UNIVERSIDAD DE CONCEPCIÓN

FACULTAD DE INGENIERÍA
DEPARTAMENTO DE INGENIERÍA METALÚRGICA

PROFESORES PATROCINANTES

SR. ROBERTO FUSTOS TORIBIO
SR. FRANCISCO MUÑOZ GUTIÉRREZ

MODELOS DE REGRESIÓN EN EVALUACIÓN DE YACIMIENTOS

PAULA FRANCISCA GONZÁLEZ FARIÑA

INFORME DE MEMORIA DE TÍTULO
PARA OPTAR AL TÍTULO DE
INGENIERO CIVIL DE MINAS

NOVIEMBRE 2020

Esta memoria de título está dedicada con cariño

a mi familia, y en especial a mi abuelo

Lucio Fariña Fernández,

que orgulloso debe estar en el cielo.

AGRADECIMIENTOS

Esta Memoria de Título concreta la culminación de un desafío que decidí comenzar el 2014, un desafío no exento de sentimientos, que me tuvo al límite más de alguna vez, pero que me preparó y me guió para ser la persona y la profesional que soy hoy en día.

Primero quiero agradecer el apoyo de mi familia, en especial de mi madre Mónica Fariña quien confió en mi desde el principio y me desafío para lograr mis objetivos.

Segundo quiero agradecer a todos mis amigos de la universidad partiendo por Camila Godoy y Zoraya González quienes fueron las primeras personas que conocí y hasta el día de hoy me alientan a ser mejor, a los grandes amigos que hice en mi carrera, Pamela Mella, Valentina Neira, Cristian Garrido y a su posterior remplazo Rafael Lara quienes celebraron conmigo en los mejores momentos, me ayudaron en los más difíciles, estuvieron conmigo hasta el último minuto y además me enseñaron que todo siempre puede ser peor así que todo hay que tomárselo con un poco humor y ligereza.

Además, quiero agradecer a todos los profesores y profesionales que conocí en la universidad y me enseñaron más que solo conocimientos técnicos, en especial a Roberto Fustos por su guía y comprensión.

Por último, quisiera agradecer a todas y cada una de las personas que compartieron conmigo mi experiencia en la universidad y que no necesito nombrar porque tanto ellos como yo sabemos que desde lo más profundo les agradezco haberlos conocido y el haberme dado todo el apoyo, ánimo y sobre todo amistad.

RESUMEN

La estimación de recursos minerales (ERM) es un proceso indispensable para el desarrollo de un proyecto minero, esto producto que en esta etapa se puede determinar la continuidad o no continuidad del proyecto, basados en si el depósito es capaz de proporcionar un beneficio económico.

En consecuencia, es necesario que la ERM sea lo más precisa posible, pero en la realidad producto de la baja cantidad de información que se disponen provenientes de sondajes de exploración se hace imposible que el proceso de la adquisición, procesamiento e interpretación de estos datos este lleno de subjetividades por parte del personal que maneja la información lo que puede conllevar a tener una errónea o pobre interpretación de la distribución real del mineral del depósito estudiado.

La problemática que se aborda en esta Memoria de Título se relaciona directamente con los modelos de regresión lineal utilizados en la actualidad para la realización de la ERM analizando las principales limitaciones y complicaciones de estos modelos a la hora de realizar la ERM.

La metodología consiste en la construcción de dos escenarios, el posterior análisis de los resultados y una categorización de recursos; el primero de estos escenarios se basa en la construcción de un conjunto de datos simulado con el que se obtendrá un modelo de bloques, de este se selecciona un porcentaje datos y se aplican los modelos de regresión: Regresión Beta, Regresión Geográficamente Ponderada, Kriging Bayesiano, Inverso de la distancia y Kriging ordinario generando nuevos modelos de bloques, finalmente se comparan ambos modelos de bloques a través de la raíz del error cuadrático medio (RMSE); en el segundo escenario se realiza una ERM a un conjunto de datos reales de campo repitiendo el procedimiento anterior.

Para el caso simulado se destaca la gran eficacia de los cinco métodos para la estimación, presentado los menores resultados de RMSE el método Kriging Bayesiano, seguido de Kriging Ordinario utilizando un menor costo computacional, y por el caso contrario presentando los mayores resultados de RMSE se encuentra el método Regresión beta.

En cuanto al caso real fue necesario optimizar los parámetros de los algoritmos a través de validación cruzada teniendo el menor resultado de RMSE el método Inverso de la distancia y el mayor valor de RMSE el método Regresión Beta.

ABSTRACT

Estimation of mineral resources (EMR) is an essential part in the development of a mining project, this stage it is critical because the project it will be determined if can continue or not, based on the criteria if the deposits are capable of providing an economic benefit.

It is necessary for EMR to be as accurate as possible, but in reality, due to the low amount of information available from exploration drilling, it is impossible that the process of acquisition, processing and interpretation of the data aren't full of subjectivity because of the staff who handle the information, which can lead to an erroneous or subjective interpretation of the actual mineral distribution in the deposit studied.

The problem addressed in this work is directly related to the linear regression models currently used to perform the EMR, analyzing the mainly limitations and complications of these models when performing the EMR.

The methodology consists in the construction of two scenarios, the subsequent analysis of the results and then a resources classification; the first of these scenarios is based on the construction of a simulated data set with which a block model is built, from this set a portion of data is selected and then regression models are applied, these are: Beta Regression, Geographically Weighted Regression, Kriging Bayesian, Inverse of the distance and ordinary Kriging, generating new block models, finally the resulting block models are compared through the root mean square error (RMSE); in the second scenario, an EMR is performed on a real data set repeating the previous procedure.

For the simulated case, the great effectiveness of the five estimation methods is emphasized, getting the lowest RMSE results with the Bayesian Kriging method followed by Ordinary Kriging using a lower computational cost, and for the opposite case getting the highest RMSE results with the Beta Regression method.

As for the real case it was necessary to optimize the parameters of the algorithms through cross validation, getting the lowest RMSE result from the Inverse of the distance and the highest RMSE value from the Beta Regression method.

ÍNDICE DE CONTENIDOS

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

ÍNDICE DE ANEXOS

NOMENCLATURA

ERM	:	Estimación de recursos minerales
RMSE	:	Raíz del error cuadrático medio
GLM	:	Modelo Lineal Generalizado
BREG	:	Regresión Beta
CV	:	Validación cruzada
AIC	:	Criterio de información de Akaike
AICc	:	Criterio de información corregido de Akaike
MV	:	Máxima Verosimilitud
OLS	:	Mínimos cuadrados ponderados
GWR	:	Regresión Ponderada Geográficamente
OK	:	Kriging ordinario
KU	:	Kriging Universal
BKR	:	Kriging Bayesiano
IDW	:	Inverso de la distancia
POR	:	Textura Porfídica
AFA	:	Textura Afanítica
MAC	:	Textura Maciza
BRE	:	Textura Brechosa
MET	:	Tipo de roca Metandesitas
HIE	:	Tipo de roca Hierro de mena
BRH	:	Tipo de roca Brecha
AND	:	Tipo de roca Andesita
DIO	:	Tipo de roca Diorita

CAPÍTULO 1

INTRODUCCIÓN

1.1 ANTECEDENTES

La Geoestadística es una rama de las matemáticas aplicadas que se originó en la industria minera a principios de los años cincuenta, para ayudar a mejorar los cálculos de las reservas minerales. Los primeros pasos se dieron en Sudáfrica, con el trabajo del ingeniero de minas D. G. Krige y el estadístico H. S. Sichel [1], este último observó la naturaleza asimétrica de la distribución del contenido de oro en las minas sudafricanas, la equiparó a una distribución de probabilidad log normal y desarrolló las fórmulas básicas para esta distribución. Ello permitió una primera estimación de las reservas, pero bajo el supuesto de que las mediciones eran independientes, una clara contradicción con la experiencia, dado que los fenómenos geológicos se rigen por patrones determinados y no por un proceso aleatorio.

Una primera aproximación a la solución de este problema fue dada por D. G. Krige quien propuso una variante del método de medias móviles, que puede considerarse como el equivalente al método conocido como Kriging Simple, que es uno de los métodos de estimación lineal espacial con mayores cualidades teóricas [2].

A finales de los años cincuenta, las técnicas atrajeron la atención de ingenieros, en particular de G. F. Matheron quien desarrolló los conceptos innovadores de D. G. Krige y los puso en un marco único, en la teoría: “Variables regionalizadas” [3], nombre con el que se les denomina a las variables aleatorias que exhiben relaciones de estructura espacial.

En los últimos 30 años la geoestadística ha probado su superioridad como método de estimación de los recursos minerales en la mayoría de los tipos de minas, donde ha demostrado su utilidad para modelar y simular la compleja heterogeneidad geológica interna de una gran variedad de reservas minerales.

Su empleo también ha sido extendido a otros campos, tales como, las ciencias medio ambientales y ecológicas, hidrogeología, agricultura, geografía e incluso campos como la pesca, donde el factor tiempo, al igual que la variabilidad espacial, juegan un rol importante [3].

En la actualidad, el desarrollo de la informática ha propiciado condiciones para su divulgación y aplicación a un grupo mayor de problemas, es así como se pueden encontrar en el mercado programas informáticos profesionales de geoestadística que ofrecen opciones para la aplicación de estas técnicas.

1.2 MOTIVACIÓN EN MINERÍA

En un proyecto minero (ver **Figura 1**), luego de la etapa inicial de prospección minera le sigue una etapa de exploración, esta última tiene como principal objetivo lograr un conocimiento detallado del depósito mineral, determinando de la manera más precisa posible sus dimensiones y enriquecimiento mineral.

Al finalizar la etapa de exploración se obtiene como resultado un modelo de recursos, con el cual se procederá a realizar un estudio de prefactibilidad, la cual es una etapa crucial dado que en esta se determina la continuidad del proyecto, basados en si el depósito es capaz de proporcionar un beneficio económico.

El modelo de recursos resultante del paso anterior corresponde al modelo que entrega la cuantificación de recursos minerales, que se obtiene del modelo de bloques; a su vez el modelo de bloques se construye utilizando métodos de estimación como son los métodos de regresión lineal presentes en esta Memoria de Título , por lo tanto, se puede afirmar que contar con una estimación de recursos minerales (ERM), lo más cercano a la realidad de depósito, es vital en la viabilidad de un proyecto minero.

La elección del mejor método de estimación a utilizar es el modelo que aproveche al máximo la escasa información proveniente de la etapa de prospección (principalmente obtenida de sondajes de exploración) y que pueda minimizar el error de estimación, lo cual no es una labor sencilla dado que estos métodos están condicionados a la cantidad de información disponible y a las restricciones intrínsecas que posee cada uno.

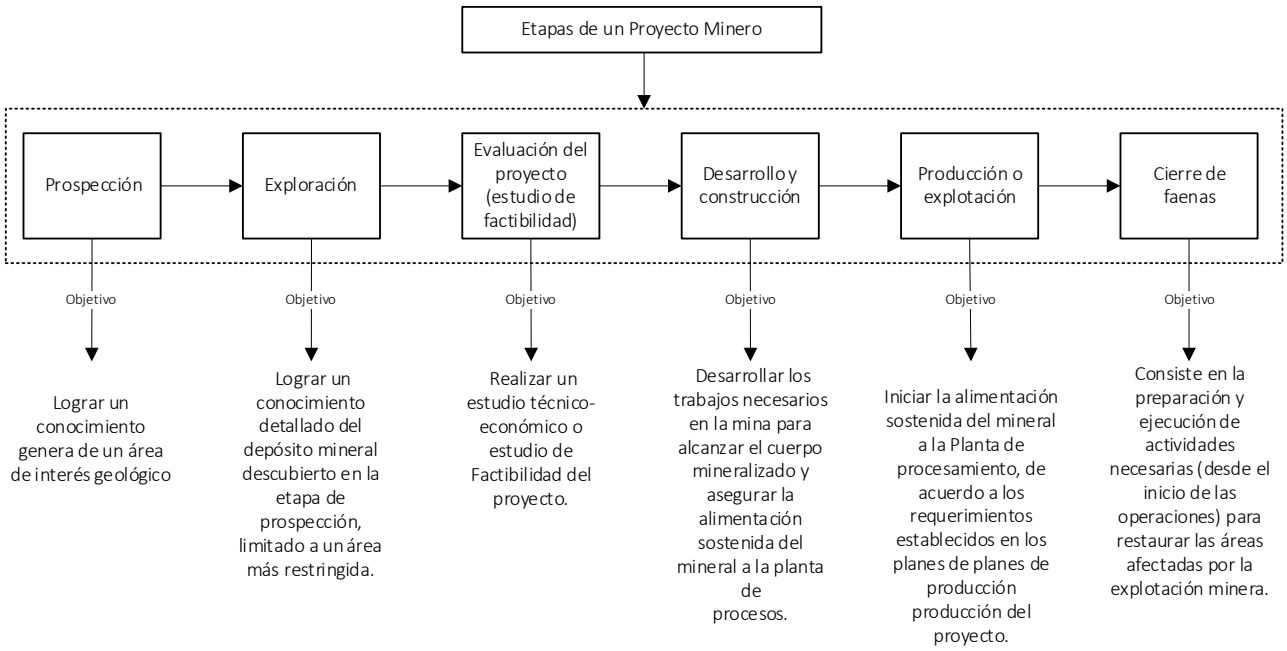


Figura 1: Etapas de un proyecto minero

1.3 DESCRIPCIÓN DEL PROBLEMA

La problemática que se aborda en esta Memoria de Título se relaciona directamente con los modelos de regresión lineal utilizados en la actualidad para la realización de la ERM, además se explicarán y se analizarán las principales limitaciones y complicaciones de estos modelos a la hora de realizar la ERM.

Como se introdujo anteriormente, la cuantificación de recursos es un proceso de gran incidencia en un proyecto minero, por lo que es necesario que la ERM sea lo más precisa y exacta posible, pero en la realidad producto de la baja cantidad de información disponible, provenientes de sondajes de exploración, se hace imposible que el proceso de la adquisición, procesamiento e interpretación de estos datos este lleno de vicios y subjetividades por parte del personal que lo maneja la información.

Lo descrito anteriormente puede conllevar a tener una errónea o pobre interpretación de la distribución real del mineral del depósito estudiado y a su vez ocasiona que el modelo de bloques resultante posea un alto nivel de incertidumbre.

