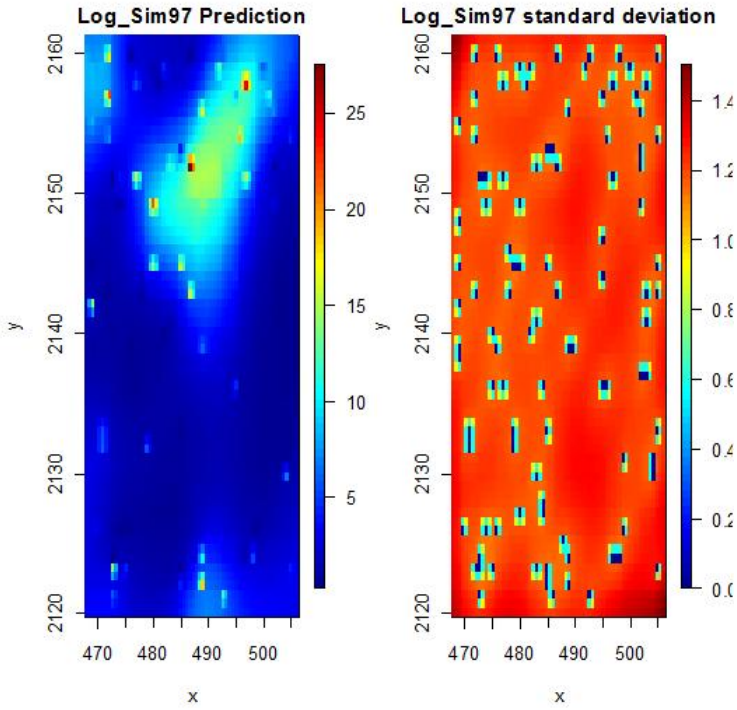

Figura 6.52: Mapa de estimaciones con kriging de logD3MAc100

6.6. Base de datos del tipo 3 con muestreo combinado y 100 observaciones (D3MCc100)

Se inicia el análisis exploratorio de datos para el escenario D3MCc100 mostrando la distribución de los datos (Figura 6.53) y sus estadísticas básicas (Cuadro 6.22).

En las estadísticas básicas se muestra que la media es mayor que la mediana, lo cual indica una asimetría positiva, que se confirma con el coeficiente de asimetría. Además, el 50 % de la información se encuentra en el intervalo [0.25, 3.5] mientras que el restante 50 % se encuentra en el intervalo (3.5, 25.98].

En el histograma (Figura 6.54) se observa visiblemente la asimetría de la muestra, así como los 16 datos atípicos. El q-q plot (Figura 6.55) destaca que la muestra no tiene distribución normal. Por lo tanto, se propone una transformación logarítmica para reducir la escala y mejorar la asimetría de la muestra. Se inicia nuevamente el análisis exploratorio de datos.

Nombre	**Estadísticas**
Número total	100
Distancia max	51.623638
Distancia min	1
Media	5.468
Varianza	44.34510819
Desviación estándar	6.65921228
Coeficiente var	1.217791264
Rango min	0.25
1er cuantil	0.2768
Mediana	3.535
3er cuantil	5.927
Máximo rango	25.98
Asimetría	1.366029478
Curtosis	3.758373242

Cuadro 6.22: Estadísticas básicas D3MCc100

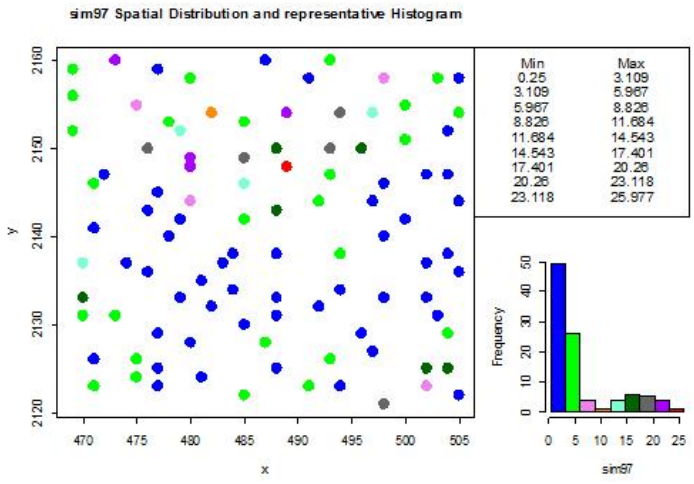

Figura 6.53: Distribución de D3MCc100

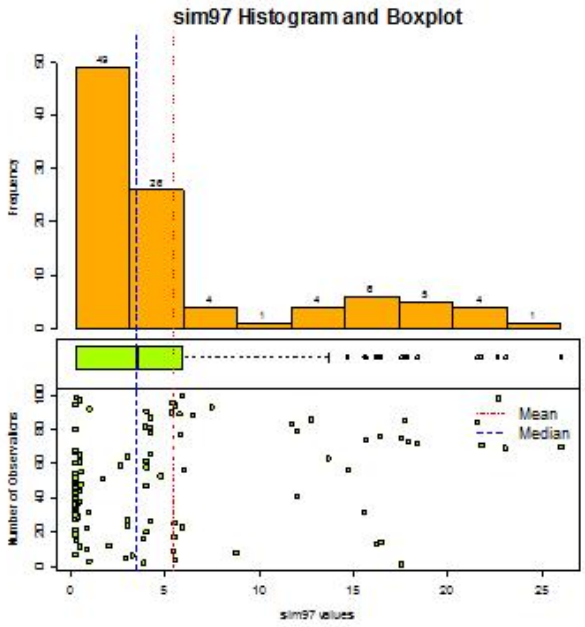

Figura 6.54: Histograma de D3MCc100

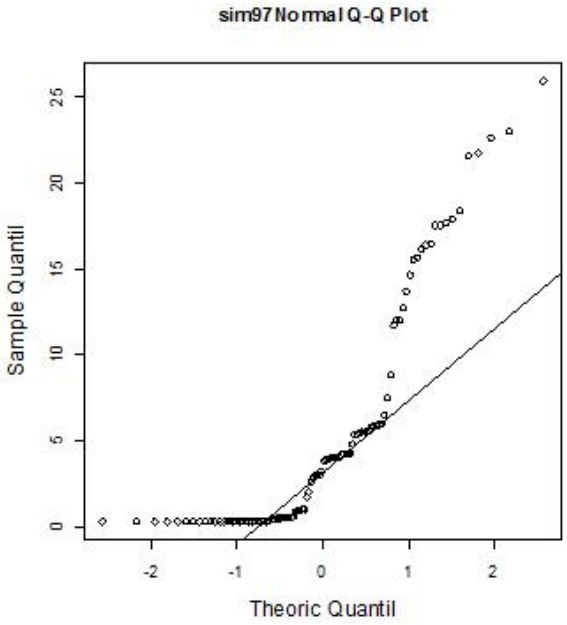

Figura 6.55: Q-Q plot de D3MCc100

Se muestra la distribución de los datos (Figura 6.56) ya transformados. Las estadísticas básicas (Figura 6.23) muestran que la media ahora es menor que la mediana y son ahora mucho más cercanas, lo que indica una ligera asimetría negativa que también se confirma con el coeficiente de asimetría. Además, se observa la reducción de escala ya que ahora el 75 % de la información se encuentra en el intervalo [-1.38, 1.25] mientras que el restante se encuentra en el intervalo (1.25, 3.25].

El histograma (Figura 6.57) muestra que ya no se tienen datos atípicos y una mejora significativa en la asimetría de la muestra, aunque el q-q plot (Figura 6.58) muestra que no se tiene una distribución normal.