En la actualidad se conocen diversos algoritmos lo cuales se utilizan para realizar la ERM, pero estos están condicionados, por un lado a la cantidad de información confiable que se tenga disponible de los sondajes de explotación obtenidos en la etapa de prospección minera; y por el otro lado, también están condicionados intrínsecamente por las construcciones propias de las estimaciones, ejemplo de esto es el sesgo condicional en el método de Kriging o la falta de una media del error en el caso de la Interpolación por Inverso de la Distancia (ambos métodos son utilizados con frecuencia para la ERM y serán comparados con otros métodos propuestos más adelante). Estos problemas generan complicaciones en la construcción del modelo de bloques, los que acarrean problemas que pueden tener incidencia significativa en los procesos posteriores, produciendo así un impacto económico en el proyecto minero.

El caso ideal sería contar con una construcción del modelo de bloques precisa y con la menor incertidumbre posible, tomando en cuenta la baja cantidad de información disponible. Para esto es necesario utilizar un conjunto de algoritmos que permitan sortear los problemas antes descritos y que puedan entregar la mayor cantidad de información sobre la distribución mineral, haciendo que el riesgo económico sea mínimo.

1.4 HIPÓTESIS

Es posible disminuir la incertidumbre resultante durante la construcción del modelo de bloques y además obtener una medida del riesgo de estimación utilizando diversos algoritmos de predicción basados en Modelo de Regresión Lineal que son aplicados a datos provenientes de sondajes de exploración.

1.5 OBJETIVOS

1.5.1 OBJETIVO GENERAL

El objetivo general de esta Memoria de Título corresponde a la aplicación de tres modelos de regresión lineal para realizar la construcción de un modelo de bloques y la obtención de una medida de la incertidumbre de la distribución mineral.

1.5.2 OBJETIVOS ESPECÍFICOS

1. Aplicar los modelos de regresión lineal (Kriging Bayesiano, regresión geográficamente ponderada y regresión beta) para la construcción de modelos de bloques y cuantificación de la incertidumbre.
2. Analizar el impacto del uso de estos algoritmos de regresión lineal en la estimación y categorización de recursos minerales.
3. Realizar una comparación de diferentes escenarios de estimación utilizando los métodos actuales (Método de Kriging, Inverso de la distancia e interpolación) con los propuestos.

CAPÍTULO 2

MARCO TEÓRICO

2.1 CONCEPTOS Y DEFINICIONES

2.1.1 MODELO DE REGRESIÓN

La regresión abarca una amplia gama de métodos para modelar la relación entre una variable dependiente y un conjunto de una o más variables independientes. La variable dependiente se conoce como la variable y, variable respuesta o regresiva y las variables independientes se conocen como variables x, variable predictora o regresora. En su forma más simple, un modelo de regresión lineal puede tomar la forma:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad con\ i = 1 \dots n \qquad \text{Ec. 2.1}$$

Donde,

- y_i: Variable dependiente en la ubicación i.
- x_i: Variable independiente.
- ε_i : Error asociado en la ubicación i.
- β: Coeficiente de regresión.

Los parámetros por estimar deben ser de manera que el valor $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ se minimice sobre las n observaciones, donde y_i e $\hat{y}_i$ son los valores estimado y real respectivamente para la i-ésima observación. La expresión se conoce como residual para la i-ésima observación [4].

La capacidad del modelo para replicar los valores y observados se mide a través de la bondad del ajuste, esta se expresa comúnmente por el valor de r^2 y toma valores entre 0 y 1 (tomando valores más cercanos a 1 cuando el modelo se ajusta mejor a los valores observados), este valor cuantifica la proporción de variación entre el valor y observado y el real.

2.1.2 REGRESIÓN CON DATOS ESPACIALES

Una gran cantidad de modelos estadísticos simples consideran muestras aleatorias, estas suponen variables aleatorias independientes e idénticamente distribuidas, donde la independencia es un supuesto conveniente que hace que la teoría estadística sea más tratable. Sin embargo, los modelos que involucran dependencia estadística en las variables son más reales, esta dependencia puede verse también en los residuos del modelo.

La existencia de estructura espacial en los datos significa que el valor de la variable dependiente en una unidad espacial se ve afectado por las variables independientes en las unidades cercanas. Esto conduce a estimaciones de parámetros que son parciales e ineficientes.

2.1.3 VARIABLE REGIONALIZADA

La idea de variable regionalizada se utiliza para cualificar un fenómeno que se desarrolla en el espacio y/o tiempo y que presenta una cierta estructura de autocorrelación. De manera más formal se puede definir como un proceso estocástico con dominio contenido en un espacio euclidiano d-dimensional R^d, $\{Z(x): x\, \epsilon D\ \subset R^d\}$. Si $d = 2$, $Z(x)$ puede asociarse a una variable medida en un punto x del plano. En términos prácticos $Z(x)$ puede verse como una medición de una variable aleatoria.

Un proceso estocástico es una colección de variables aleatorias indexadas; esto es, para cada x en el conjunto de índices $D, Z(x)$ es una variable aleatoria. En el caso de que las mediciones sean hechas en una superficie, entonces $Z(x)$ puede interpretarse como la variable aleatoria asociada a ese punto del plano (x representa las coordenadas, planas o geográficas, y Z la variable en cada una de ellas). Estas variables aleatorias pueden representar la magnitud de una variable ambiental medida en un conjunto de coordenadas de la región de estudio [5].

2.1.4 ESTACIONARIEDAD

La variable regionalizada es estacionaria si su función de distribución conjunta es invariante respecto a cualquier translación del vector h, o lo que es lo mismo, la función de distribución del vector aleatorio $\vec{Z}(x) = [Z(x_1), Z(x_2), \dots, Z(x_n)]^T$ es idéntica a la del vector $\vec{Z}(x) = [Z(x_1 + h), Z(x_2 + h), \dots, Z(x_n + h)]^T$ para cualquier h.

2.1.5 VALIDACIÓN CRUZADA

Para evaluar los resultados de la interpolación se utiliza el método de la validación cruzada. La idea central de este procedimiento consiste en simular la no existencia de los puntos con datos (uno por uno) y realizar la interpolación para estimar los valores en estos puntos., esto dará a conocer el error para cada punto [6].

Existen métricas utilizadas frecuentemente para la evaluación del desempeño de los modelos y comparación entre ellos, algunas de estas métricas son (para concepto de la Memoria de Título se utiliza RMSE):

- Raíz del error cuadrático medio (RMSE).
- Error absoluto medio ponderado (MAPE).
- Error absoluto medio (MAE).
- Coeficiente de determinación (R2).
- Validación cruzada generalizada (GCV).

2.1.6 RAÍZ DEL ERROR CUADRÁTICO MEDIO (RMSE)

RMSE (*Root Mean Square Error)* es una medida de uso frecuente de la diferencia entre los valores pronosticados por un modelo y los valores realmente observados. Estas diferencias individuales son también llamadas residuos y el RMSE sirve para agregar en una sola medida la capacidad de predicción [7].

Se define como la raíz cuadrada de la media de los errores al cuadrado:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Z_i^* - Z_i)^2} \qquad \text{Ec. 2.2}$$

Donde,

- n: Número de muestras.
- Z_i^*: Valor estimado.
- Z_i: Valor real.

2.2 REGRESIÓN BETA

El modelo de regresión beta es comúnmente utilizada para modelar variables que asumen valores en el intervalo unitario estándar $(0, 1)$. Se basa en el supuesto de que la variable dependiente está distribuida en beta y que su media está relacionada con un conjunto de regresores a través de un predictor lineal con coeficientes desconocidos y una función de enlace (ver anexo 8.1). El modelo también incluye un parámetro de precisión que puede ser constante o depender de un conjunto de regresores a través de una función de enlace [4].

2.2.1 ANTECEDENTES

El modelo de regresión lineal para analizar datos y estudiar la relación entre algunas variables, es de uso común en diversas aplicaciones; sin embargo, dicho modelo es inadecuado cuando la distribución natural de la variable respuesta no es bien aproximada por la distribución de Gauss [8] (ver anexo 8.2).

La práctica habitual utilizada para realizar un análisis de regresión, en la cual la variable dependiente y asume valores en el intervalo de unidad estándar $(0, 1)$, estaba basada en transformar los datos de forma que la respuesta transformada sea $\tilde{y}$, luego se le aplicaba un análisis de regresión lineal estándar, sin embargo, el enfoque nombrado anteriormente presenta algunos inconvenientes que son [8]:

- Los parámetros de regresión son interpretables en términos de la media de $\tilde{y}$, y no en términos de la media de y, dada la desigualdad de Jensen[1].
- Las regresiones que involucran datos del intervalo $(0, 1)$ como las proporciones son típicamente heterocedásticas (ver título 2.2.2.2).
- Las distribuciones de las proporciones son típicamente asimétricas, y por lo tanto las aproximaciones con base gaussiana para la estimación de intervalo y la prueba de hipótesis puede ser bastante inexacta en muestras pequeñas.

[1] En algo que es cóncavo o convexo, al tocar dos de sus puntos con una recta, todo aquello que se sitúe entre dichos dos puntos quedará siempre por encima o por debajo del segmento con el que se trabaja.

Ferrari & Cribari-Neto (2004) propusieron un modelo de regresión para variables aleatorias continuas que asumen valores en el intervalo unidad estándar, como es el caso de las proporciones. Dado que el modelo se basa en la suposición de que la respuesta posee una Distribución Beta (ver Ec. 2.4) y le otorgaron el nombre modelo de Regresión Beta.

La motivación principal para el modelo de regresión beta radica en la flexibilidad entregada por la ley beta asumida. La densidad beta puede asumir un número de formas diferentes dependiendo de la combinación de valores de parámetros, incluyendo la forma izquierda y derecha sesgada o la forma plana de la densidad uniforme. Esto se ilustra en la Figura 2 que representa diferentes densidades beta.

Las densidades se parametrizan en términos de la media μ y el ϕ parámetro de precisión. La flexibilidad evidente hace que la distribución beta un candidato atractivo para la modelización estadística basada en datos.

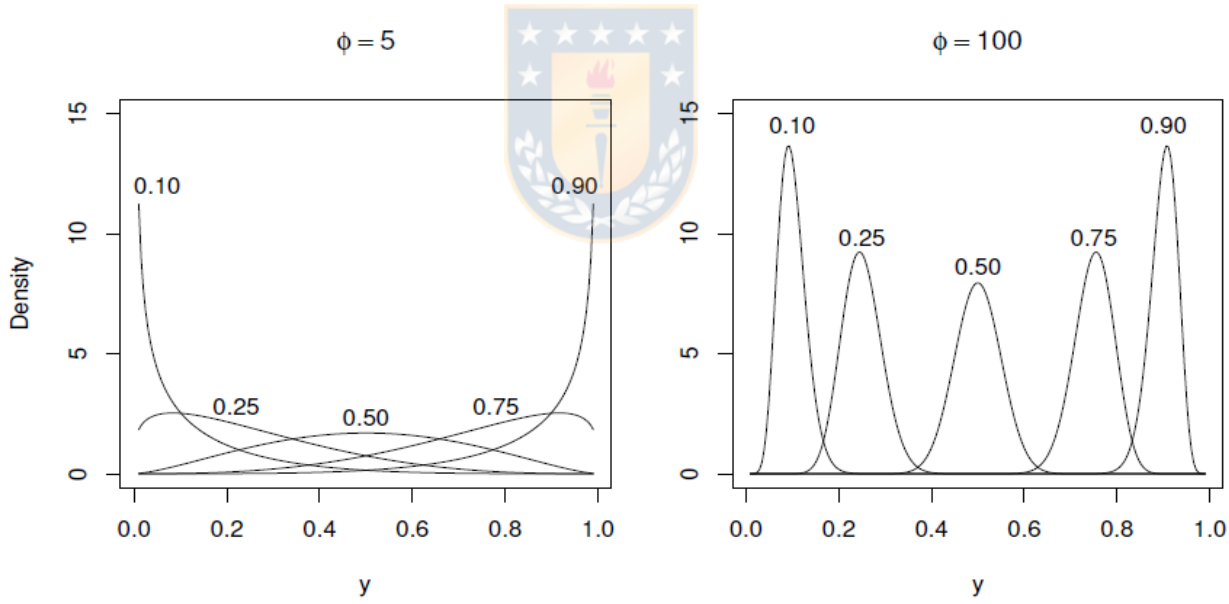


Figura 2: Función de densidad de probabilidad con distribución beta

Simas, Barreto-Souza & Rocha (2010) propusieron una variante del modelo de regresión beta que permite no linealidades y dispersión variable. En particular, en este modelo más general, no se supone que el parámetro que explica la precisión de los datos sea constante en las observaciones, pero se permite que varíe, lo que lleva al modelo de Regresión Beta de dispersión variable [4].

2.2.2 DESCRIPCIÓN

El análisis de regresión con un resultado acotado es un problema común en la estadística aplicada. Se considera entonces la Regresión Beta [9], que está directamente relacionada con el enfoque de los Modelos Lineales Generalizados[2], para el modelado conjunto de medias y dispersiones descritas por McCullagh y Nelder (1989). Smyth (1989) desarrolló este enfoque para variables aleatorias de las familias de la distribución exponencial (ver anexo 8.3) y como la distribución beta esta indexada por dos parámetros se adapta fácilmente a ella. El enfoque clásico para ajustar un modelo de regresión beta es utilizar la estimación de máxima verosimilitud (ver anexo 8.4) con la posterior selección de variables basadas en Criterio de información de Akaike AIC[3].

2.2.2.1 CONDICIONES DE LOS DATOS

Se considera una variable respuesta que está acotada por un valor máximo de referencia, estos valores han sido transformados mediante una función lineal para que la escala original de puntajes se mueva en el intervalo $(1, 0)$. Teniendo en cuenta que la variable toma valores en (a, b) (con $a < b$ conocido) y si y también asume los extremos $(0, 1)$, una transformación útil en la práctica es:

$$y' = \frac{(y - a)(N - 1)}{(b - a)N} + \frac{1}{2N} \qquad \text{Ec. 2.3}$$

Donde N es el tamaño de la muestra e y'tiene las propiedades deseadas expuestas en Smithson & Verkuilen (2006) [10].

2.2.2.2 HETEROCEDASTICIDAD

Se entiende heterocedasticidad como el grado en que la varianza de los residuos depende de la variable predictora (ver **FIGURA 3**). Teniendo en cuenta que los residuos también pueden variar en el modelo, se establece que los datos son heterocedastico si la cantidad de los residuos que varían en el modelo se modifica a medida que cambia la variable predictora. Es importante identificar la heterocedasticidad en el tratamiento de datos ya que complica el análisis de los resultados al ejecutar el análisis de regresión [11].