Nombre	**Estadísticas**
Número total	100
Distancia max	51.623638
Distancia min	1
Media	0.6889
Varianza	2.634256782
Desviación estándar	1.623039366
Coeficiente var	2.355898168
Rango min	-1.386
1er cuantil	-1.286
Mediana	1.259
3er cuantil	1.78
Máximo rango	3.257
Asimetría	-0.074660298
Curtosis	1.501328788

Cuadro 6.23: Estadísticas básicas logD3MCc100

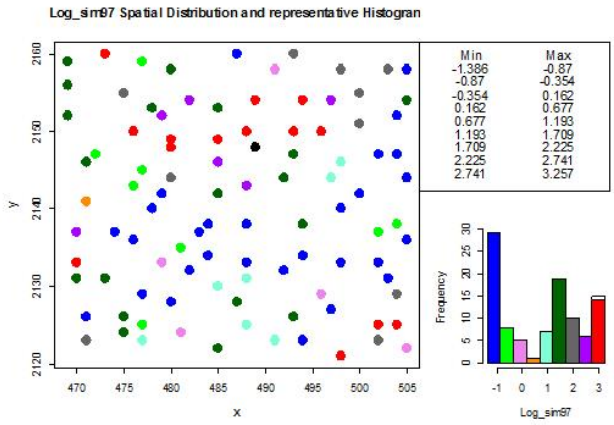

Figura 6.56: Distribución de logD3MCc100

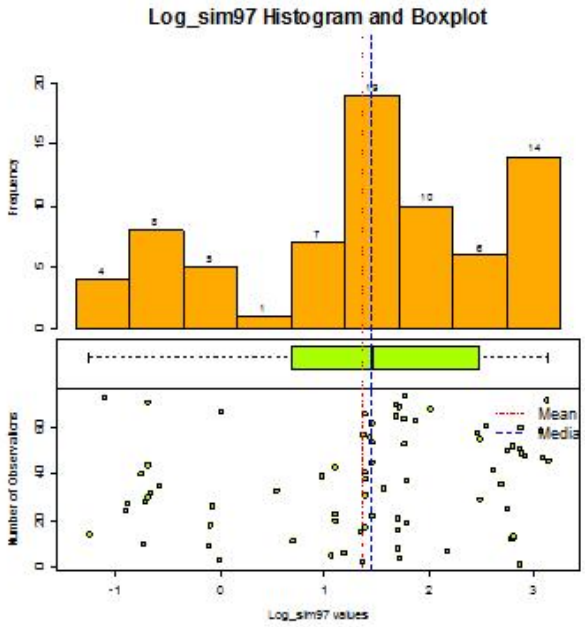

Figura 6.57: Histograma de logD3MCc100

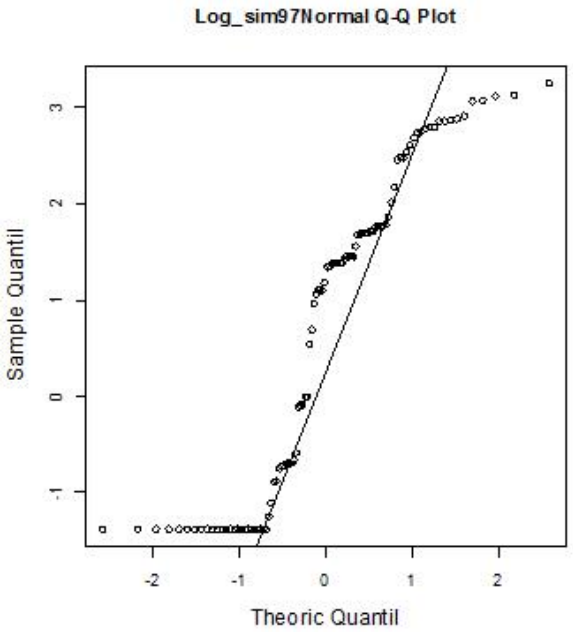

Figura 6.58: Q-Q plot de logD3MCc100

En el gráfico de estacionariedad (Figura 6.60) no se observan indicios de que no se cumpla con estacionariedad y el gráfico con respecto a las coordenadas (Figura 6.59) no muestra tendencia. Se procede a realizar el análisis variográfico.

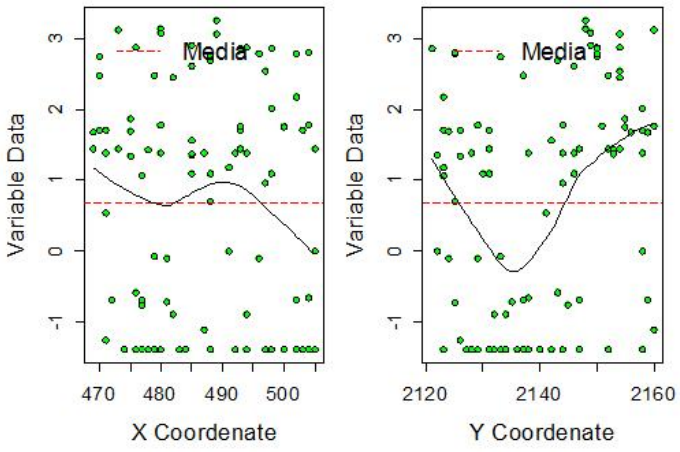

Figura 6.59: Gráfico con respecto a las coordenadas de logD3MCc100

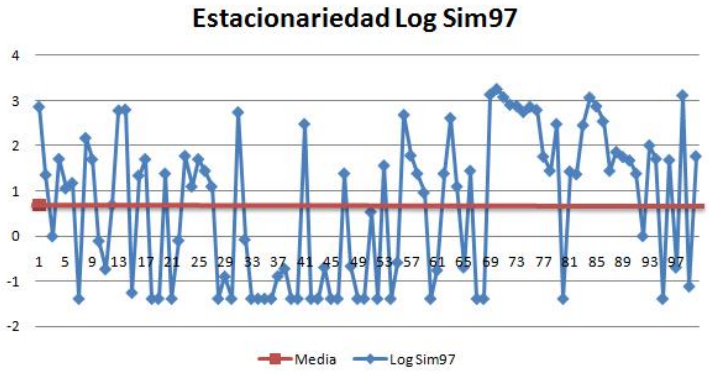

Figura 6.60: Estacionariedad de logD3MCc100

En el variograma adireccional (Cuadro 6.24 Figura 6.61) no se observa comportamiento de h^2, lo cual confirma que no hay tendencia significativa. Sin embargo, el variograma en cuatro direcciones (Cuadro 6.25 Figura 6.62) muestra diferentes alcances para las diferentes direcciones lo cual indica que se tiene anisotropía.

El mapa de anisotropía (Figura 6.63) muestra posibles elipses con eje mayor en dirección de 0°, ó en dirección cercana a 45°. Por lo tanto, se revisan cuidadosamente los diferentes variogramas para varias direcciones cercanas a lo observado.

Disctancia max	51.623638
Distancia min	1
Dirección	0°
Tolerancia	90°
Intervalos	11
Distancia Lag	2

Cuadro 6.24: Variograma adireccional de logD3MCc100

Disctancia max	51.623638
Distancia min	1
Dirección	0°,45°,90°,135°
Tolerancia	22.5°
Intervalos	11
Distancia Lag	2

Cuadro 6.25: Variograma 4 direcciones logD3MCc100

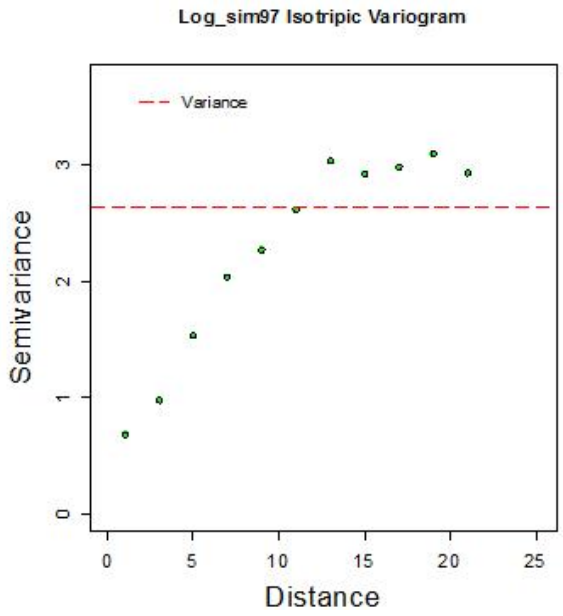

Figura 6.61: Variograma de logD3MCc100

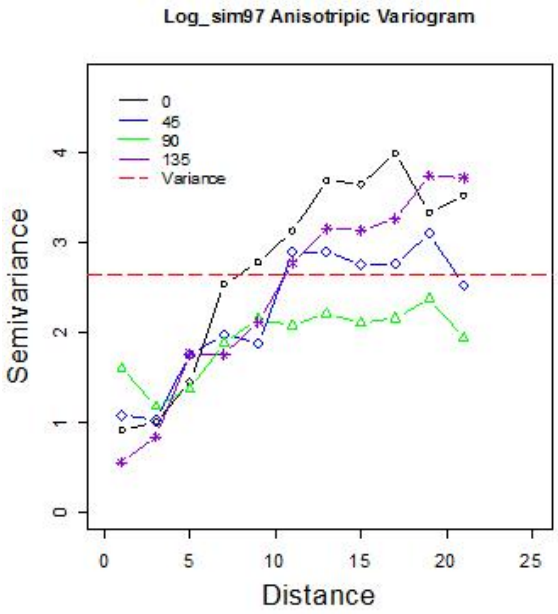

Figura 6.62: Variograma en 4 direcciones de logD3MCc100

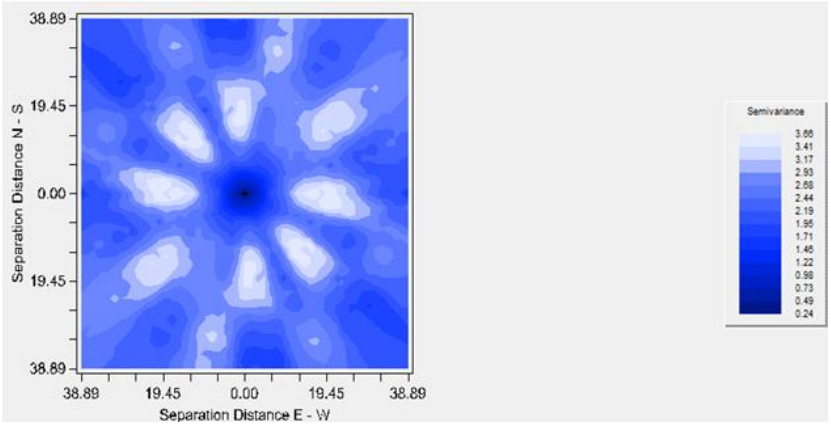

Figura 6.63: Mapa de anisotropía de logD3MCc100

En este escenario, las direcciones de anisotropía no coinciden con el mapa de anisotropía ya que pueden estar afectadas por los dos tipos de elipses presentados y la información con la que se cuenta. Al revisar con detenimiento cada variograma realizado se encontró que el eje de menor alcance se encuentra en la dirección de 30° (Cuadro 6.26 Figura 6.64) y el de mayor alcance a 120° (Cuadro 6.27 Figura 6.65). Una vez elegidas las direcciones de los ejes de anisotropía se muestran los ajustes para cada uno de ellos.

Modelo	**Nugget**	**Sill+Nugget**	**Rango**	**SCE**
Esférico	0.35	3.1	15.4	0.724

Cuadro 6.26: Variograma anisotrópico eje menor 30° logD3MCc100

Modelo	**Nugget**	**Sill+Nugget**	**Rango**	**SCE**
Esférico	0.4	3.17	22	0.729

Cuadro 6.27: Variograma anisotrópico eje Mayor 120° logD3MCc100

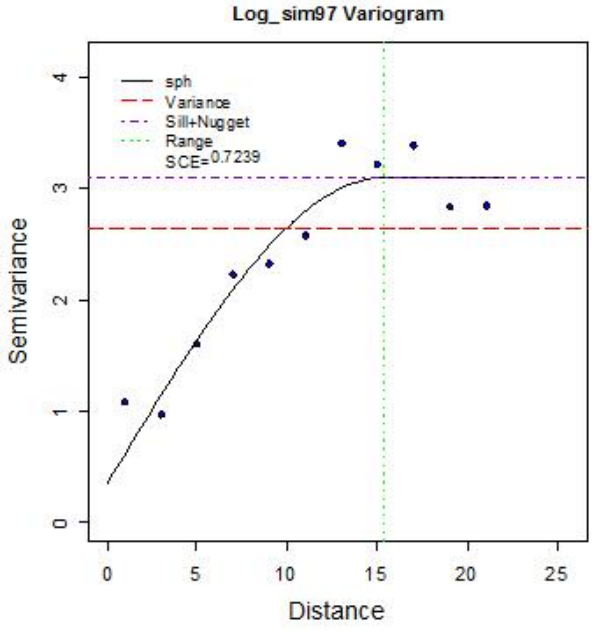

Figura 6.64: Variograma anisotrópico eje menor 30° logD3MCc100

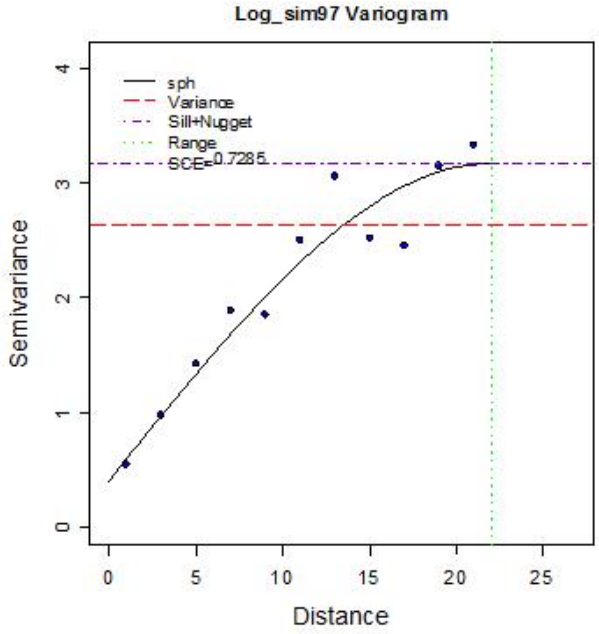

Figura 6.65: Variograma anisotrópico eje Mayor 120° logD3MCc100

Ya que se ajustan los variogramas de los ejes de anisotropía, se elige el de mayor alcance como modelo de variograma y se realiza la validación cruzada (Cuadro 6.28) para ver si es adecuado.

La validación cruzada muestra que la media de los errores es cercana a cero, y la distribución de los estimados no es tan similar a la de los datos. Se continúa con el análisis de residuales.

Nombre	**logD3MCc100**	**Estimados**	**Error**
Número total	100	100	100
Distancia max	51.623638	51.623638	51.623638
Distancia min	1	1	1
Media	0.6889	0.6813	0.007644
Varianza	2.634256782	1.444428804	1.514743022
Desviación estándar	1.623039366	1.201843919	1.230748968
Coeficiente var	2.355898168	1.764092988	160.9992008
Rango min	-1.386	-1.459	-2.988
1er cuantil	-1.286	-0.293	-0.7503
Mediana	1.259	0.6429	0.1188
3er cuantil	1.78	1.707	0.8592
Máximo rango	3.257	2.952	2.482
Asimetría	-0.074660298	0.090321915	-0.3785473
Curtosis	1.501328788	1.89822801	2.751886274

Cuadro 6.28: Validación cruzada logD3MCc100

El gráfico de valores reales contra estimados (Figura 6.66) muestra que los datos se aproximan a una línea de 45°. El histograma (Figura 6.67) y q-q plot (Figura 6.68) de los residuales muestran una distribución simétrica y cercana a la normal.