[2] Modelo Lineal Generalizado (GLM) es una generalización flexible de la regresión lineal ordinaria que permite variables de respuesta que tienen modelos de distribución de errores distintos de una distribución normal.

[3] Criterio de información de Akaike es una medida de la bondad de ajuste de un modelo estadístico, describe la relación entre el sesgo y varianza en la construcción del modelo.

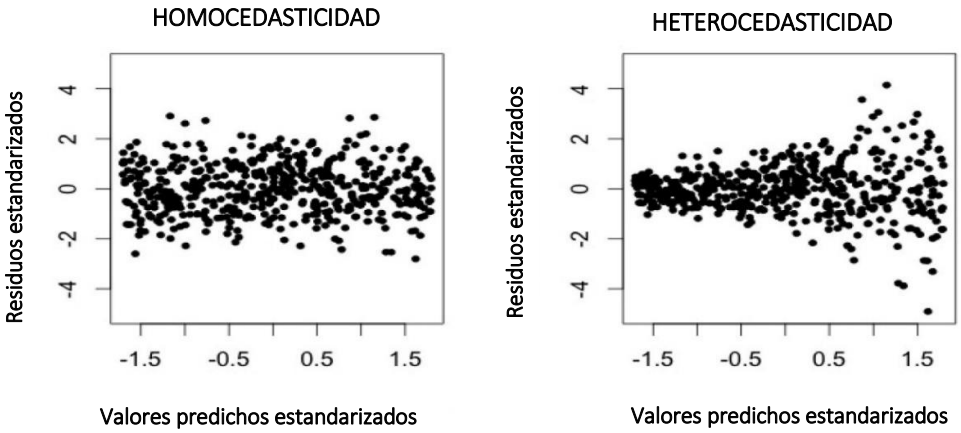


FIGURA 3: Homocedasticidad vs heterocedasticidad en residuos estandarizados.

2.2.2.3 DISTRIBUCIÓN BETA

La distribución beta es flexible ya que su densidad puede asumir diferentes formas dependiendo de los valores de los dos parámetros indexados en la distribución. Una variable aleatoria y se dice que sigue una distribución beta de parámetros $\alpha > 0$ y $\beta > 0$, $y \sim \text{Beta}(\alpha, \beta)$, si la función de densidad tiene la siguiente forma:

$$f(y;\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1} \quad , \quad 0 < y < 1 \qquad \text{Ec. 2.4}$$

$$f(y;\alpha,\beta) = \frac{1}{B(\alpha,\beta)} X^{\alpha-1}(1-y)^{\beta-1}$$

Donde $\Gamma(\cdot)$ es la función gamma y $B(\alpha,\beta)$ es la función beta:

$$B(\alpha,\beta) = \int_0^1 X^{\alpha-1}(1-y)^{\beta-1}\,dx = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \qquad \text{Ec. 2.5}$$

La media y la varianza de y son respectivamente,

$$E(y) = \frac{\alpha}{\alpha+\beta} \qquad \text{Ec. 2.6}$$

$$Var(y) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \qquad \text{Ec. 2.7}$$

Para la función de densidad (ver Ec. 2.4) cuando $\alpha = \beta$ se le da el nombre de "Distribución Beta Simétrica" (la distribución uniforme es un caso particular cuando $\alpha = \beta = 1$), por el contrario, cuando $\alpha \neq \beta$ se le da el nombre de "Distribución beta asimétrica".

2.2.3 ALGORITMO

El modelo de Regresión Beta se basa en una parametrización alternativa de la densidad beta en términos de la media variable y un parámetro de precisión. Ferrari y Cribari-Neto (2004) propusieron una parametrización diferente estableciendo $\mu = \alpha/(\alpha + \beta)$ y $\emptyset = \alpha + \beta$. Además, considerando $0 < \mu < 1$ y $\emptyset > 0$ se obtiene $y \sim B(\mu, \emptyset)$ con esperanza y varianza mostradas respectivamente en las ecuaciones Ec. 2.9 2.9 y Ec. 2.10:

$$f(y;\mu,\emptyset) = \frac{\Gamma(\emptyset)}{\Gamma(\mu\emptyset)\Gamma\big((1-\mu)\emptyset\big)} y^{\mu\emptyset-1}(1-y)^{(1-\mu)\emptyset-1}, \quad 0 < y < 1 \qquad \text{Ec. 2.8}$$

$$E(y) = \mu \qquad \text{Ec. 2.9}$$

$$VAR(y) = \frac{\mu(1-\mu)}{(1+\emptyset)} \qquad \text{Ec. 2.10}$$

Donde, $\emptyset$ es el parámetro de precisión ya que cuanto mayor sea el parámetro μ menor será la varianza de y.

Sea $y_1, y_2, y_3, \dots, y_n$ una muestra aleatoria tal que $y_i \sim B(\mu_i, \emptyset)$, con $i = 1, 2, 3, \dots, n$. El modelo de Regresión Beta se define como:

$$g(\mu_i) = x_i^T \beta = \eta_i \qquad \text{Ec. 2.11}$$

Donde $\beta = \big(\beta_1, \dots, \beta_k\big)^T$ es un vector k x 1 de parámetros de regresión desconocidos $(k < n)$, $x_i = (x_{i1}, \dots, x_{ik})^T$ es un vector de k regresores (variables independientes o covariables) y $\eta_i$es un predictor lineal, es decir, $\eta_i = \beta_1 x_{i1} + \cdots + \beta_k$.

Típicamente $x_{i1} = 1$ para todo i para que el modelo tenga una intersección. Aquí, $g(\cdot): (0,1) \mapsto \mathbb{R}$ es una función enlace, que es estrictamente creciente y dos veces diferenciable. La motivación para usar una función enlace en la estructura de la regresión es doble:

- Ambos lados de la ecuación de regresión asumen valores en la recta real cuando se aplica la función enlace μ_i.
- Mayor flexibilidad, ya que el profesional puede elegir la función que produce el mejor ajuste.

Algunas funciones de enlace útiles son: $logit\ g(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$; $probit\ g(\mu) = \Phi^{-1}(\mu)$, donde $\Phi(\cdot)$ es la función de distribución normal estándar; $log - \log\ g(\mu) = \log\{-\log(1-\mu)\}$; $log - \log\ g(\mu) = -\log\{-\log(\mu)\}$; y Cauchy $g(\mu) = \tan\{\pi(\mu - 0.5)\}$.

La varianza de y es una función que hace que el modelo de regresión basado en esta parametrización sea naturalmente heterocedastico, en particular:

$$VAR(y_i) = \frac{\mu_i(1-\mu_i)}{1+\emptyset} = \frac{g^{-1}(x_i^T\beta)[1-g^{-1}(x_i^T\beta)]}{1+\emptyset} \qquad \text{Ec. 2.12}$$

La función log-verosimilitud es $\ell(\beta, \emptyset) = \sum_{i=1}^{n} \ell_i(\mu_i, \emptyset)$, donde:

$$\ell_i(\mu_i, \emptyset) = \log\ \Gamma(\emptyset) - \log\Gamma(\mu_i\emptyset) - \log\Gamma((1-\mu_i)\emptyset + (\mu_i\emptyset - 1)logy_i(1-\mu_i)\emptyset - 1\}\log(1-y_i) \qquad \text{Ec. 2.13}$$

Nótese que $\mu_i = g^{-1}(x_i^T\beta)$ es una función de β, vector de parámetros de regresión. La estimación de parámetros se realiza por máxima verosimilitud (ML, por sus siglas en ingles *maximum likelihood*) (ver anexo 8.4).

Una extensión del modelo de Regresión Beta anterior empleada por Smithson & Verkuilen (2006) e introducida formalmente por Simas (2010) es el modelo de Regresión Beta de dispersión variable (ver anexo 8.5).

Hay varios tipos de residuales disponibles para los modelos de Regresión Beta. Los residuos de respuesta sin procesar $y_i - \hat{\mu}_i$ generalmente no se usan debido a la heterocedasticidad inherente en modelo (ver Ec. 2.12). Por lo tanto, una alternativa natural son los residuos de Pearson que Ferrari y Cribari-Neto (2004) llaman "Residuos ordinarios estandarizados", este modelo está definido por la siguiente ecuación:

$$r_{P,i} = \frac{y_i - \hat{\mu}_i}{\sqrt{\widehat{VAR}(y_i)}} \qquad \text{Ec. 2.14}$$

Donde,

$$\widehat{VAR}(y_i) = \frac{\hat{\mu}_i(1-\hat{\mu}_i)}{1+\hat{\emptyset}_i} \qquad \text{Ec. 2.15}$$

$$\hat{\mu}_i = g_1^{-1}(x_i^T \hat{\beta}) \quad \text{Ec. 2.16}$$

$$\widehat{\emptyset}_i = g_2^{-1}(z_i^T \hat{\gamma}) \quad \text{Ec. 2.17}$$

Del mismo modo, los residuos de desviación pueden definirse de la manera estándar a través de contribuciones para obtener un incremento de la probabilidad [4].

Espinheira, Ferrari y Cribari-Neto (2008) propusieron otros residuos, en particular uno residual con mejores propiedades que denominaron "Residuo ponderado estandarizado 2" (ver anexo 8.6).

2.2.4 Campo de Aplicación

La distribución beta es usada para modelar el comportamiento de variables aleatorias limitadas por intervalos de longitud finita. En particular, es una distribución adecuada para porcentajes y proporciones también tiene uso en la inferencia Bayesiana, ya que produce una familia de distribuciones de probabilidad a priori conjugadas. Algunos estudios reales relacionados con el modelo de la Regresión Beta son:

En el campo de la salud estudios como: Breton et al. (2014) *"Prenatal Tobacco Smoke Exposure Is Associated with Childhood DNA CpG Methylation*" [12] y Swearingen et al. (2011) "*Application of Beta Regression to Analyze Ischemic Stroke Volume in NINDS rt-PA Clinical Trials*" [13]. En el campo de la ecología como el estudio de Warton & Hui (2011) *"The arcsine is asinine: the analysis of proportions in ecology"* [14]; Peterson & Urquhart (2006) "*Predicting water quality impaired stream segments using landscape-scale data and a regional geostatistical model: A case study in Marylan*" [15]. En el sector del transporte como el estudio de Ferrari & Cribari-Neto (2004) "*Prater's gasoline yield data*" [5] entre otros.

2.2.5 Resultados Obtenidos

El resultado que se obtiene al realizar la estimación de los datos con el modelo de Regresión Beta se caracteriza por ser puntual y pertenecer al intervalo abierto $(0, 1)$ producto del supuesto de distribución beta inicialmente planteado para el modelo, en consecuencia, este valor debe ser transformado a la recta real haciendo uso del inverso de la ecuación utilizada en el tratamiento de datos inicial (ver Ec. 2.3).

2.2.6 VENTAJAS Y DESVENTAJAS

En la literatura estadística, la regresión beta se ha establecido como una técnica poderosa para modelar, sin embargo, a pesar de presentar una serie de beneficios hay aplicaciones en las que la metodología clásica de Regresión Beta todavía tiene una serie de limitaciones, algunas ventajas y desventajas del método son [9]:

- Modelo de Regresión Beta es flexible para proporciones, tasas y concentraciones [16].
- Puede captar sesgo y heteroscedasticidad.
- Debido al diseño, los métodos de inferencia estándar se pueden reutilizar fácilmente.
- Las funciones de ajuste se pueden conectar a filtros más complejos.
- Los parámetros de regresión son interpretables en términos de la media de y y no de $\tilde{y}$.
- Las bases de datos científicas a menudo involucran un gran número de posibles variables predictoras que podrían incluirse en un modelo de regresión. En consecuencia, si se usa la estimación de máxima verosimilitud para ajustarse a un modelo de regresión beta, el modelo puede volverse demasiado complejo y, por lo tanto, sobre ajustar los datos. Esto generalmente conduce a una gran variación y a una alta incertidumbre sobre las relaciones predictor-respuesta.
- Los modelos estadísticos a menudo sufren problemas de multicolinealidad, lo que significa que las variables predictoras están altamente correlacionadas.
- En muchas aplicaciones, las relaciones predictor-respuesta son de naturaleza no lineal esto significa que el predictor lineal $X^T\beta$ del modelo clásico de regresión beta necesita ser reemplazado por una función más flexible que permita una cuantificación adecuada del efecto de predictores no lineales.
- Los modelos de regresión beta clásicos explican convenientemente la sobre dispersión al incluir un parámetro de precisión $\emptyset$. Por otro lado, a menudo se observa que la sobre dispersión depende de los valores de una o más variables predictoras [17]. En el contexto de un modelo de regresión beta, esto implica que $\emptyset$ no es constante, pero debe ser regresado a las variables predictoras este problema hace que la selección de variables sea aún más complicada.

2.3 REGRESIÓN PONDERADA GEOGRÁFICAMENTE

Un método de análisis de datos espaciales recientemente desarrollado es la regresión ponderada geográficamente (GWR, siglas en inglés de "*Geographically Weighted Regression"*) que es útil para modelar procesos espacialmente heterogéneos. GWR consiste en una forma "local" de regresión lineal que está específicamente diseñada para analizar interrelaciones que varían específicamente. El uso de GWR permite explorar si las relaciones entre la variable dependiente y las variables explicativas cambian entre lugares.

2.3.1 ANTECEDENTES

Hay una serie de supuestos subyacentes al modelo de regresión básico descrito (ver Ec. 2.1), uno de los cuales son que las observaciones deben ser independientes entre sí, esto no es siempre el caso con datos para unidades espaciales ya que no solo las variables en el modelo exhiben dependencia espacial, sino que también los residuos del modelo pueden exhibir dependencia. Estas características de los datos espaciales tienen implicaciones para las estimaciones de los parámetros en el modelo básico. Si hay una estructura espacial en los residuos del modelo, esto conducirá a estimaciones ineficientes de los parámetros [18].

La heterogeneidad espacial es otro fenómeno en el modelado espacial. Cuando se ajustan los modelos de regresión básicos, se espera que las relaciones a modelar sean las mismas en todas partes dentro del área de estudio de donde se extraen los datos. Este supuesto se conoce como homogeneidad. Sin embargo, este hecho se cuestionó para el tratamiento de datos espaciales, ya que los procesos que los generan podrían variar en el espacio. Esta condición se conoce como heterogeneidad espacial [18] (ver anexo 8.7).

Para hacer frente a la heterogeneidad espacial Brunsdon y otros en 1996 introdujeron el término GWR para aludir a una familia de modelos de regresión "ajustados al espacio", donde es posible observar las variaciones espaciales de los parámetros estimados y con ello saber dónde y cuánto influye el efecto de una variable explicativa sobre la dependiente.