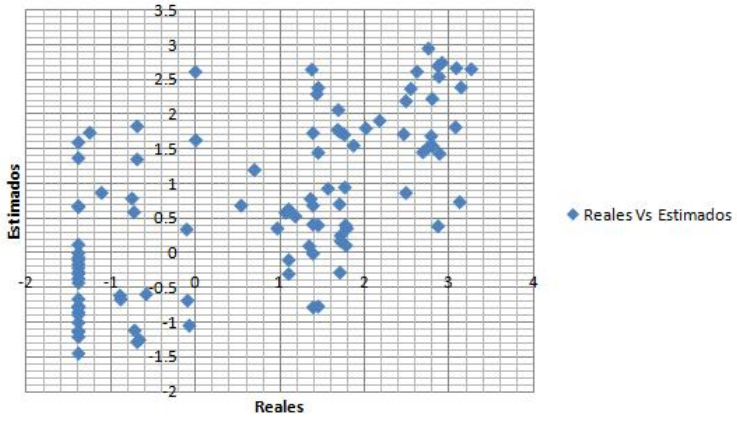

Figura 6.66: Valores reales contra estimados de logD3MCc100

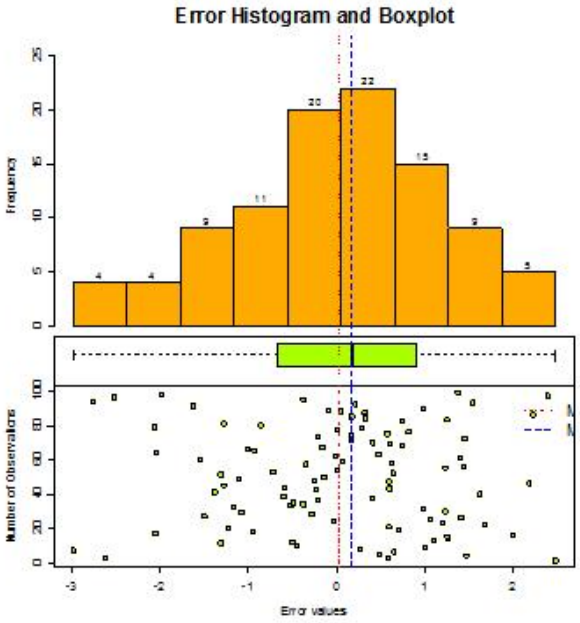

Figura 6.67: Histograma de errores de logD3MCc100

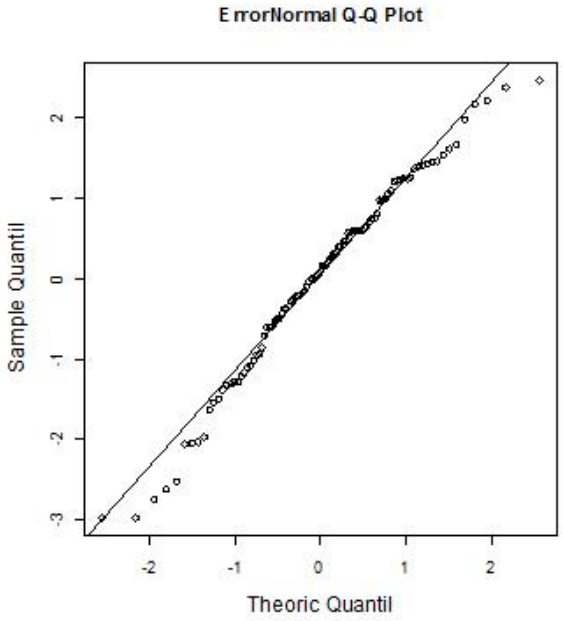

Figura 6.68: Q-Q plot de errores de logD3MCc100

Por lo tanto, se procede a realizar la estimación con Kriging. Se obtiene un mapa de valores estimados (Figura 6.69) a 1km por 1km, con un índice de anisotropía de 0,7.

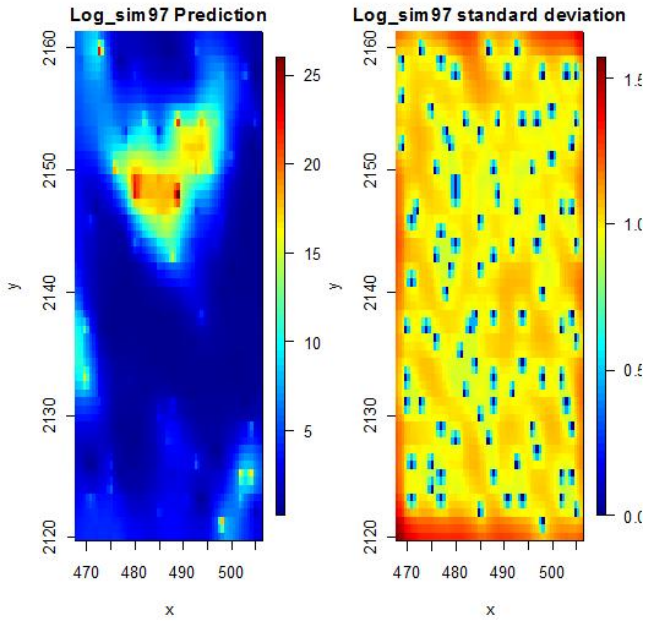

Figura 6.69: Mapa de estimaciones de logD3MCc100

6.7. Conclusiones de los modelos de datos 3

- Todos los modelos presentaron asimetría positiva y posteriormente a la transformación presentaron una ligera asimetría negativa.
- Todos los modelos presentaron anisotropía, por lo que en todos los escenarios se utilizaron modelos de variograma anisotrópicos.
- En todos los escenarios se utiliza una transformación de logaritmo para reducir escala y la asimetría que presentan los datos 3 y cumplir con los supuestos del modelo.
- No fue necesario eliminar los datos atípicos debido a que la transformación reduce la asimetría y permite que ya no existan datos atípicos.
- El lag de distancia para calcular los variogramas varía dependiendo del espaciamiento que existe entre muestras.
- Todos los modelos son ajustados con un modelo esférico y parámetros del variograma muy similares.
- Se realizó un ajuste visual con criterio propio para los ajustes de variograma tanto para el eje menor como para el eje mayor en cada uno de los escenarios.
- Se realizó validación cruzada en todos los escenarios y los modelos se aceptaron mediante el criterio de media de los errores cercana a cero, y se verificó que los residuales se asemejen una distribución normal o por lo menos simétrica.
- En todos los escenarios se estimó con Kriging ordinario.
- Cuando se tienen 400 observaciones, el muestreo utilizado tiene menos importancia debido a que la información cubre una gran parte de la región y los mapas de estimación son muy parecidos a los de los datos originales del tipo 3.
- Los escenarios con 36 observaciones son mucho más complejos para ajustarles un modelo adecuado y en ocasiones se logra únicamente con ajuste visual. El escenario D3MAc036 destaca que la característica de anisotropía es difícil de visualizar correctamente cuando las ubicaciones no son significativas.
- Los escenarios con 100 observaciones son los más representativos, ya que tienen una cantidad suficiente de información y no tienen tantos datos que la estimación requiera de mucho tiempo, aunque la información en ocasiones no es significativa y puede variar la estimación si no se encuentran adecuadamente los ejes de anisotropía.
- Los modelos son adecuados dependiendo de la precisión con la que se realice el ajuste del modelo, así como la información que se tenga a la mano.

Capítulo 7

Discusión y resultados de los casos de estudio

Dentro del siguiente capítulo se expresan los resultados obtenidos para cada caso de estudio, las diferencias más significativas y los problemas más frecuentes dentro de los escenarios presentados para cada tipo de datos. La relevancia radica en que cada caso de estudio presenta diferentes características las cuales pueden o no ser observadas dependiendo de la base de datos utilizada y la información con la que se cuenta, así como la manera en la cual fueron tratadas para lograr la mejor estimación posible.

7.1. Discusión y resultado del caso de estudio de Datos del tipo 1

El caso de estudio de los datos originales del tipo 1, pretendía ejemplificar el proceso de estimación cuando existía tendencia. Sin embargo, en este caso de estudio la tendencia es tan pequeña que no es suficiente para ser observada y mucho menos tratada en el proceso de estimación de los datos originales del tipo 1. Por lo cual, este caso de estudio resultó en una ejemplificación de la metodología la cual es equivalente a una muestra que no contiene características de gran influencia para la estimación. La muestra únicamente presenta asimetría visible, lo cual es fácilmente tratable con una transformación. Cuando no se presentan características de gran influencia es posible realizar una estimación de manera más acertada ya que no es afectada por la dirección y no presenta una tendencia tal que pudiera sesgar el modelo. Dentro de cada uno de los escenarios, se encontró una gran relación entre los modelos utilizados para la estimación por lo cual los mapas de estimación fueron muy parecidos a los datos originales del tipo 1. Conforme aumentaba el número de observaciones, la cercanía con el mapa de los datos originales fue aumentando.