El modelo GWR trata de ajustar tantas regresiones como observaciones (unidades espaciales) se consideren en el análisis en base al concepto de *distance decay* (se da más peso a las observaciones más próximas y menos a las más lejanas), operacionalizado por medio de una función *Kernel*, que simule el efecto de caída con la distancia. En consecuencia, se pueden realizar estimaciones ajustadas a cada observación, aplicando su correspondiente ecuación [19].

El desarrollo de GWR comenzó a partir de la regresión local y técnicas de suavizado (*smoothing*) y se hizo cada vez más sofisticado, considerando, por ejemplo, la estimación de máxima verisimilitud de los anchos de banda *Kernel Bandwidths*, la autocorrelación espacial entre los residuos, las especificaciones de los métodos lineales generalizados o el modelo bayesiano GWR [20].

2.3.2 DESCRIPCIÓN

GWR busca analizar la no-estacionariedad de los datos. Esto es posible porque una regresión geográficamente ponderada permite la estimación de parámetros locales y no sólo globales. Un estimado local es computado "tomando prestada" información de las unidades dentro de una distancia previamente establecida.

En vez de calibrar una única ecuación, GWR genera una ecuación de regresión por separado para cada observación, estimando un conjunto de coeficientes de regresión para cada observación, por lo tanto, permite a los parámetros variar de forma continua en el espacio geográfico. Para el cálculo de estos coeficientes, GWR calcula en primer lugar un conjunto de pesos que se utilizará para la calibración de la ecuación en cada una de las observaciones. En dicho cálculo de pesos GWR utiliza las funciones que se denominan de ponderaciones (espacial) tipo *kernel*, la cual pondera con mayor peso las observaciones más cercanas que las lejanas [20].

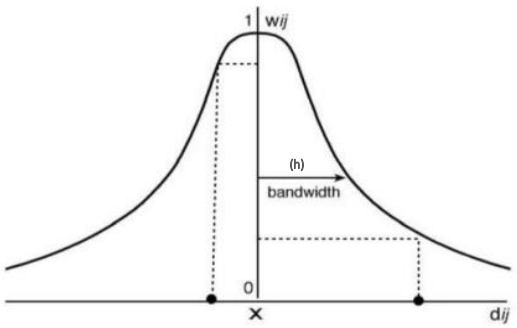


Figura 4: Esquema de función de ponderación típica del modelo GRW.

En la **Figura 4** se visualiza el esquema de una función de ponderaciones típica, donde:

- x: Punto de regresión.
- •: Punto de observación.
- h: Ancho de banda.
- w_{ij} : Peso del punto muestral j en el punto de regresión i.
- d_{ij}: Distancia entre el punto de regresión y el punto muestral.

En este contexto, la función *Kernel* toma como dato de entrada la distancia entre dos localizaciones i y j, posee un ancho de banda ("h", *bandwidth*) que determina el rango espacial que abarca el núcleo (punto central de búsqueda de localizaciones vecinas), devolviendo un peso entre dos localizaciones que es inversamente proporcional a la distancia. Existen dos tipos principales de funciones *Kernel* utilizadas en GWR:

1. **Ancho de banda fijo** (*Fixed kernel*): El número de observaciones puede variar para calibrar cada ecuación, pero el área de búsqueda permanece constante (ver **FIGURA 5**). Suele ser un método más apropiado cuando la distribución de sus observaciones es relativamente estable a través del espacio (por ejemplo, tamaño, número de vecinos). La fórmula utilizada es:

$$W_{ij} = \exp\left[-\frac{1}{2}\left(\frac{d_{ij}}{h}\right)^2\right] \qquad \text{Ec. 2.18}$$

Donde, W_{ij}, es el peso de la observación j respecto a la observación i, este peso cambia en función de la distancia d_{ij} y h es el ancho de banda.

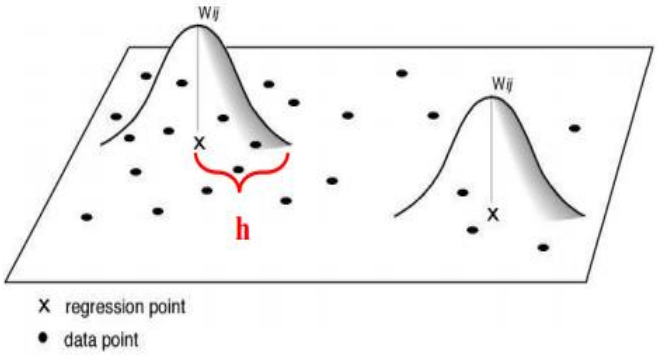


Figura 5: Esquema de búsqueda de vecindad de la función kernel con ancho de banda fijo.

2. **Ancho de banda variable** (*adaptative distance kernel*): En este caso, el número de observaciones permanece fijo mientras que el área de búsqueda varía (ver **FIGURA 6**). Suele ser una técnica más apropiada cuando la distribución varia a través del espacio (por ejemplo, los eventos que se agrupan o polígonos que son heterogéneos). La fórmula utilizada es:

$$W_{ij} = 1 - \left(\frac{d_{ij}}{h}\right)^2 \qquad \text{Ec. 2.19}$$

Donde, W_{ij}, es el peso de la observación j respecto a la observación i, este peso cambia en función de la distancia d_{ij} y h es el ancho de banda.

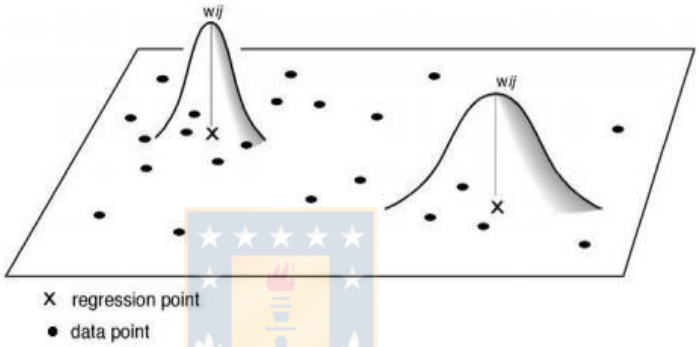


FIGURA 6: Esquema de búsqueda de vecindad de la función kernel con ancho de banda variable.

El ancho de banda de kernel se expresa en las mismas unidades que las coordenadas utilizadas en el conjunto de datos. A medida que el ancho de banda se hace más grande se va acercando a la unidad y el modelo local de GWR se acerca cada vez más al modelo global de mínimos cuadrados ordinarios [18] (ver anexo 8.8).

En ambos tipos de funciones de *kernel* existe un parámetro desconocido que es el ancho de banda kernel el cual debe ser definido o estimado a partir de las observaciones disponibles. Actualmente existen tres métodos distintos para calcularlo: Asignación directa del ancho de banda del número de vecinos más cercanos, validación cruzada (CV) y el Criterio de información corregido de Akaike (AICc), los cuales son descritos en el título 2.3.3.2.

El AICc proporciona una medida de la distancia de información entre el modelo que se ha ajustado y el modelo "verdadero" desconocido. Esta distancia no es una medida absoluta, sino una medida relativa conocida como la distancia de información Kullback-Leibler (ver anexo 8.10).

Se considera que dos modelos separados que se comparan son equivalentes si la diferencia entre los dos valores de AICc es menor que 3. Esta es una regla general ampliamente aceptada, sin embargo, también se podría usar 4 en su lugar. Como el AICc es una medida relativa, los valores reales que se informan en la salida de GWR pueden ser contra intuitivamente grandes o pequeños, esto no importa, ya que son las diferencias en los valores de AICc importantes [18].

2.3.3 ALGORITMO

Motivado por la idea de los métodos de regresión no paramétricos, Brunsdon et al (1996; 1998) propusieron GWR para explorar la no estacionariedad espacial de una relación de regresión para datos espaciales ajustando localmente un coeficiente espacial variable. El modelo es de la forma:

$$y_i = \sum_{j=1}^{p} \beta_i(u_i, v_i) x_{ij} + \varepsilon_i \,, \quad i = 1, 2, \dots, n \qquad \text{Ec. 2.20}$$

Donde, $(y_i; x_{i1}, \dots, x_{i1})$ son observaciones de la respuesta y y las variables explicativas $x_1, \dots, x_p$ en la ubicación (u_i, v_i) en la región geográfica estudiada, $\beta_i(u_i, v_i)(j = 1, 2, \dots, p)$ son p funciones desconocidad de ubicaciones geográficas y $\varepsilon_i(i = 1, 2, \dots, n)$ son términos de error con media cero y varianza común σ^2. $\beta_i(u_i, v_i)(j = 1, 2, \dots, p)$ se estiman localmente en cada ubicación (u_i, v_i) mediante el procedimiento de mínimos cuadrados ponderados (OLS) (ver anexo 8.9) en el que se utilizan algunos pesos de decaimiento de distancia.

Cada conjunto de los coeficientes estimados en n ubicaciones puede producir un mapa de variación que puede proporcionar información útil sobre la no-estacionalidad de la relación de regresión. La técnica GWR tiene un gran atractivo en el análisis de datos espaciales ya que ha aplicado con éxito a muchos problemas prácticos [21].

2.3.3.1 MÉTODO DE AJUSTE

Los parámetros en el modelo GWR se estiman localmente mediante el enfoque de mínimos cuadrados ponderados. Los pesos en cada ubicación (u_i, v_i) se toman en función de la distancia desde (u_i, v_i) a otras ubicaciones donde se recopilan las observaciones. Suponga que los pesos en la ubicación (u_i, v_i) son $w_j(u_i, v_i), j = 1,2,..,n$. Luego, los parámetros en la ubicación (u_i, v_i) se estima minimizando:

$$\sum_{j=1}^{n} w_j(u_i, v_i)\left[y_j - \beta_1(u_i, v_i)x_{j1}\beta_2(u_i, v_i)x_{j2} - \cdots - \beta_p(u_i, v_i)x_{jp}\right]^2 \quad \text{Ec. 2.21}$$

Donde,

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}, Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \quad \text{Ec. 2.22}$$

$$W(u_i, v_i) = \begin{pmatrix} w_1(u_i, v_i) & 0 & \cdots & 0 \\ 0 & w_2(u_i, v_i) & 0 & 0 \\ 0 & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & w_3(u_i, v_i) \end{pmatrix} \quad \text{Ec. 2.23}$$

$$W(u_i, v_i) = Diag[w_1(u_i, v_i), w_2(u_i, v_i), \dots, w_n(u_i, v_i)]$$

Luego, de acuerdo con la teoría OLS, los parámetros estimados (u_i, v_i) son:

$$\hat{\beta}(u_i, v_i) = [X^T W(u_i, v_i)X]^{-1} X^T W(u_i, v_i)Y \quad \text{Ec. 2.24}$$

Donde $x_i^T = (x_{i1}, x_{i2}, \dots, x_{ip})$ es la i-esima fila de X. Entonces el valor ajustado de y en (u_i, v_i) se obtiene:

$$\hat{y}_i = x_i^T \hat{\beta}(u_i, v_i) = x_i^T [X^T W(u_i, v_i)X]^{-1} X^T W(u_i, v_i)Y \quad \text{Ec. 2.25}$$

Denotado respectivamente por $\hat{Y} = (\hat{y}_1, \hat{y}_2, \dots, \hat{y}_n)^T$ y $\hat{\varepsilon} = (\hat{\varepsilon}_1, \hat{\varepsilon}_2, \dots, \hat{\varepsilon}_n)^T$ el valor de valores ajustados de y y el vector de residuos en n ubicaciones $(u_i, v_i), i = 1, 2, \dots, n$.

Entonces,

$$\begin{cases} \hat{Y} = LY; \\ \hat{\varepsilon} = Y - \hat{Y} = (I - L)Y, \end{cases} \quad \text{Ec. 2.26}$$

Donde,

$$L = \begin{pmatrix} x_1^T [X^T W(u_1, v_1)X]^{-1} X^T W(u_1, v_1) \\ x_2^T [X^T W(u_2, v_2)X]^{-1} X^T W(u_2, v_2) \\ \vdots \\ x_n^T [X^T W(u_n, v_n)X]^{-1} X^T W(u_n, v_n) \end{pmatrix} \quad \text{Ec. 2.27}$$

Es una matriz de $n\ x\ n$ e I es la matriz identidad de orden n.

2.3.3.2 ELECCIÓN DE LOS PESOS

La función de los pesos es proporcionar énfasis diferentes en observaciones diferentes para generar los parámetros estimados. Cuando se trabajan con datos que son espaciales, las observaciones cercanas a una ubicación (u_i, v_i) generalmente ejercen más influencia sobre las estimaciones de los parámetros en la ubicación (u_i, v_i) que aquellas que se encuentran más alejadas. Al igual que los pesos en la regresión no paramétrica, una opción es [21]:

$$w_j(u_i, v_i) = \exp\left[-\left(\frac{d_{ij}}{h}\right)^2\right], \quad j = 1, 2, \dots, n \qquad \text{Ec. 2.28}$$

Donde d_{ij}[4] es la distancia desde la ubicación (u_i, v_i) a (u_j, v_j) y h se llama ancho de banda (*Bandwidth*).

Otra elección de los pesos es la siguiente.

$$w_j(u_i, v_i) = \begin{cases} \left[1 - \left(\frac{d_{ij}}{h}\right)^2\right]^2, & si\ d_{ij} \le h \\ 0, & si\ d_{ij} > h \end{cases}, \quad j = 1, 2, \dots, n. \qquad \text{Ec. 2.29}$$

El ancho de banda h puede determinarse mediante el procedimiento de validación cruzada. Entonces:

$$\Delta(h) = \sum_{i=1}^{n} \left(y_i - \hat{y}_{(i)}(h)\right)^2 \qquad \text{Ec. 2.30}$$

Donde $\hat{y}_{(i)}(h)$ es el valor ajustado de y_i de la observación en la ubicación (u_i, v_i) omitida del proceso de ajuste. Se elije h_0 como valor deseable de ancho de banda de modo que:

$$\Delta(h) = min\Delta(h_0) \qquad \text{Ec. 2.31}$$

El ancho de banda h puede determinarse también mediante el criterio de información corregido de Akaike AIC_c. En un caso general, la AIC es:

$$AIC = 2k - 2\ln(L) \qquad \text{Ec. 2.32}$$

Donde k es el número de parámetros en el modelo estadístico, y L es el máximo valor de la función de verosimilitud para el modelo estimado.

[4] Son las distancias euclidianas y distancias *Geate Circle* cuando se usan coordenadas esféricas, sin embargo, de igual manera pueden utilizarse distancias no euclidianas.

En el caso concreto del modelo GWR, el ancho de banda podría ser optimizado utilizando un algoritmo que busca minimizar el valor del AICc dado por:

$$AIC_c = 2n\ log_e(\hat{\sigma}) + n\ log_e(2\pi) + n\frac{n + trance(H)}{n - 2 - trance(H)} \quad \text{Ec. 2.33}$$

Donde $\hat{\sigma}$ es la estimación de la desviación estándar del error, H es la matriz gorro (*Hat Matrix*[5]) y $trance(H)$ es la suma de los elementos diagonales de la matriz, Cada fila de la matriz gorro viene dada por $H_i = X_i\left(X^T W^i X\right)X^T W_i$ [20].

2.3.4 CAMPO DE APLICACIÓN

Aunque la GWR es una técnica muy joven, ha sido aplicada en campos de investigación muy distintos, por ejemplo, en el transporte, como el trabajo de Clark, *"Estimating local car ownership models. Journal of Transport Geography"* [22]; en el mapeo estudios como el de Dong Song, "Mapping soil organic carbon content by Geographically Weighted regression" [23]. Por otro lado, también este modelo ha sido útil para analizar la geografía de la pobreza por Sánchez, "Métodos para el análisis espacial. Una aplicación al estudio de la geografía de la pobreza" [24] y estimación de precios de vivienda como el estudio de Duque, "Infraestructura pública y precios de vivienda" [25].

2.3.5 RESULTADOS OBTENIDOS

Como mínimo, GWR producirá estimaciones de parámetros y sus errores estándar asociados en los puntos de regresión. Si los puntos de regresión son los mismos que los puntos de muestra, GWR producirá predicciones para la variable dependiente (valores ajustados), los residuos y los residuos estandarizados. Algunas implementaciones también generarán valores locales de r^2 e influirán en las estadísticas basadas en la matriz gorro.

Si los puntos de regresión no son los mismos que los puntos de muestra, y no hay variables independientes disponibles para los puntos de regresión, entonces habrá poco más que estimaciones de parámetros y errores estándar disponibles: valores ajustados, residuos y una matriz de gorros no estarán disponible. Si hay variables independientes disponibles, los valores ajustados estarán disponibles; si también hay una variable dependiente presente, entonces se puede crear todo el rango de salidas [4].

[5] La matriz que tiene esta forma $\mathrm{A(ATA)} - \mathrm{1AT}$ se suele presentar en estadísticas como matriz gorro.

2.3.6 VENTAJAS Y DESVENTAJAS

Si bien GWR se ha establecido como un método eficaz producto que permite mejorar los ajustes y neutralizar la dependencia espacial en los residuos también presenta una serie de otras ventajas y desventajas, por ejemplo:

- Permite moverse desde una perspectiva global a un análisis local del problema, obteniendo un mayor grado de detalle y precisión.
- Los coeficientes de cada uno de los predictores (elasticidades) varían de una unidad espacial a otra (inestabilidad espacial).
- La posibilidad de estimar coeficientes de determinación locales para cada unidad espacial a partir de los valores de un conjunto de observaciones vecinas permite conocer la forma en que se combinan localmente las variables de la regresión para obtener el "ajuste específico" en una localización.
- La desagregación del coeficiente de determinación (R^2) global en coeficientes locales y el análisis de su distribución geográfica permiten reconocer dónde las variables independientes tienen un mayor o peor poder explicativo.
- En la gran mayoría de los casos, esta clase de regresión genera errores de estimación más pequeños que el modelo tradicional, además de anular o reducir el problema de la autocorrelación espacial.
- Es posible generar superficies interpoladas para conocer la distribución espacial continua de los parámetros y aplicar los principios de la "predicción espacial" para hallar los valores de las observaciones que faltan.
- Facilita explorar la estructura espacial del modelo, es decir, medir el grado de dependencia espacial presente en el modelo, pudiendo ser positiva o negativa, o detectar clústeres de datos.
- Esta técnica cuestiona el supuesto implícito en las regresiones estándares de que un modelo explicativo puede aplicarse por igual a toda el área geográfica analizada, cuando en realidad puede haber importantes variaciones tanto en el modelo completo como en la relación específica entre la variable dependiente y una de sus variables explicativas.

2.4 KRIGING ORDINARIO

Kriging es la técnica de interpolación avanzada utilizada en geoestadística donde el resultado final del Kriging es un mapa con los valores interpolados de la variable. Cada interpolación lleva asociado un grado de incertidumbre que puede ser representado en el espacio (en forma de varianza o desviación estándar) de esta manera el resultado del Kriging representa la probabilidad de que la variable alcance un determinado valor [26].

Los métodos Kriging se aplican con frecuencia con el propósito de predicción, sin embargo, estas metodologías tienen diversas aplicaciones, dentro de las cuales se destacan la simulación y el diseño de redes óptimas de muestreo [27].

2.4.1 ANTECEDENTES

El Kriging es un método estadístico desarrollado en la década de los sesenta por el ingeniero de minas Daniel G. Krige y formalizado por Georges Matheron. Daniel G. Krige propuso una primera aproximación a través de una variante del método de medias móviles, que puede considerarse como el equivalente al método conocido como Kriging Simple, que es uno de los métodos de estimación lineal espacial con mayores cualidades teóricas [28].

El método de Krige es denominado como el mejor estimador lineal insesgado, lineal porque el método está representado como la combinación lineal de las observaciones por los ponderadores, e insesgado porque la media de los errores es igual cero. Para la implementación de este método, no es necesario contar con datos que provengan de una distribución normal, pero un supuesto que sí se debe cumplir es que exista correlación espacial entre las coordenadas vecinas.

Kriging utiliza el modelo de covarianza espacial para estimar las variables desconocidas en una nueva ubicación basada en sus vecinos y considera la función de covarianza entre los datos conocidos y desconocidos. Los métodos de Kriging son algoritmos ampliamente utilizados para la estimación de datos espaciales, ya que adopta la función de covarianza basada en la distancia de dos puntos para obtener los coeficientes lineales y predecir valores en nuevas ubicaciones. Actualmente existe una amplia gama de variantes del método de Kriging dependiendo de la información que se posea y el tipo de predictor (ver Tabla 1).

Tabla 1: Tipos de predictores Kriging.

Tipo de Predictor	Nombre
Lineal	Simple Ordinario Universal
No lineal	Indicador Probabilístico Log-Normal, Multi-Gaussiano Disyuntivo

2.4.2 CONSTRUCCIÓN DEL KRIGING

La resolución de un problema de estimación por Kriging se articula siempre en torno a las mismas etapas. Las diferentes variantes sólo radican en las hipótesis realizadas sobre la función aleatoria $Z = \{Z(x) \in D\}$ que representa la variable regionalizada.

2.4.2.1 RESTRICCIÓN DE LINEALIDAD

El estimador tiene que ser una combinación lineal ponderada (promedio ponderado) de los datos, que se denota de la siguiente forma:

$$Z^*(x_0) = a + \sum_{i=1}^{n} \lambda_i Z(x_i) \qquad \text{Ec. 2.34}$$

donde x_0 es la posición donde se busca tener una estimación, $\{x_\alpha,\ i = 1, \dots, n\}$ son las posiciones con datos, mientras que los ponderadores $\{\lambda_i,\ i = 1, \dots, n\ \}$ y el coeficiente a son las incógnitas del problema de Kriging. En rigor, se debería escribir $\lambda_i(x_0)$ en lugar de λ_i, puesto que los ponderadores de Kriging dependerán de la posición x_0 a estimar.

Esta restricción se debe a la decisión de considerar sólo los primeros momentos de las distribuciones de probabilidad (esperanza y covarianza / variograma). La construcción de estimadores más sofisticados que no sean combinaciones lineales de los datos, requeriría la especificación de la distribución espacial de la función aleatoria más allá de su variograma. Esto se puede realizar con métodos de geoestadística no lineal [29].

El estimador $Z^*(x_0)$ así definido es una combinación de variables aleatorias y, por lo tanto, es una cantidad aleatoria. Para obtener una estimación numérica, basta con aplicar la fórmula anterior a la realización particular que constituyen los datos experimentales [29]:

$$z^*(x_0) = a + \sum_{i=1}^{n} \lambda_i z(x_i) \qquad \text{Ec. 2.35}$$

2.4.2.2 RESTRICCIÓN DE INSESGO

Esta etapa consiste en expresar que el error de estimación tiene esperanza nula, es decir:

$$E[Z^*(x_0) - Z(x_0)] = 0 \qquad \text{Ec. 2.36}$$

Se puede interpretar esta restricción, reemplazando la esperanza matemática por una media en el espacio si se calcula sobre numerosas configuraciones de Kriging idénticas, la media de los errores de estimación cometidos se acerca a cero. La ausencia de sesgo no garantiza que los errores sean bajos, sino sólo que su media global es aproximadamente nula [29].

2.4.2.3 RESTRICCIÓN DE OPTIMALIDAD

Al superar las etapas anteriores, el estimador está sometido a una o varias restricciones, pero no está totalmente especificado. La última etapa consiste en buscar los ponderadores que minimizan la varianza del error de estimación:

$$VAR[Z^*(x_0) - Z(x_0)]\ es\ minima$$

En términos intuitivos, esta restricción significa que, si se calcula sobre numerosas configuraciones de Kriging idénticas, la varianza estadística de los errores de estimación cometidos es la más baja posible. Este criterio de precisión equivale a la minimización del error cuadrático promedio.

2.4.3 CONCEPTOS PARA ANÁLISIS VARIOGRÁFICO

Los valores de una variable regionalizada no son independientes. En la interpretación probabilística de la variable regionalizada, esta noción intuitiva de dependencia está descrita por la distribución espacial de la función aleatoria, que modela la manera como se relacionan los valores observados en distintos sitios por una distribución de probabilidad multivariable.

2.4.3.1 VARIOGRAMA EXPERIMENTAL

El variograma experimental de las coordenadas observadas, representa el proceso de dependencia espacial entre las coordenadas, a mayor distancia mayor semivarianza. Es calculado en base a la distancia en a la que se encuentran las coordenadas geográficas según los criterios de tolerancia [30].

Se considera una variable regionalizada z conocida en n posiciones $\{x_1, \dots, x_n\}$. El estimador tradicional del variograma para un vector de separación h dado, se define de la siguiente manera:

$$\hat{\gamma}(h) = \frac{1}{2|N(h)|} \sum_{N(h)} \left[Z(x_\alpha) - Z(x_\beta)\right]^2 \qquad \text{Ec. 2.37}$$

Donde,

- $N(h)$: (α, β) tal que $x_\alpha - x_\beta = h$
- $|N(h)|$: Es el número de pares contenidos en el conjunto $N(h)$

Se ve que el estimador anterior consiste en reemplazar la esperanza en la expresión del variograma teórico (ver Ec. 2.38) Por la media aritmética sobre los pares de datos separados por el vector h.

$$\gamma(h) = \frac{1}{2} E\{[z(x+h) - z(x)]^2\} \qquad \text{Ec. 2.38}$$

El estimador así definido lleva el nombre de variograma experimental. No se trata de una función propiamente tal, sino que, de una serie de valores, pues sólo se puede calcular para vectores h tales que $N(h)$ no es vacío.

El variograma experimental para un vector h puede interpretarse como el momento de inercia de la nube de correlación diferida (ver anexo 8.14) que mide la distancia cuadrática promedio entre los puntos de la nube y la línea diagonal. Mientras más densa es la nube de correlación diferida en torno a la diagonal, más pequeña su inercia [29].

2.4.3.2 TOLERANCIAS EN LOS PARÁMETROS DE CÁLCULO

Si los datos están irregularmente distribuidos en el campo D, el número de pares $|N(h)|$ que interviene en el cálculo de $\hat{\gamma}(h)$ para un vector h dado, es generalmente muy pequeño (incluso igual a 0 ó 1). El variograma experimental tiene entonces un aspecto muy errático y resulta imposible interpretarlo y modelarlo. Para que sea más robusto, se suele permitir algunas tolerancias de cálculo, sobre las distancias y las direcciones:

$$\hat{\gamma}^{+}(h) = \frac{1}{2|N^{+}(h)|} \sum_{N^{+}(h)} \left[Z(x_\alpha) - Z(x_\beta))\right]^2 \qquad \text{Ec. 2.39}$$

Donde, $N^{+}(h)$: $\{(\alpha, \beta)\text{tal que } x_\alpha - x_\beta \ \in T(h)\} = \ \cup_{h' \in T(h)} N(h')$

$T(h)$ es una región de tolerancia alrededor de h, de la forma $[h - \Delta h, h + \Delta h]$ en el caso unidimensional. En el caso bi o tridimensional, existen tolerancias tanto sobre la longitud de h como sobre su orientación, tal como se ilustra en la Figura 7.

El ancho de banda limita la separación del cono de tolerancia a una extensión máxima. En el espacio de tres dimensiones, se introduce dos anchos de banda: uno horizontal y otro vertical.

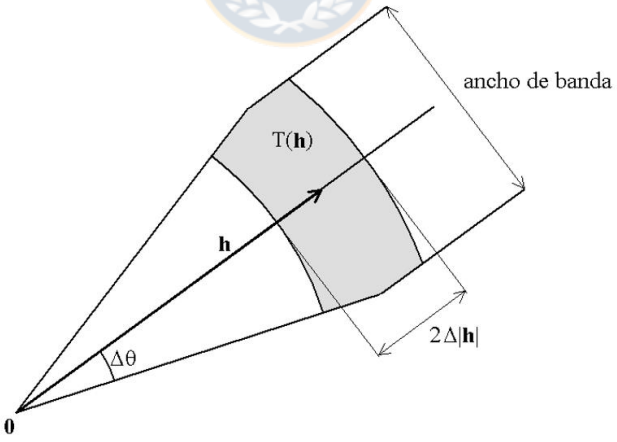


Figura 7: Región de tolerancia T(h) alrededor del vector h (caso bidimensional)

2.4.3.3 PROPIEDADES DEL VARIOGRAMA EXPERIMENTAL

El variograma experimental $\hat{\gamma}(h)$ es un estimador insesgado del variograma teórico:

$$E[\hat{\gamma}(h)] = \hat{\gamma}(h) \qquad \text{Ec. 2.40}$$

Un indicador de la robustez de $\hat{\gamma}(h)$ es su varianza relativa:

$$\frac{Var[\hat{\gamma}(h)]}{[\hat{\gamma}(h)]^2} \qquad \text{Ec. 2.41}$$

Mientras más elevada dicha varianza, más susceptible es el variograma experimental de fluctuar en torno a su valor esperado (el variograma teórico $\gamma(h)$) y más difícil se vuelve la inferencia estadística [29].