El proceso de estimación de los datos originales del tipo 1 se resume de la siguiente manera. Se inicia con el análisis exploratorio de datos, en el cual se observó que la muestra no tiene distribución normal, y además tiene una

asimetría positiva y datos atípicos, por lo que se le aplicó una transformación de raíz cuadrada para reducir la asimetría y la escala. Utilizando el gráfico de estacionariedad y el gráfico con respecto a las coordenadas se verificó que no existieran indicios de tendencia que pudieran afectar la estimación. Posteriormente, se procede a realizar el análisis estructural, el cual se basa en el análisis variográfico donde se corrobora que no exista tendencia o comportamiento de h^2 en el variograma y se buscan indicios de anisotropía generando también el mapa de anisotropía para corroborar que no existe esta característica. Después, se ajustó un modelo de variograma esférico con nugget de 0.2, meseta de 2.5 y alcance de 26. Luego se realiza la validación cruzada y las pruebas de residuales para verificar que el modelo sea adecuado. Por último, se realizó la estimación por el método de Kriging ordinario para obtener un mapa de estimaciones el cual resultó como se esperaba, muy cercano al gráfico de la distribución de los datos originales del tipo 1.

Recordando que cada uno de los escenarios de este caso de estudio fue realizado de manera independiente y anterior al resultado de los datos originales del tipo 1, se mencionan los siguientes puntos relevantes en común que se encontraron para todos los escenarios:

1. La base de datos de cada escenario, debido a que es un subconjunto de los datos originales del tipo 1, también presenta asimetría positiva.
2. En todos los escenarios de datos 1, la muestra es tratada con una transformación de raíz cuadrada, tanto para disminuir la asimetría como para reducir la escala.
3. La base de datos de cada escenario no cumplen normalidad.
4. La base de datos de cada escenario no presentan tendencia.
5. La base de datos de cada escenario no presentan indicios de no estacionariedad.
6. La base de datos de cada escenario no presentan indicadores de anisotropía.
7. Los modelos de variograma de cada escenario pasaron la validación cruzada y el análisis de residuales.
8. En todos los escenarios se realizó la estimación por medio de Kriging ordinario.
9. Los modelos ajustados en casi todos los escenarios son esféricos, excepto en el D1MAc400.
10. El nugget se encuentra en un intervalo de $[0, 0.28]$ para todos los escenarios.
11. La meseta se encuentra en un intervalo de $[2.34, 3.63]$ para todos los escenarios.
12. El alcance varía en un intervalo de $[21.35, 35]$ para todos los escenarios.

Más específicamente, el resultado de los escenarios con 36 observaciones resultó como se esperaba, un ajuste del modelo de variograma muy complejo para los tres tipos de muestreo debido a que el variograma no contenía información suficiente por lo que fue necesario un ajuste casi completamente visual, sin embargo, se esperaba que el muestreo regular fuera el más eficiente, pero en este caso destaca el muestreo combinado ya que mostró una ligera diferencia sobre el variograma porque su ajuste fue el menos elaborado dentro de este grupo de escenarios que contenía tan poca información y el mapa de estimación fue el más cercano al mapa de estimación de los datos originales del tipo 1. Por otro lado, el grupo de escenarios que contiene 100 observaciones resultó más próximo al mapa de estimación de los datos originales del tipo 1. Dentro de este grupo, el muestreo aleatorio (escenario D1MAc100) es el que mejor coincide en el mapa de estimación, a pesar de que el muestreo no es el óptimo, las observaciones resultaron muy significativas por lo que fue posible estimar más adecuadamente. Sólo un escenario dentro del caso de estudio de Datos 1 resultó ajustado como modelo exponencial y fue el escenario D1MAc400, esto puede ser ocasionado por el muestreo, ya que no es regular y probablemente asignó mayor peso a muestras que no eran tan significativas, lo cual derivó en un modelo diferente al utilizado en los demás escenarios. Por último, destaca la notoria similitud que existe respecto al mapa de estimación de los datos originales del tipo 1 para el grupo de escenarios con 400 observaciones. Cuando se cuenta con una gran cantidad de información, es posible generar con mayor confiabilidad una estimación. Cuando se tiene un aumento considerable en la información, el muestreo resulta menos importante debido a que existen menos regiones dentro del área total en las cuales no se cuenta con información y también se cuenta con suficiente información significativa. Justo como se esperaba, el variograma es muy parecido entre lo variogramas de este grupo y estos son muy parecidos al variograma de los datos originales del tipo 1, además de que su ajuste no es tan complejo, con lo cual los modelos de variograma elegidos son muy similares entre sí y por lo tanto los mapas de estimación también.

Para todos los escenarios dentro del caso de estudio de datos del tipo 1, aunque no se conocía por adelantado la característica que presentaba la base de datos, los gráficos más importantes para tendencia son el gráfico de estacionariedad ya que muestra si la media no es constante, con lo cual es un posible indicador de tendencia, así como el gráfico con respecto a las coordenadas ya que también muestra el comportamiento de la información. Además, dentro del proceso de estimación destacan también gráficos como el histograma, debido a que muestra los datos atípicos, la asimetría o normalidad de la muestra y da una idea de la distribución de los datos. Es muy importante el gráfico del variograma adireccional, ya que permite identificar o confirmar si existe tendencia, así como el variograma en cuatro direcciones porque permite identificar si existe anisotropía.

Dentro de los procedimientos utilizados, se menciona que era posible utilizar una transformación diferente a la utilizada, eso es criterio del geoestadístico[1], sin embargo se recomienda que si se utiliza una transformación sobre la información sea una que se apega a las características de la muestra, en este caso de

[1] Como hace referencia en el artículo *"Variance of Geostatisticians"* Ref. Bibliográfica [22]

estudio, la diferentes bases de datos de los escenarios presentados y de los datos originales del tipo 1 presentan valores de cero, asimetría positiva y la escala es grande por lo cual se decidió utilizar una transformación de raíz cuadrada la cual mejora estas características, incluyendo el hecho de que eliminó los datos atípicos en casi todos los escenarios. También se recomienda que para el modelo elegido en el ajuste del variograma, se utilicen y den gran importancia a las técnicas de validación cruzada y el análisis de residuales ya que con ellas es posible medir intuitivamente la confiabilidad del modelo.

7.2. Discusión y resultado del caso de estudio de Datos del tipo 2

Para el caso de los datos del tipo 2, es muy importante mencionar que existió una diferencia de gran trascendencia dentro de los escenarios de este caso de estudio, debido a que los datos tienen una característica de estructura anidada y esta no fue observada ni tratada en los procesos de estimación.

Por el contrario, los datos muestran un comportamiento no estacionario y lineal ascendente el cual es posible asociarlo a que la estructura tenía una parte de gaussiano y otra parte de exponencial lo cual ocasionó una elevación de manera exponencial sobre la información que se podría interpretar como tendencia. Por lo tanto, el proceso de estimación para los datos del tipo 2 contiene la ejemplificación del proceso de ajuste de tendencia. Debido a que desde la base de Datos originales del tipo 2 es observada la tendencia, se concluye que no necesariamente es error realizar este procedimiento cuando se cuenta con indicadores que lo respaldan.

Recordando, la tendencia es un patrón de comportamiento de los elementos de un entorno particular durante un período, es decir, en este caso es la dirección que toman los valores respecto a cada una de las coordenadas y la importancia radica en que si existe tendencia entonces no se cumple con la hipótesis intrínseca, por lo que no se tiene media constante y entonces no cumple con los supuestos de la metodología mencionada.

Dentro del proceso de estimación, se observa que este tipo de datos no pueden ser utilizados directamente ya que no cumplen uno de los supuestos de la metodología de estimación con kriging, el cual requiere de estacionariedad. Por lo tanto, es necesario realizar una transformación o algún cambio que permita continuar el proceso de estimación. En general, dentro de la metodología existe un modelo de tendencia el cual se utiliza en estos casos, se utilizan los residuales del nuevo modelo los cuales deben ser estacionarios y con ellos se realiza la estimación. Sin embargo, en cada uno de los escenarios se analiza y observa el comportamiento de la información y se define si la muestra presenta características de tendencia o no, así como las implicaciones que conllevan las diferentes afectaciones que ocurren cuando no se tiene información suficiente o no se cuenta con un muestreo adecuado.