2.4.3.4 PARÁMETROS PARA CONSTRUCCIÓN DE VARIOGRAMA

En la Tabla 2 se muestran los parámetros con los cuales son necesario para la construcción de un semivariograma, la Figura 8 muestra un ejemplo de un semivariograma y la Tabla 3 muestra algunos modelos de semivariogramas.

Tabla 2: Parámetros de construcción para un semivariograma.

Parámetro	Definición
Pepita	Definido por el símbolo C_0, también conocida como efecto pepita, se lo atribuye al error de medición de una muestra de coordenadas.
Meseta	Definido por el símbolo $(C_0 + C)$. Es el valor en donde la distancia se vuelve contante, también es conocido como la varianza asintótica del proceso
Rango	Definido por el símbolo a, y representa la distancia por el cual los valores de la variable dejan de estar correlacionadas.

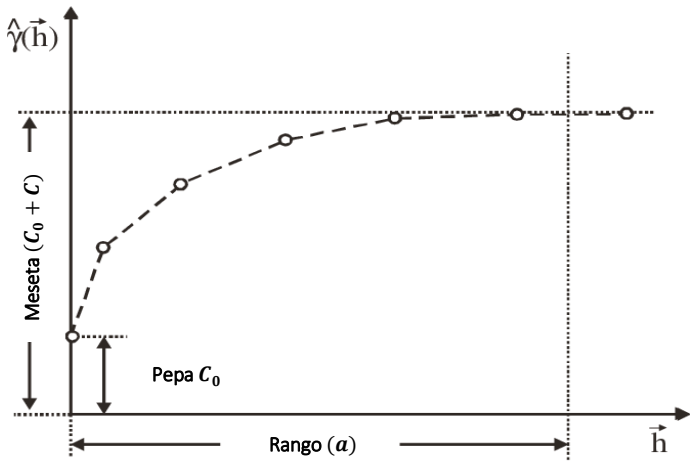


Figura 8: Semivariograma con sus parámetros de construcción

Tabla 3: Semivariogramas teóricos.

Modelo	$\boldsymbol{\gamma(h)}$	Rango
Esférico	$C_0 + C\left[1.5\left(\frac{h}{a}\right) + 0.5\left(\frac{h}{a}\right)3\right]$	$0 \leq h \leq a$ $h > a$
Exponencial	$C_0 + C\left[1 - e^{-\left(\frac{h}{a}\right)}\right]$	$h => 0$
Circular	$C_0 + C\left[\frac{2h}{\pi a}\sqrt{1 - \left(\frac{h}{a}\right)^2} + \frac{2}{\pi} arcsin\left(\frac{h}{a}\right)\right]$	$0 \leq h \leq a$ $h > a$
Gaussiano	$C_0 + C\left[1 - e^{-\left(\frac{h}{a}\right)^2}\right]$	$h => 0$
logarítmico	$C_0 + C\left[Log\left(\frac{h}{a}\right)\right]$	$h > 0$

2.4.4 ALGORITMO

Suponga que se hacen mediciones de la variable de interés Z en los puntos x_i, $i = 1,2, \dots, n$, de la región de estudio, es decir se tienen realizaciones de las variables $Z(x_1), \dots, Z(x_n)$, y se desea predecir $Z(x_0)$, en el punto x_0 donde no hubo medición. En esta circunstancia, el método Kriging Ordinario propone que el valor de la variable puede predecirse como una combinación lineal de las n variables aleatorias así:

$$Z^*(x_0) = \lambda_1 Z(x_1) + \lambda_2 Z(x_2) + \cdots + \lambda_n Z(x_n) = \sum_{i=1}^{n} \lambda_i Z(x_i) \qquad \text{Ec. 2.42}$$

Donde, λ_i representa los pesos, ponderaciones de los valores originales. Dichos pesos se calculan en función de la distancia entre los puntos muestreados y el punto donde se va a hacer la correspondiente predicción. La suma de los pesos debe ser igual a uno para que la esperanza del predictor sea igual a la esperanza de la variable. Esto último se conoce como el requisito de insesgo y es mostrada en Ec. 2.36.

Se asume que el proceso es estacionario de media m (desconocida) y utilizan las propiedades del valor esperado, se demuestra que la suma de las ponderaciones debe ser igual a uno:

$$E\left(\sum_{i=1}^{n} \lambda_i \, Z(x_i)\right) = m \qquad \text{Ec. 2.43}$$

$$\sum_{i=1}^{n} \lambda_i E\big(Z(x_i)\big) = m \qquad \text{Ec. 2.44}$$

$$\sum_{i=1}^{n} \lambda_i \, m = m \qquad \text{Ec. 2.45}$$

$$m \sum_{i=1}^{n} \lambda_i \; = m \; \Rightarrow \sum_{i=1}^{n} \lambda_i \; = 1 \qquad \text{Ec. 2.46}$$

Se dice que $Z^*(x_0)$ es el mejor predictor, lineal en este caso, porque los pesos se obtienen de tal manera que minimicen la varianza del error de predicción, es decir que minimicen la expresión:

$$V[Z^*(x_0) - \mathrm{Z}(x_0)]$$

Sujeto a,

$$\sum_{i=1}^{n} \lambda_i \; = 1$$

Se tiene,

$$V[Z^*(x_0) - \mathrm{Z}(x_0)] = V[Z^*(x_0)] - 2COV[Z^*(x_0), \mathrm{Z}(x_0)] + V[\mathrm{Z}(x_0)] \qquad \text{Ec. 2.47}$$

Desagregando las componentes de la ecuación anterior se obtiene lo siguiente:

$$V[Z^*(x_0)] = V\left[\sum_{i=1}^{n} \lambda_i \, \mathrm{Z}(x_i)\right] = \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j COV\big[\mathrm{Z}(x_i), \mathrm{Z}\big(x_j\big)\big] \qquad \text{Ec. 2.48}$$

En adelante se utiliza,

$$\mathrm{C}_{ij} = COV\big[\mathrm{Z}(x_i), \mathrm{Z}\big(x_j\big)\big] \qquad \text{Ec. 2.49}$$

$$V[Z(x_0)] = \sigma^2 \qquad \text{Ec. 2.50}$$

Entonces,

$$V[Z^*(x_0) - \mathrm{Z}(x_0)] = \sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j\mathrm{C}_{ij} - 2\sum_{i=1}^{n}\lambda_i\mathrm{C}_{i0} + \sigma^2 \qquad \text{Ec. 2.51}$$

Esta ecuación por minimizar se resuelve mediante el método de multiplicadores de Lagrange

$$\sigma_k^2 = \sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j\mathrm{C}_{ij} - 2\sum_{i=1}^{n}\lambda_i\mathrm{C}_{i0} + \sigma^2 + 2\mu\left(\sum_{j=1}^{n}\lambda_i - 1\right) \qquad \text{Ec. 2.52}$$

Donde,

- μ: Multiplicador de Lagrange

Obteniendo los valores extremos de una función, derivando e igualando a cero, en este caso respecto a λ_n :

$$\frac{\partial\sigma_k^2}{\partial\lambda_n} = 2\sum_{j=1}^{n}\lambda_j\mathrm{C}_{nj} - 2\mathrm{C}_{n0} + 2\mu = 0 \qquad \text{Ec. 2.53}$$

$$\sum_{j=1}^{n}\lambda_j\mathrm{C}_{nj} + \mu = \mathrm{C}_{n0} \qquad \text{Ec. 2.54}$$

Luego se deriva respeto a μ:

$$\frac{\partial\sigma_k^2}{\partial\mu} = 2\sum_{i=1}^{n}\lambda_i - 2 = 0 \qquad \text{Ec. 2.55}$$

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

De los resultados anteriores resulta un sistema de $(n-1)$ ecuaciones con $(n+1)$ incógnitas, que matricialmente puede ser escrito como:

$$\begin{pmatrix} \mathrm{C}_{11} & \cdots & \mathrm{C}_{1n} & 1 \\ \vdots & \ddots & \vdots & \vdots \\ \mathrm{C}_{11} & \cdots & \mathrm{C}_{nn} & 1 \\ 1 & \cdots & 1 & 0 \end{pmatrix}\begin{pmatrix} \lambda_i \\ \vdots \\ \lambda_n \\ \mu \end{pmatrix} = \begin{pmatrix} \mathrm{C}_{10} \\ \vdots \\ \mathrm{C}_{n0} \\ 1 \end{pmatrix} \qquad \text{Ec. 2.56}$$

$$\mathrm{C}_{ij} \bullet \lambda = \mathrm{C}_{i0} \qquad \text{Ec. 2.57}$$

Por lo cual los pesos que minimizan el error de predicción se determinan mediante la función de covariograma a través de:

$$\lambda = {\mathrm{C}_{ij}}^{-1} \bullet \mathrm{C}_{i0} \qquad \text{Ec. 2.58}$$

Encontrando los pesos se calcula la predicción en el punto x_0 y de forma análoga se procede para cada punto donde se quiera hacer predicción.

2.4.5 Varianza de predicción del Kriging Ordinario

Multiplicando Ec. 52 por λ_i se obtiene:

$$\lambda_j \left(\sum_{j=1}^{n} \lambda_j \mathrm{C}_{ij} + \mu \right) = \lambda_i \mathrm{C}_{i0}, \qquad \forall\, i,j = 1,2,\dots,n. \qquad \text{Ec. 2.59}$$

Sumando las n ecuaciones,

$$\sum_{i=1}^{n} \lambda_i \sum_{j=1}^{n} \lambda_j \mathrm{C}_{ij} + \sum_{j=1}^{n} \lambda_j \, \mu = \sum_{i=1}^{n} \lambda_i \mathrm{C}_{i0} \qquad \text{Ec. 2.60}$$

Sustituyendo Ec. 2.51 se obtiene:

$$\sigma_k^2 = \sigma^2 - \sum_{i=1}^{n} \lambda_i \mathrm{C}_{i0} - \mu \qquad \text{Ec. 2.61}$$

Los pesos λ pueden ser estimados a través de la función de semivarianza, para lo cual se requiere conocer la relación entre las funciones de covariograma y de semivarianza (ver anexo 8.12).

2.5 Kriging Bayesiano

La regresión Kriging bayesiano (BKR, siglas en inglés de "*Bayesian Kriging Regression*") es un modelo estadístico de Kriging con una suposición gaussiana en el marco Bayesiano. Dado que el marco Bayesiano considera los parámetros del modelo como variables aleatorias, BKR tiene la capacidad de cuantificar las incertidumbres asociadas con los parámetros del modelo.

2.5.1 DESCRIPCIÓN

El Kriging Bayesiano en comparación con los métodos de Kriging convencionales (Kriging simple, ordinario y universal) se lleva a cabo mediante la integración del conocimiento previo (denominado como "*prior*") sobre la observación para ser considerada como una suposición calificada en el procedimiento de estimación. La suposición calificada permite obtener una observación más realista bajo la evaluación de la incertidumbre.

El Kriging Bayesiano ajusta el semivariograma experimental de modo automático mediante simulaciones, tomando como valor inicial los parámetros de entrada que son probabilidades a priori. Además, asume implícitamente que la semivarianza experimental es el verdadero semivariograma para las zonas a pronosticar. La incertidumbre de las observaciones se representa como una distribución posterior para anular pequeñas regiones poco realistas dentro de las observaciones; la distribución previa se incorpora a través del Teorema de Bayes (ver anexo 8.15) para evaluar la incertidumbre de los parámetros (pepa, meseta y rango) en la función de covarianza.

Este tipo de Kriging es similar al Kriging Universal (ver anexo 8.13), pero existe una suposición adicional de que existen conocimientos previos sobre el valor de los coeficientes en la tendencia, además si la varianza en el Kriging Bayesiano es cero, el modelo es idéntico al Kriging Simple. Por otro lado, este método es estable para cualquier número de coeficientes y datos esto significa que se puede usar un modelo de tendencia flexible, aunque existen pocos datos de sondajes [30].

2.5.1.1 MARCO BAYESIANO

En la práctica, los parámetros utilizados son desconocidos, en la inferencia bayesiana utiliza a estos parámetros como variables aleatorias. Basándose en esta incertidumbre, con simulaciones, se ajustan mejor estos paramentos, que luego serán usados en las predicciones. Considerando un vector aleatorio Y con probabilidad de distribución dado por la función $pr\left(\frac{\gamma}{\vartheta}\right)$. Fijado por un vector ϑ. Considerando que de una muestra $Y = \gamma$ puede ser escrita como $L(\gamma|\vartheta) \equiv pr(\gamma|\vartheta)$, L es una función del parámetro ϑ y se llama, función de verosimilitud.

Considerando la distribución de Y como:

$$(Y|\beta, \sigma^2, \emptyset, \tau^2) \sim N\left(X\beta; \tau^2 I + V_\gamma(\sigma^2, \emptyset)\right) \qquad \text{Ec. 2.62}$$

La verosimilitud es una función de $\vartheta = (\beta, \sigma^2, \emptyset, \tau^2)$

$$L(\vartheta|\gamma) = \left|\tau^2 I + V_\gamma(\sigma^2, \emptyset)\right|^{-\frac{1}{2}} exp\left\{-\frac{1}{2}(\gamma - X\beta)'(\tau^2 I + V_\gamma(\sigma^2, \emptyset))^{-1}(\gamma - X\beta)\right\} \quad \text{Ec. 2.63}$$

La estadística bayesiana, el parámetro ϑ y la variable Y son cantidades aleatorias distribución de probabilidad conjunta (ver anexo 8.16) $pr(\gamma, \vartheta) = pr\left(\frac{\gamma}{\vartheta}\right) pr(\vartheta)$. La información sobre los parámetros del modelo de los datos es reflejada en la distribución a priori $pr(\vartheta)$.

El Teorema de Bayes combina la información a priori y la probabilidad de tal forma que los conocimientos a priori de los parámetros se van actualizando, después de recoger la muestra, siendo esta de la forma $pr(\vartheta|Y) = pr(\vartheta)pr(\gamma|\vartheta)$.

La distribución $pr(\vartheta|Y)$ es conocida como la distribución a posteriori en base a la inferencia Bayesiana. Para el modelo a posteriori:

$$pr(\beta, \sigma^2, \emptyset, \tau^2|\gamma) = |\tau^2 I + V_\gamma(\sigma^2, \emptyset)|^{-\frac{1}{2}} * exp\left\{-\frac{1}{2}(\gamma - X\beta)'(\tau^2 I + V_\gamma(\sigma^2, \emptyset))^{-1}(\gamma - X\beta)\right\} \quad \text{Ec. 2.64}$$

De las distribuciones a priori que conllevan a posteriori se les llama distribuciones previas conjugadas. Estas distribuciones previas pueden ser ajustadas de manera sencilla con el poder computacional de la actualidad. Existen dos casos extremos para ajustar las distribuciones son:

1. Cuando los parámetros son desconocidos.
2. Cuando el conocimiento de los parámetros es aproximado, los a priori son llamados no informativos, o a priori planos.