El proceso de estimación de los datos originales del tipo 2 se resume de la siguiente manera. Primero se realiza el análisis exploratorio de datos del tipo 2, donde se encuentra que se tiene una asimetría significativa, así como datos atípicos; entonces se procede a realizar una transformación logarítmica para mejorar la asimetría y reducir la escala. Luego se realiza nuevamente el análisis exploratorio. En el momento que se llega al gráfico de tendencia (o gráfico respecto de las coordenadas) se observan indicios de tendencia con respecto a la variable y, y se procede a realizar el análisis variográfico para verificarla. El variograma adireccional muestra un comportamiento lineal, por lo que se procede a tratar la tendencia con otra transformación. Sin embargo, si se realiza la transformación al modelo de tendencia en este punto del proceso se tendrían dos transformaciones, es más adecuado probar el modelo de tendencia sobre los datos originales del tipo 2 y verificar nuevamente con el análisis exploratorio de datos si disminuyó la escala, asimetría y datos atípicos. Una vez verificado que el nuevo modelo cumple los supuestos, se realiza nuevamente el análisis variográfico y se confirma que la tendencia ya no se presenta. También se revisa que no haya indicios de anisotropía. Ahora, se propone el nuevo modelo de variograma calculado con 25 intervalos y un lag de 1 y se realiza un ajuste visual que conlleva a un modelo esférico con nugget de 9.8, meseta de 26.4 y alcance de 13.2. Luego, se verifica si es adecuado con la validación cruzada y el análisis de residuales. Y por último, se procede a realizar la estimación con Kriging ordinario.

Resaltando nuevamente que cada uno de los escenarios de este caso de estudio fue realizado de manera independiente y anterior al resultado de los datos originales del tipo 2, se mencionan los siguientes puntos relevantes en común que se encontraron para todos los escenarios:

1. Las bases de datos de cada escenario presentan asimetría positiva.
2. En todos los escenarios de datos 2, la muestra es tratada con una transformación de logaritmo en principio y posteriormente al encontrarse tendencia se regresa a los datos originales del tipo 2 y con ellos se utiliza el modelo de residuos para tratar la tendencia. El caso excluyente es el de D2MAc036 el cual únicamente utilizó la transformación de logaritmo y no el modelo de tendencia, debido a que no pasó las pruebas de validación del modelo.
3. Las bases de datos de cada escenario no cumplen normalidad.
4. Los escenarios presentan indicios de no estacionariedad por lo que son tratados con el modelo de tendencia.
5. Los escenarios no presentan indicadores de anisotropía.
6. Los modelos de los escenarios pasaron la validación cruzada y el análisis de residuales, excepto el escenario D2MAc036 el cual no pasó las pruebas para verificar que el modelo fuera adecuado, por lo que se ajustó un modelo sin eliminar la tendencia.
7. En todos los modelos se realizó la estimación por medio de Kriging ordinario.

8. Los modelos ajustados en casi todos los escenarios son esféricos, excepto en el caso D2MAc036.

9. El nugget está en un intervalo de [0, 13.5] para todos los escenarios.

10. La meseta se encuentra en un intervalo de [11.5, 32] en los escenarios con esféricos y 4.5 para el escenario con exponencial.

11. El alcance varía en un intervalo de [13.2, 15.5] para los escenarios esféricos y 24 para el escenario exponencial.

Para destacar las diferencias entre los escenarios presentados, se agrupa la información en común dependiendo del número de observaciones que contiene la base de datos de cada escenario. Se comienza con el grupo de escenarios de 36 observaciones, donde el escenario D2MRc036 fue bastante complejo al ajustarle un modelo y poco confiable como lo es en los escenarios de poca información. El escenario D2MCc036 tiene el mapa de estimación más alejado respecto al mapa de estimación de los datos originales del tipo 2 aunque no fue el más complejo en el ajuste del modelo. El escenario que destaca es el D2MAc036 debido a que al ajustarle un modelo de variograma con la transformación de tendencia no se encuentra ningún modelo que pase las pruebas de validación cruzada y residuales, por lo cual es necesario regresar el modelo únicamente con la transformación logarítmica y a ésta ajustarle el modelo de variograma para la estimación. Dentro de este grupo, resalta que las bases de datos tan pequeñas son complejas para permitir calcular un variograma adecuadamente y por lo mismo también es difícil ajustar el modelo de variograma, por lo cual la combinación de una estructura anidada que se percibe como tendencia, más un muestreo aleatorio con una cantidad insuficiente de información para definir todas las características resulta en un proceso laborioso y difícilmente confiable en los resultados, aunque sorprendentemente acertado para esta situación en particular. Se continúa con el grupo de escenarios con 100 observaciones, en los cuales es más notoria la tendencia sobre los diversos gráficos y el ajuste del variograma no es tan complejo como en los escenarios de 36 observaciones. El escenario D2MRc100 es el escenario que tiene un mapa de estimación más cercano al de los datos originales del tipo 2. El escenario D2MCc100 tiene un mapa de estimación cercano al de los datos originales del tipo 2 pero sólo en las regiones donde los valores son más significativos, ya que en regiones donde los valores no son tan definidos pierde una adecuada estimación. El escenario D2MAc100 es el más complejo dentro de este grupo ya que como se esperaba, la ubicación de las observaciones varía de tal manera que no permite tener un mapa de estimación suficientemente adecuado y confiable. Para los escenarios con 400 observaciones la característica de tendencia se muestra visiblemente, lo cual permite identificarla más fácilmente. De igual manera que en el caso de estudio anterior, al aumentar el número de observaciones hace que las diferencias entre los tipos de muestreo de este grupo de escenarios sean menores. Los variogramas entre los tres tipos de muestreo con 400 observaciones se vuelven muy similares, por lo que los ajustes en los modelos también lo son y por ende los mapas de estimación también.

Para todos los escenarios dentro del caso de estudio de datos del tipo 2, resultan de gran importancia los gráficos para tendencia, como el gráfico de estacionariedad, el gráfico respecto a las coordenadas y con importancia máxima

el variograma. Esto debido a que los mencionados gráficos son los que indican una falta sobre los supuestos de la metodología y debido a ellos la estructura anidada es tratada como tendencia. Es importante destacar que al realizar la transformación de tendencia, también cambia el variograma y por lo tanto la estructura anidada ya no es observada, con lo cual es posible ajustarle un modelo diferente al que se contenía como característica en la base de datos del tipo 2. En adición, gráficos como el histograma, resultan muy importantes ya que la tendencia mostrada podría ser causada por una asimetría grave o por muchos valores extremos.

Dentro de los procedimientos utilizados, se menciona la transformación inicial de logaritmo la cual podría ser diferente, ya que esa transformación es a criterio del geoestadístico[2]. De igual manera, la transformación del modelo de tendencia es elegida por el geoestadístico ya que debe de probar con polinomios de diferentes grados para llegar al que se ajuste mejor al patrón de tendencia que presenta y confirmar que los residuos son estacionarios para aceptarlo; así como verificar que no afecte otros supuestos de la metodología dentro del análisis exploratorio, como asimetría significativa o datos atípicos de gran influencia. Por último, el modelo elegido de variograma también puede variar dependiendo de la decisión del geoestadístico, por lo que tendrá que evaluar en base a las mismas pruebas de validación cruzada y de ahí definir qué tan confiable es la estimación.

7.3. Discusión y resultado del caso de estudio de Datos del tipo 3

El caso de estudio de los datos originales del tipo 3, como se esperaba ejemplificó la identificación y el proceso de estimación cuando existe anisotropía. La muestra de igual manera que en los casos de estudio anteriores presenta asimetría visible, lo cual fue fácilmente tratable con una transformación logarítmica. Cuando se identifican características de gran influencia es posible realizar una estimación de manera más acertada siempre y cuando sea detectada y tratada de manera adecuada. Dentro de cada uno de los escenarios, se encontró una gran relación entre los modelos utilizados para la estimación, sin embargo, debido a que la característica de anisotropía depende de la dirección, los mapas de estimación fueron significativamente influenciados por el tipo de muestreo y el número de observaciones que contenían.