Para el último caso se utiliza la estadística bayesiana clásica para ajustarlos.

Las bases para el pronóstico bayesiano son también llamadas como distribución predictiva $pr(\gamma_0|\gamma)$. La distribución predictiva considera a los parámetros desconocidos como el promedio de los parámetros resultantes de la distribución condicional $pr(\gamma_0|\gamma, \vartheta)$, con pesos dados por la distribución a posteriori por los parámetros del modelo $pr(\vartheta|\gamma)$.

$$pr(\gamma_0|\gamma) = \int pr(\gamma_0, \vartheta|\gamma)\, d\vartheta \qquad \text{Ec. 2.65}$$

$$pr(\gamma_0|\gamma) = \int pr(\gamma_0|\gamma, \vartheta) * pr(\vartheta|\gamma) d\vartheta \qquad \text{Ec. 2.66}$$

La distribución predictiva también se puede escribir en base a su función de verosimilitud dada anteriormente.

$$pr(\gamma_0|\gamma) = \int \frac{pr(\gamma_0|\gamma, \vartheta) * pr(\vartheta)}{\int pr(y|\vartheta) pr(\vartheta) d\vartheta} d\vartheta \qquad \text{Ec. 2.67}$$

En la práctica, lo que realiza esta metodología es primero estimar un semivariograma experimental a partir de un subconjunto de las coordenadas observadas, teniendo este semivariograma como base, se simulan los puntos nuevos a predecir dentro del mismo subconjunto, se estiman semivariogramas nuevos a partir del subconjunto pronosticado, este proceso se repite un número determinado de veces, hasta terminar de pronosticar todos los puntos a pronosticar [30].

2.5.2 ALGORITMO

Los principios bayesianos se pueden usar para justificar el modelamiento de un proceso determinista desconocido mediante un proceso aleatorio. El modelo de media no estacionaria:

$$Z(s) = \mu(s) + \delta(s), \qquad s \in D \qquad \text{Ec. 2.68}$$

Donde,

- $\delta(\cdot)$: Proceso aleatorio intrínsecamente estacionario de media cero útil para analizar procesos físicos que son espacialmente heterogéneos.
- $\mu(\cdot)$: Función media que a menudo no se conoce exactamente, por lo que se lleva a la parametrización subsiguiente:

$$\mu(s) = \sum_{j=1}^{p+1} \beta_{j-1}\ f_{j-1},, \qquad s \in D \qquad \text{Ec. 2.69}$$

La cual es una combinación lineal desconocida de funciones conocidas. Se puede asumir que $\beta \equiv (\beta_0, \dots, \beta_p)'$ es un vector de parámetros fijos pero desconocidos. Un enfoque alternativo es expresar la incertidumbre, a gran escala, de manera bayesiana; en otras palabras, suponer que el parámetro $\mu(\cdot)$ es un proceso aleatorio independiente de $\delta(\cdot)$.

Se puede suponer que $\mu(\cdot)$ es (intrínsecamente o de segundo orden) estacionaria o que $\mu(\cdot)$ es un proceso aleatorio general con todos los parámetros de segundo orden conocidos. Omre (1987) toma este último enfoque y utiliza el término Kriging Bayesiano para describir la predicción óptima de $Z(B)$ a partir de datos $Z(S_1), \dots, Z(S_n)$. Las ecuaciones de Kriging se derivan de las expresiones que da para $E\big(Z(S_i)\big)$, $E\big(Z(B)\big)$, $var\left(Z(S_i) - Z\big(S_j\big)\right)$, y $var\left(Z(S_i) - Z\big(S_j\big)\right)$ en términos del variograma de $\delta(\cdot)$, las medias conocidas y la función del variograma de $\mu(\cdot)$. Finalmente, da una estimación del variograma para $\delta(\cdot)$.

El modelo *prior* se puede especializar suponiendo que $\mu(\cdot)$ tiene la expresión lineal de la Ec. 2.69 donde β es un vector aleatorio. En este o en el entorno más general, se puede adoptar un enfoque empírico de Bayes e intentar estimar los parámetros del proceso aleatorio $\mu(\cdot)$ en función de la distribución (marginal) de $Z(S_1), \dots, Z(S_n)$. Estos son sustituidos en el predictor Bayes de $Z(B)$ para producir un predictor empírico Bayes.

Siempre que se pueda encontrar una clase conveniente de *priors*, se puede calcular el predictor óptimo $E(Z(s_0)|Z)$ (en principio) y sería superior al predictor obtenido usando la metodología de Kriging Universal.

Suponga que $f(\cdot \mid \cdot)$ denota una densidad genérica, cuyos argumentos también representan las cantidades aleatorias consideradas. Además, $Z(s_0)$ y Z son aleatorios y que los parámetros θ del modelo son aleatorios con $f(\theta), \theta \in \Theta$ prior.

Entonces, la densidad predictiva es:

$$f(Z(s_0)|Z) = \int_{\Theta} f(Z(s_0)|Z,\theta) f(\theta|Z)\, d\theta \qquad \text{Ec. 2.70}$$

Sin embargo, si las probabilidades posteriores $f(\theta|Z)$ y $f(\theta|Z(s_0), Z)$ son fáciles de calcular, se podría usar una extensión de la fórmula descrita en Besag (1989):

$$f(Z(s_0)|Z) = f(Z(s_0)|Z,\theta)\frac{f(\theta|Z)}{f(\theta|Z(s_0), Z)} \qquad \text{Ec. 2.71}$$

Nótese que el primer término en el lado derecho en la distribución condicional modelada de $Z(s_0)$ dada Z. A pesar de las apariencias, el lado derecho no depende de θ, sino solo de parámetros priors [31].

2.5.3 CAMPO DE APLICACIÓN

El desarrollo de la técnica del Kriging Bayesiano ha incrementado el uso de la geoestadística en trabajos de diferentes áreas, este incremento ha dado origen a evaluaciones comparativas con otros tipos de Kriging destacándose diversos trabajos, algunos de estos son:

En estimación de la calidad del aire se destacan estudios como: Nannavecchia et. Al (2015) *"Air Quality in Taranto: Multivariate Bayesian Kriging*" [32]. En el campo de la biología estudios como el de Arroyo et. al. (2003) "*Diferencias entre modelos geoestadísticos aplicados en el análisis de la distribución espaciotemporal de especies biológicas*" [33]. En estudios para la caracterización del subsuelo ejemplo de esto es: Nobre & Sykes (1992) "*Application of Bayesian Kriging to subsurface characterization*" [34]. En el campo agrícola estudios como Niyibizi et. al. (2018) "*Using Bayesian Kriging for Spatial Smoothing of Trends in the Means and Variances of Crop Yield Densities*" [35], entre otros.

2.5.4 VENTAJAS Y DESVENTAJAS

El uso de Kriging Bayesiano se ha utilizado en diferentes estudios dado que es estable para cualquier número de coeficientes y tiene una tendencia flexible. Algunas ventajas y desventajas del método son:

1. El modelado interactivo requerido es mínimo.
2. Los errores estándar de la predicción son más precisos que en otros métodos Kriging.
3. Permite realizar predicciones precisas de datos moderadamente no estacionarios.
4. Es más preciso que otros métodos Kriging para los *datasets* pequeños.
5. Usando el Kriging Bayesiano es posible incorporar en el modelo fuentes de incertidumbre asociadas a los parámetros de predicción y de esta forma encontrar estimaciones más realistas
6. El tiempo de procesamiento aumenta rápidamente a medida que el número de puntos de entrada, el tamaño del subconjunto o el factor de superposición se incrementan. Aplicar una transformación aumentará también el tiempo de procesamiento.
7. El procesamiento es más lento que en otros métodos Kriging, especialmente cuando los resultados se envían a un ráster.

2.6 INVERSO DE LA DISTANCIA

Inverso de la distancia (IDW, siglas en inglés de *"Inverse Distance Weighted"*) combina el concepto de vecindad entre lugares con disponibilidad de datos con un cambio gradual de las superficies definidas con una tendencia. Se supone que el valor del atributo Z en una posición donde el valor del atributo no es conocido es un promedio de los valores de sus vecinos [36].

La ecuación general del método es:

$$Z^*(x) = \sum_{i=1}^{n} \lambda_i * Z(x)_i \qquad \text{Ec. 2.72}$$

Donde,

- $Z^*(x)$: Valor estimado de la variable.
- $Z(x)_i$: Valor de la variable en el punto conocido i.
- λ_i: Peso de la estación i.
- n: Número de estaciones vecinas.

Donde los vecinos más cercanos tienen más peso que los más lejanos. Los pesos se calculan en función de la distancia entre las estaciones y el punto a interpolar, de la siguiente forma:

$$\lambda_i = \frac{\frac{1}{{d_{ij}}^{\beta}}}{\sum_{i=1}^{n} \frac{1}{{d_{ij}}^{\beta}}} \qquad \text{Ec. 2.73}$$

Donde,

- λ_i: Peso de la estación i.
- d_{ij}: Distancia geométrica entre la estación y el punto a interpolar.
- β: Coeficiente de ponderación.

Mediante IDW es posible controlar la importancia de los puntos conocidos sobre los valores interpolados en base a la distancia y la potencia del peso. Al definir una potencia alta, mayor será el énfasis sobre los puntos más cercanos en tanto que al definir una potencia menor, se les dará una mayor importancia a los puntos más distantes [6].

2.7 DEFINICIÓN DE RECURSOS MINERALES Y RESERVAS MINERAS

2.7.1 RECURSOS MINERALES

Es una concentración u ocurrencia de un material natural sólido inorgánico u orgánico fosilizado de interés económico que se encuentra en o bajo la corteza terrestre, de tal forma que el tonelaje, calidad o ley tengan perspectivas razonables para una eventual extracción económica.

Los Recursos Minerales se clasifican de acuerdo con el incremento de la confianza geológica del depósito en:

- Recursos Inferidos
- Recursos Indicados
- Recursos Medidos

2.7.1.1 RECURSO INFERIDO

Es aquella porción del recurso mineral para el cual el tonelaje y ley se estiman basándose en una limitada evidencia geológica y de muestreo. La evidencia geológica es suficiente para dar a entender la existencia de una mineralización, pero no garantiza la continuidad geológica y los contenidos de leyes presentes.

Un Recurso Mineral Inferido tiene un menor nivel de confianza que el aplicado a un Recurso Indicado y no se debe convertir en Reserva Minera. Existe una razonable expectativa de que la mayoría de los recursos minerales Inferidos con la continuación de trabajos de exploración, podrían ser actualizados (Re-categorizados) a recursos minerales indicados.

2.7.1.2 RECURSO INDICADO

Es aquella porción del Recurso Minero para el cual el tonelaje, ley o calidad, densidad, forma y características físicas son estimadas o interpretadas con suficiente certeza, que permiten aplicar los "factores modificatorios" lo suficientemente detallados para apoyar la planificación minera y la evaluación de la viabilidad económica del depósito.

Un Recurso Mineral Indicado tiene un menor nivel de confianza que el aplicado a un Recurso Mineral Medido y sólo se puede convertir en una Reserva Mineral Probable.

2.7.1.3 RECURSO MEDIDO

Es aquella porción del Recurso Minero para el cual el tonelaje, ley o calidad, densidad, forma y características físicas son estimadas o interpretadas con suficiente certeza permitiendo aplicar los "factores modificatorios" para respaldar la planificación minera detallada y la evaluación final de la viabilidad económica del depósito.

Un Recurso Mineral Medido tiene un mayor nivel de confianza que el aplicado a un Recurso Indicado o Recursos Inferidos. Se puede convertir a una Reserva Minera Probada o a una Reserva Minera Probable.

2.8 RESERVAS MINERAS

Es la parte económicamente explotable de un Recurso Mineral Medido o Indicado. Incluye dilución de materiales y tolerancias por pérdidas que se puedan producir cuando se extraiga el material.

Las Reservas Mineras se subdividen en orden creciente de confianza en Reservas Probables y Reservas Probadas. En la Figura 9 se muestra la relación entre recursos y reservas.

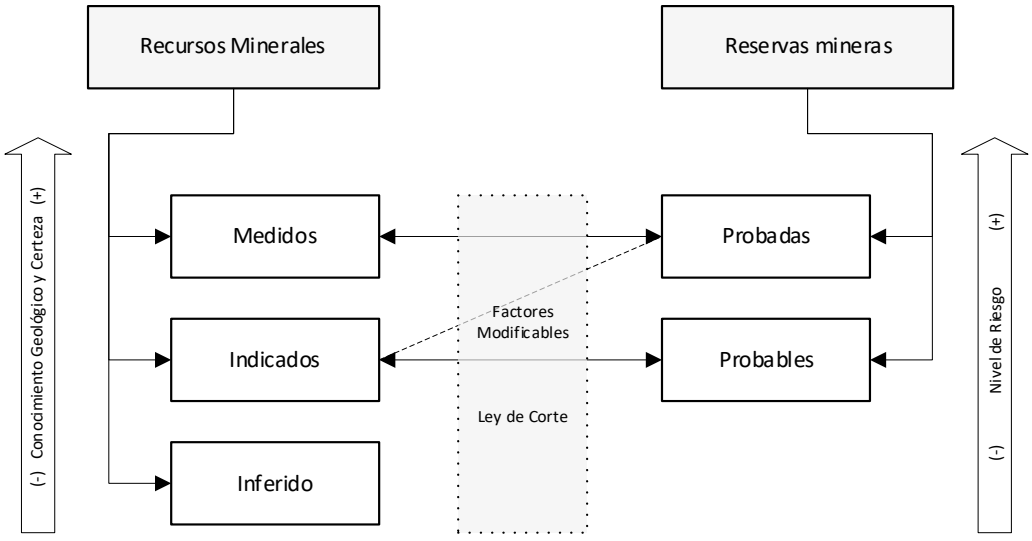


Figura 9: Relación entre Recursos y Reservas

CAPÍTULO 3

METODOLOGÍA

3.1 DESCRIPCIÓN

Para poder cumplir con los objetivos señalados la metodología consiste en la construcción de dos escenarios y el posterior análisis de estos; el primero de estos escenarios se basa en construir un conjunto de datos simulado con el que se obtendrá un modelo de bloques simulado, de este se tendrá toda información necesaria de manera tal que al realizar una ERM (sobre un conjunto de datos del modelo) se pueda comparar el verdadero valor entregado por el modelo de bloques simulado con los resultados de los modelos de regresión propuestos: Regresión Beta, Regresión Geográficamente Ponderada, Kriging Bayesiano y para los métodos tradicionales inverso de la distancia y Kriging ordinario; en el segundo escenario se realizará una ERM a un conjunto de datos reales de campo aplicando los cinco modelos mencionados.