El proceso de estimación de los datos originales del tipo 3 se resume de la siguiente manera. Primero se realiza un análisis exploratorio de los datos del tipo 3, donde se observa asimetría positiva y varios datos atípicos. Entonces se aplica una transformación logarítmica para mejorar la simetría de la muestra y reducir la escala. Luego se verifica que no exista tendencia significativa y que cumpla con estacionariedad. Después se procede a realizar el análisis variográfico donde se genera el variograma adireccional el cual no muestra indicios de tendencia y después se generan los variogramas en cuatro direcciones en los cuales se observa anisotropía. Una vez detectada la anisotropía es importante realizar

[2]Como se menciona en el artículo *"Variance of geostatisticians"* Ref. Bibliográfica [22]

numerosos variogramas en diversas direcciones para encontrar la dirección de máxima continuidad y con ellos los ejes de anisotropía. Dentro de los diversos variogramas se eligen los de mayor y menor alcance, tomando en cuenta que deben estar en direcciones perpendiculares, es decir, el eje mayor se encuentra a $\pm 90°$ del eje menor y viceversa. Una vez elegidos los ejes de anisotropía, se realizan los ajustes de los variogramas correspondientes para conocer el alcance definido. Cabe destacar que los modelos deben tener un nugget y meseta muy cercanos y sólo variar en el alcance. Ahora se procede con un modelo de variograma anisotrópico el cual es en la dirección de mayor alcance. Para este caso referente a los datos originales del tipo 3 se utilizó un variograma en la dirección de $20°$, con un modelo esférico con nugget de 0.6, meseta de 3.3 y alcance de 24.5. Posteriormente se realiza la validación cruzada y se verifica que el modelo propuesto cumpla con el análisis de residuales. Por último, se realiza la estimación con kriging ordinario y si es necesario se utiliza el índice de anisotropía que es un valor entre [0, 1], ya que por definición es el alcance del eje menor entre el alcance del eje mayor. Con lo que finalmente se obtiene un mapa de valores estimados con elipses en la dirección del eje mayor de anisotropía.

Reiterando que cada uno de los escenarios de este caso de estudio fue realizado de manera independiente y anterior al resultado de los datos originales del tipo 3, se mencionan los siguientes puntos relevantes en común que se encontraron para todos los escenarios:

1. En todos los escenarios de datos 3, la muestra es tratada con una transformación de logaritmo para reducir la escala y asimetría.
2. La distribución de las muestras tiene asimetría positiva y posterior a la transformación logarítmica se tiene una ligera asimetría negativa, la cual ya no es significativa.
3. Las bases de datos de los escenarios del tipo 3 presentaron anisotropía.
4. Las bases de datos de los escenarios del tipo 3 no cumplen normalidad.
5. Todos los modelos de los escenarios del tipo 3 pasaron la validación cruzada y análisis de residuales.
6. En todos los modelos de los escenarios del tipo 3 se realizó la estimación por medio de Kriging ordinario.
7. Los modelos ajustados en todos los escenarios del tipo 3 son esféricos.
8. El nugget de los modelos ajustados se encuentra dentro del intervalo [0.3, 1] para todos los escenarios.
9. La meseta de los modelos ajustados se encuentra en un intervalo de [2.9, 3.4] para todos los escenarios.
10. El alcance de los modelos ajustados varía en un intervalo de [13.4, 30] para todos los escenarios.

Continuando con el resultado de los escenarios del caso de estudio del tipo 3, se agrupa la información obtenida para los escenarios con 36 observaciones. Primero se confirma un ajuste visual y muy complejo del modelo de variograma para los tres tipos de muestreo. Luego, se ve altamente afectado el mapa de estimaciones en los tres tipos de muestreo debido a que la información de las observaciones no permite definir confiable y acertadamente la dirección de anisotropía. En particular, el escenario del muestreo combinado (D3MCc036) destaca debido a que fue necesario ajustar la tolerancia a 30° para calcular el variograma y poder identificar los ejes de anisotropía, sin embargo, debido a la poca información con la que se cuenta el mapa de estimación no es suficientemente acertado. El escenario D3MAc036 destaca debido a que la ubicación de las muestras no permite identificar acertadamente la dirección de anisotropía y por lo tanto la estimación no es confiable ni adecuada. Por último, el escenario D3MRc036 es el más acertado dentro de los tres tipos de muestreo con 36 observaciones ya que el mapa de estimación es el más cercano dentro de este grupo al mapa de estimación de los datos originales del tipo 3. Es adecuado pensar que el número tan reducido de información con la que se cuenta y claramente siendo poco significativa ocasione que la dirección de máximo y mínimo alcance se vea altamente desviada. Ahora sigue el grupo de escenarios de 100 observaciones, en el cual también destaca el escenario de D3MRc100 debido a que permite encontrar los ejes de anisotropía adecuadamente, ajustar un modelo adecuado y ser el escenario del grupo que tiene el mapa de estimación más cercano al de los datos originales del tipo 3. Después sigue el escenario D3MAc100 el cual destaca que se pueden encontrar los ejes de anisotropía pero las observaciones con las que se cuenta no son suficientemente significativas por lo que la estimación no es tan cercana como se esperaría. Por último, se tiene el escenario de D3MCc100, el cual destaca una gran importancia debido a que no permite encontrar adecuadamente los ejes de anisotropía. Aunque se cuenta con una cantidad de observaciones suficiente y un muestreo que se consideraría adecuado, la dirección de anisotropía es casi contraria a la de los datos originales del tipo 3 y la razón posible es que a pesar de que la muestra tiene suficientes observaciones, éstas no son suficientemente significativas y por lo tanto no permite identificar adecuadamente la desviación, con lo cual se afecta considerablemente a la estimación. Continuando con el grupo de escenarios de 400 observaciones es notable la diferencia en el ajuste y cálculo de los variogramas debido a la gran cantidad de información adicional con la que se cuenta respecto al grupo de escenarios de 36 observaciones. En este grupo se vuelve visible la dirección de anisotropía de igual manera que en los datos originales del tipo 3. Debido a que se cuenta con mayor cantidad de información se hace notable nuevamente que el tipo de muestreo cuando se cuenta con una gran cantidad de información es menos relevante. El escenario D3MCc400 tiene el mapa de estimación más cercano respecto al de los datos originales del tipo 3. Los escenarios D3MRc400 y D3MAc400 también tienen un mapa de estimación bastante cercano al de los datos originales del tipo 3. Por lo tanto, para la característica de anisotropía destaca mucho la importancia de la ubicación de las observaciones cuando se cuenta con poca información y disminuye esta importancia cuando el número de observaciones es grande.

Para todos los escenarios dentro del caso de estudio de datos del tipo 3, aunque no se conocía por adelantado la característica que presentaba la base

de datos, los gráficos más importantes para identificar anisotropía son los variogramas en 4 direcciones y el mapa de anisotropía que utiliza el cálculo de estos variogramas. Una vez identificada la existencia de anisotropía es de gran importancia realizar un mayor número de variogramas en diferentes direcciones para encontrar la dirección de máxima continuidad y por ende los ejes de anisotropía. También son importantes gráficos como el histograma, ya que la asimetría puede afectar la visibilidad sobre la adecuada dirección de anisotropía, así como los gráficos mencionados para identificar tendencia, debido a que esta característica puede influir sobre la dirección de anisotropía.

Debido a que para todos los escenarios la asimetría fue notoria, también destaca la posibilidad de utilizar una transformación diferente a la logarítmica, con la cual sea posible disminuir esta asimetría, pero eso queda a decisión del geoestadístico. Por el lado de la característica de anisotropía, es recomendable también conocer información sobre la variable en cuestión, así como el área sobre la cual son tomadas las muestras, debido a que la anisotropía puede ser identificada también en base a información cualitativa.

7.4. Resultados Generales

Es importante reiterar que cada escenario de cada uno de los casos de estudio fue realizado independientemente de cualquier escenario presentado en este trabajo, por lo cual es posible observar las similitudes y diferencias entre el conjunto de todos los escenarios presentados.

Iniciando con la importancia del análisis exploratorio de datos, destaca una característica recurrente dentro de las bases de datos de todos los escenarios y de los datos originales de cualquier tipo y esta es la asimetría significativa que se observa, independientemente de si son pocos o muchos datos o si es uno u otro tipo de muestreo la asimetría se muestra en cada escenario de manera contundente. Esta asimetría es tratada con diferentes transformaciones dependiendo de la base con la cual se está realizando el estudio. El análisis exploratorio de datos provee de herramientas para conocer mejor la distribución de los datos, así como indicadores para las diferentes características. Las herramientas indispensables como las estadísticas básicas, el histograma, q-q plot, gráfico respecto a las coordenadas, gráfico de estacionariedad, etc. muestran las características de datos atípicos, asimetría, estacionariedad, tendencia, entre otras, con lo cual se determina el proceso a seguir dentro de la metodología de estimación. Por lo tanto, es indispensable ser cuidadoso al realizar este análisis.

Continuando, el análisis estructural es donde se calculan los variogramas, los cuales son herramientas que permiten adaptar el modelo para la estimación. Los variogramas permiten confirmar características como la tendencia y anisotropía, o incluso si es suficientemente evidente las estructuras anidadas. Dentro del análisis estructural se realizan las pruebas de validación cruzada y análisis de residuales, con las cuales el modelo elegido es probado para conocer la confiabilidad o cercanía con la que se cuenta para la estimación. En particular, en este trabajo de tesis se utilizó kriging ordinario para todas las estimaciones.

Para el tamaño de muestra, resultó evidente que cuando se cuenta con una mínima cantidad de información la estimación no es muy confiable aún sin tener características influyentes sobre el modelo y hacia la estimación. Si se cuenta con suficiente información significativa es posible llegar a una estimación aceptable si se adapta un modelo adecuado. Pero si se cuenta con una gran cantidad de información, el ajuste del modelo es más fácil, evidente y confiable.

En general, para el objetivo de los escenarios con poca, suficiente y gran cantidad de información fueron evidentes las aportaciones sobre el proceso a seguir y se cumplió con lo que se esperaba al ser un análisis complejo al tener poca información, una análisis menos complejo cuando se cuenta con suficiente información y un análisis más sencillo cuando se cuenta con gran cantidad de información.

El tipo de muestreo es importante para darle mayor significancia a las observaciones con las que se cuenta. Es decir, ayuda significativamente cuando se cuenta con características influyentes que dependen de la ubicación de las observaciones, por ejemplo, en los escenarios donde se tiene anisotropía es más fácil identificarla si se cuenta con un muestreo regular o combinado que en los escenarios de muestreo aleatorio. Sin embargo, si se tiene una gran cantidad de información, el tipo de muestreo no resulta tan importante ya que se abarca gran parte del área total y por ende se cuenta con mucha información significativa. Por otro lado, se esperaba que el muestreo regular fuera el más sencillo para cada caso de estudio, debido a que se pensaría que se cuenta con mayor información significativa por la ubicación elegida, no obstante algunos escenarios con muestreo combinado o aleatorio permitieron identificar más adecuadamente las características de la muestra.

Los 3 casos de estudio presentan por separado 3 características influyentes dentro del proceso de estimación, sin embargo, en la realidad es posible que se presente más de una característica en la misma base de datos u otras características que no fueron mencionadas en este trabajo, por lo que se recomienda utilizar las herramientas, indicadores, procesos, gráficos y toda la información presentada en este trabajo como apoyo visual y analítico para identificar y tratar las diferentes afectaciones que puedan presentarse al realizar algún otro estudio.

Con todo esto, se termina resaltando que durante el proceso de aplicación de la metodología de estimación con kriging, la confiabilidad y eficiencia de la estimación geoestadística está basada en el adecuado proceso del análisis de la información, en particular del análisis exploratorio de datos, así como una cantidad suficiente de la misma informción, un acertado ajuste del modelo de variograma a utilizar y una adecuada valoración del mismo para llegar a una estiamción exitosa.

Bibliografía

[1] Edward H. Isaaks y R. Mohan Srivastava: *An Introduction to Applied Geostatistics*, New York, Oxford University Press, 1989.

[2] Martín A. Díaz Viera: *Notas de Geoestadística*, http://mmc2.geofisica.unam.mx/gmee/, Instituto de Geofísica de la UNAM, 2002.

[3] Michael Edward John: *Geostatistics and petroleum geology*, Second Edition, Kluwer Academic Publishers, 1999

[4] Olivier Dubrule: *Geostatistics in Petroleum Geology*, The American Association of Petroleum Geologists, 1998.

[5] Peter J. Diggle y Paulo J. Ribeiro Jr.: *Modeled based Geostatistics*, Springer Series in statistics, XIII, 2007.

[6] María Rosa Cañada Torrecilla: *Aplicación de la Geoestadística al estudio de la variabilidad espacial del ozono en los veranos de la comunidad de Madrid*, Departamento de Geografía, Universidad Autónoma de Madrid, 2004.

[7] Echeverría J.C., Molinero H.B., Serra J.A. y Peña Zubiate C. : *Evaluación de Recursos Naturales con Geoestadística y Kriging*, IV Jornadas Cuidemos nuestro mundo (CNM) para contribuir a la implementación de un modelo ambiental para San Luis, 1996, 155p.

[8] Geo-Eas 1.2.1, *User's Guide*, United States Enviromental Protection Agency, 1991.

[9] Richard L. Chambers, Jeffrey M. Yarus y Kirk B. Hird: *Petroleum Geostatistics for Non Geostatisticians*, The Leading Edge, 2000.

[10] Akhil Datta-Gupta: *Notas de curso*, TAMU Petroleum Engineering, Harold Vance Department of Petroleum Engineering, 1999.

[11] Clayton V. Deutsch y André G. Journel: *Geostatistical Software library and User's Guide*, Oxford University Press, 384pp, 1997.

[12] A.G. Journel: *Fundamental of Geostatistics in five lessons*, American Geophysical Union, 40pp, 1989.

[13] Tarek Ahmed y Paul D. McKinney: *Advanced Reservoir Engineering*, Elsevier, 2005.

[14] Ronald Paul Barry: *A Diagnostic to Asses the Fit of a Variogram Model to Spatial Data*, Department of Math Sciencies.

[15] F.J. Moral García: *Aplicación de la geoestadística en las ciencias ambientales*, Ecosistemas, 78-86. Enero 2004, http://www.revistaecosistemas.net/articulo.asp?Id=167.

[16] *Moran's I Anlaysis Window*: GS+ for Windows Overview, Gamma Design Software, Professional Geostatistics for the Environmental Sciences, http://mmc2.igeofcu.unam.mx/cursos///geoest/Software/GS+/Manual/Geostatistics%20for%20the%20Environmental%20Sciences%20Moran's%20I%20Analysi0̃.

[17] José María Montero Lorenzo y Beatriz Larras Iribas: *Introducción a la geoestadistica lineal*, netbiblo, p.16, 2008.

[18] *Teoría Geoestadística*, http://geoestadistica.com/funcion_aleatoria.htm.

[19] Francisco Jesús Moral García: *Representación Gráfica de la distribución espacial de una plaga en una plantación mediante el uso de técnicas geoestadísticas*, Cursos de sistemas de información Geográfica por internet, 2003, http://www.mappinginteractivo.com/plantilla-ante.asp?id_articulo=266.

[20] Salvador Figueras M. y Gargallo Valero P. : *Análisis Exploratorio de Datos*, 5campus.com, Estadísitica, 2008, http://www.5campus.com/leccion/aed.

[21] Francisco Jesús Moral García y José Rafael Marques da Silva: *Ejemplo de representación gráfica de una variable regionalizada*, XIV Congreso Internacional de Ingeniería Gráfica, Santander, España 2002.

[22] Evan J. Englund: *Variance of Geostatisticians*, Mathematical Geology, 22:4, 417-455, 1990.

[23] Evan J. Englund y Dennis Weber: *Evaluation and Comparison of Spatial Interpolators*, Mathematical Geology, 22:4, 381-391, 1992.

[24] Evan J. Englund y Dennis D. Weber: *Evaluation and Comparison of Spatial Interpolators II*, Mathematical Geology, 1993.

[25] Javier Méndez Venegas: *Modelación de la distribución espacial de la precipitación en el valle de la ciudad de México usando técnicas geoestadísticas*, Tesis de Maestro en Ciencias, Colegio de Postgraduados, Estado de México, 2008.

[26] Martín Díaz Viera, Victor Hernández Maldonado y Javier Méndez Venegas, *R geoestad*, Software, 2011.

[27] *GS+*, LLC Gamma Design Software, 1988.