Posteriormente se realizará una categorización de recursos para poder realizar un análisis de los escenarios y así restablecer conclusiones respecto a la diferencia de recursos minerales que entregará el depósito dependiendo del modelo de regresión elegido.

Con esta metodología de trabajo apoyada por herramientas computacionales se espera obtener resultados y establecer conclusiones respecto a cuál será el método de regresión que entregue la mejor, peor y más probable caso de recursos minerales para el depósito.

3.2 PROCEDIMIENTO DE TRABAJO

3.2.1 PROCEDIMIENTO CON DATOS SIMULADOS

A continuación, se presenta el procedimiento que se realizará al conjunto de datos simulados, correspondientes al primer escenario, cuyo objetivo principal es comparar los tres modelos de regresión propuestos:

1. Utilizando la herramienta computacional RStudio, se comienza creando el espacio de trabajo, el cual tiene como objetivo crear un espacio tridimensional con respectivas coordenadas geográficas que representará las dimensiones de un yacimiento mineral.
2. Se simulan tres atributos continuos que representaran dos leyes y un error, luego se construye una ley respecto a las dos leyes simuladas anteriormente.
3. Se genera una base de datos simulados con atributos continuos distribuidos tridimensionalmente, correspondientes a un muestreo intensivo de una zona de estudio determinada.
4. Se particionará la base de datos en sondajes de exploración, llamada base de entrenamiento y la información restante de la base simulada toma el nombre de base de validación.
5. Con la primera se construirán modelos de bloques con el atributo continúo predicho en las localidades de la base de datos de validación y éstos se compararán con el atributo real de la misma base. Esto se realizará para los algoritmos propuestos y tradicionales.
6. Se realizará un estudio comparativo con las medidas estadísticas principales asociadas a las predicciones.
7. Se estudiará la variabilidad y las medidas de incertidumbre de las predicciones realizadas a través de los métodos de regresión.
8. En cada uno de los modelos de bloques predichos se generará una categorización de recursos (Recursos medidos, Recursos indicados y Recursos inferidos), reportando las medidas estadísticas principales.

3.2.2 Procedimiento con Datos Reales

A continuación, se presenta el procedimiento que se realizará al conjunto de datos reales, correspondientes al segundo escenario, el cual tiene como objetivo principal comparar el uso de modelos propuestos con los tradicionales:

1. Se comenzará con una descripción detallada de la base de datos y un análisis exploratorio preliminar, obteniendo las estadísticas básicas para la base de datos del caso real.
2. Se utilizan resultados de Kriging Indicador de la Memoria de Título de Valentina Neira para obtener atributos categóricos del yacimiento real.

3. Se aplicarán los algoritmos propuestos y tradicionales para la construcción del modelo de bloques y cuantificación de la incertidumbre.
4. Se realizará la categorización de recursos en cada modelo de bloques obtenido.

En cada uno de los conjuntos de datos, se realizará un análisis crítico de los resultados y se elaborarán conclusiones y discusiones parciales.

En la FIGURA 10 se puede visualizar el esquema general del procesamiento de trabajo para el caso real y para el caso simulado.

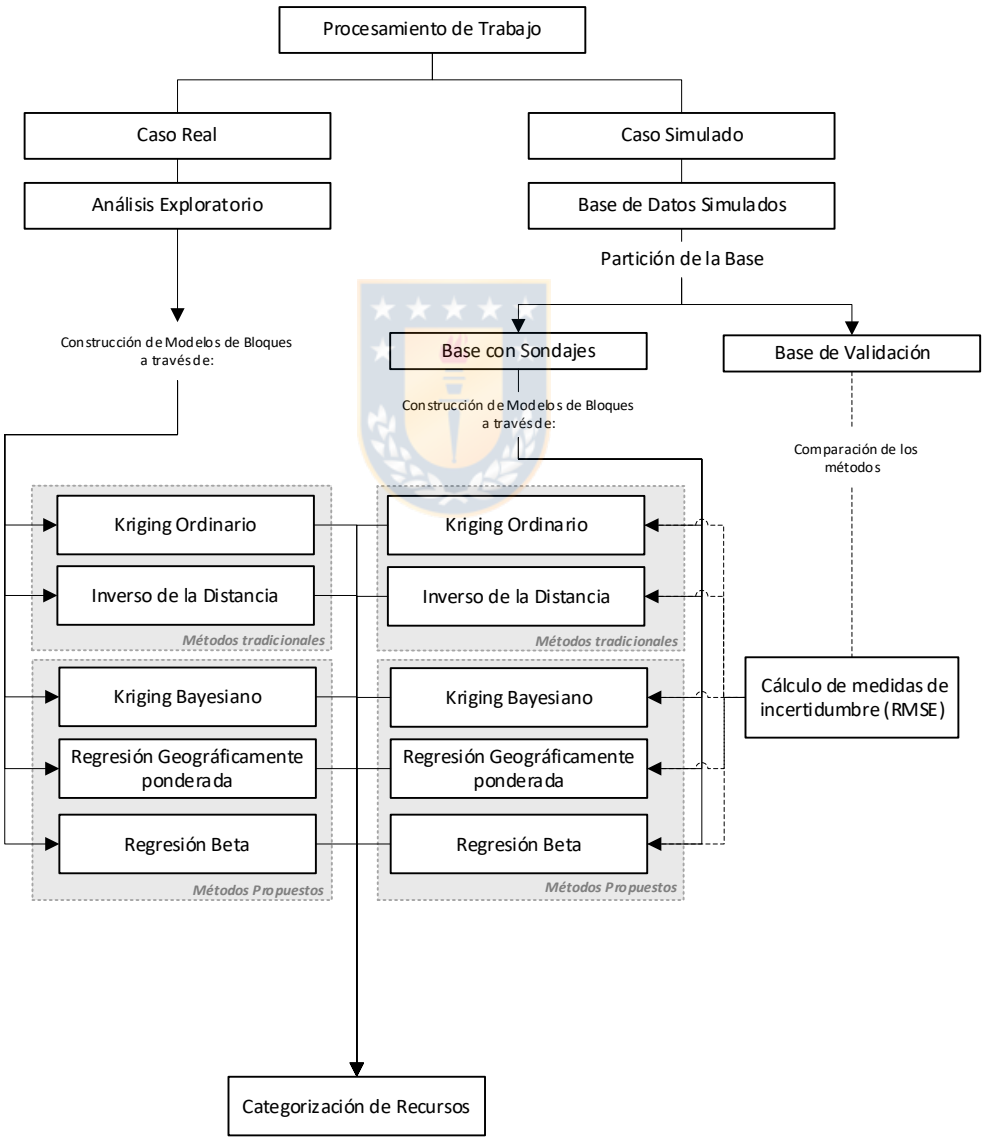


Figura 10: Esquema de procedimiento de trabajo.

3.3 HERRAMIENTAS COMPUTACIONALES

3.3.1 ENTORNO DE DESARROLLO INTEGRADO RSTUDIO

Para poder realizar la estimación de los datos con los modelos de regresión y llevar a cabo la construcción de los modelos de bloques se utilizará la herramienta RStudio que es una interfaz que permite acceder de manera sencilla a toda la potencia de R (Incluye una consola, un editor de resultado de sintaxis que admite la ejecución directa de código y una variedad de herramientas robusta) [37].

R es un lenguaje y entorno de programación, creado en 1993 por Ross Ihaka y Robert Gentleman del Departamento de Estadística de la Universidad de Auckland, cuya característica principal es que forma un entorno de análisis estadístico para la manipulación de datos, su cálculo y la creación de gráficos [38].

CAPÍTULO 4

CASO SIMULADO

4.1 CONSTRUCCIÓN BASE SIMULADA

4.1.1 GENERACIÓN ESPACIO DE TRABAJO

Para la Base Simulada se comienza con la creación del espacio de trabajo, esto se logra con la generación de coordenadas geográficas (x, y, z) a través del despliegue de grillas. Se decide generar una grilla desde la coordenada $(0, 0, 0)$ hasta la coordenada $(200, 200, 200)$, con un distanciamiento de 10 unidades, obteniendo de un total de 8000 puntos.

Cada punto generado representa el centroide de cada bloque del modelo de bloques, en consecuencia, las dimensiones del yacimiento generado son de $200 \times 200 \times 200\ [m^3]$ y cada bloque tiene dimensiones $10 \times 10 \times 10\ [m^3]$. En la **Figura 11** se visualiza el área de trabajo descrita anteriormente.

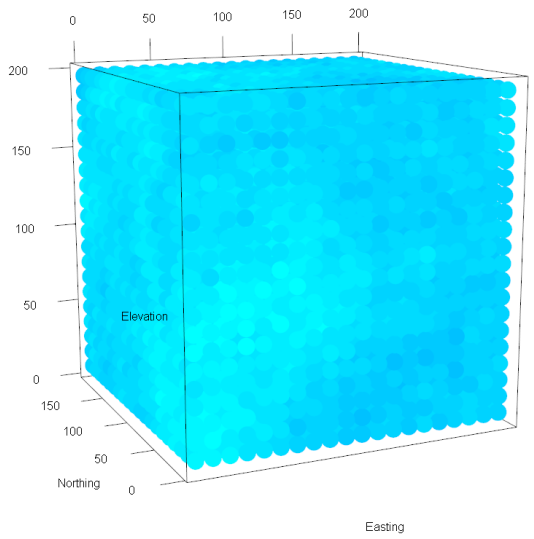


Figura 11: Visualización de área de trabajo.

4.1.2 SIMULACIÓN DE ATRIBUTOS CONTINUOS

Con la meta de poder representar la variabilidad de los atributos continuos y representar los diferentes escenarios de un yacimiento mineral se generan 100 simulaciones y en cada simulación se genera una ley artificial que se construye a partir de otras dos leyes simuladas y además se le agrega un error para poder representar de mejor manera un caso real.

En la Ec. 4.1 se muestra el modelo con el cual se construye la ley artificial Y en los 100 yacimientos simulados.

$$Y_i = a_0 + a_1 * v_1 + a_2 * v_2 + e, \qquad con\ i = 1, \dots, 100 \qquad \text{Ec. 4.1}$$

Donde,

- Y_i: Distribución de ley artificial en la simulación i.
- v_1: Distribución de ley simulada 1 en la simulación i.
- v_2: Distribución de ley simulada 2 en la simulación i.
- e: Distribución del error en la simulación i.
- a_j: Ponderadores arbitrarios, con $j = 0, 1, 2$.

Para la construcción de las leyes simuladas (v_1 y v_2) y el error (e) se utilizan variogramas simples con el fin de obtener correlación de las variables; además se escogen los parámetros de estos variogramas de manera tal que la construcción de las leyes se asemeje lo máximo posible a un yacimiento real.

En la **Tabla 4** se muestran los modelos y parámetros de los variogramas escogidos en la generación de atributos continuos en las 100 simulaciones.

Tabla 4: Parámetros de variogramas para la construcción de atributos continuos.

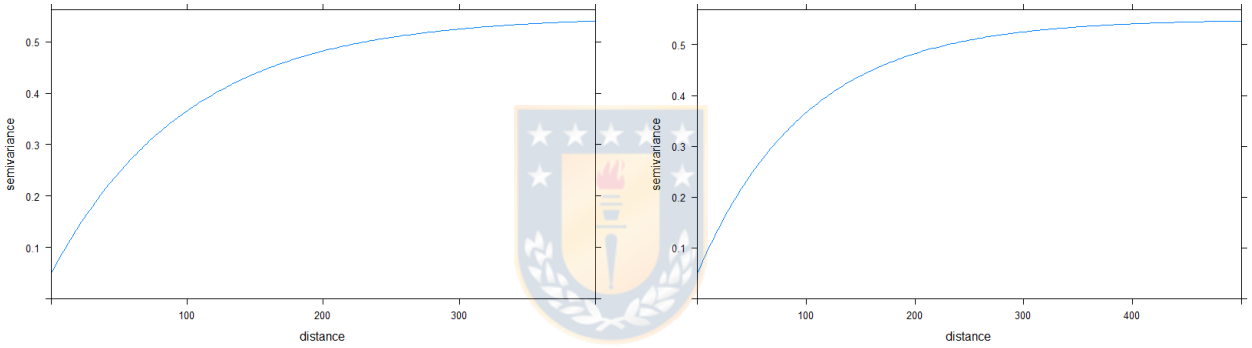
Atributo continuo	Modelo	Pepa	Meseta	Rango
Ley v_1	Exponencial	0.05	0.5	100
Ley v_2	Gaussiano	0	1.2	50
Error e	Exponencial	0.05	0.5	100

En las siguientes ecuaciones (Ec. 4.2, Ec. 4.3 y Ec. 4.4) se muestran las fórmulas de los variogramas de las leyes simuladas (v_1 y v_2) y el error (e) respectivamente; y en la **Figura 12** se visualizan las gráficas de los variogramas, donde h es la distancia.

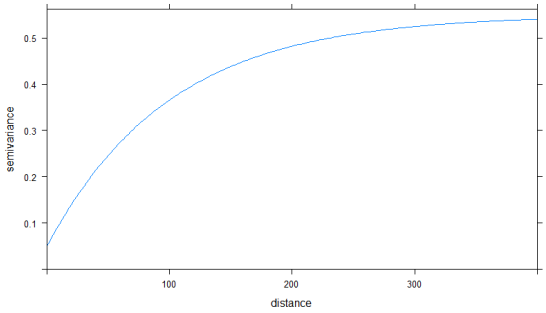
$$\gamma_{v_1}(h) = 0.05 + 0.5\left[1 - e^{-\left(\frac{h}{100}\right)}\right] \quad \text{Ec. 4.2}$$

$$\gamma_{v_2}(h) = 1.2\left[1 - e^{-\left(\frac{h}{100}\right)^2}\right] \quad \text{Ec. 4.3}$$

$$\gamma_e(h) = 0.05 + 0.5\left[1 - e^{-\left(\frac{h}{100}\right)}\right] \quad \text{Ec. 4.4}$$



a) Variograma Ley $\mathbf{v_1}$ **b)** Variograma Ley v_2



c) Variograma Error e

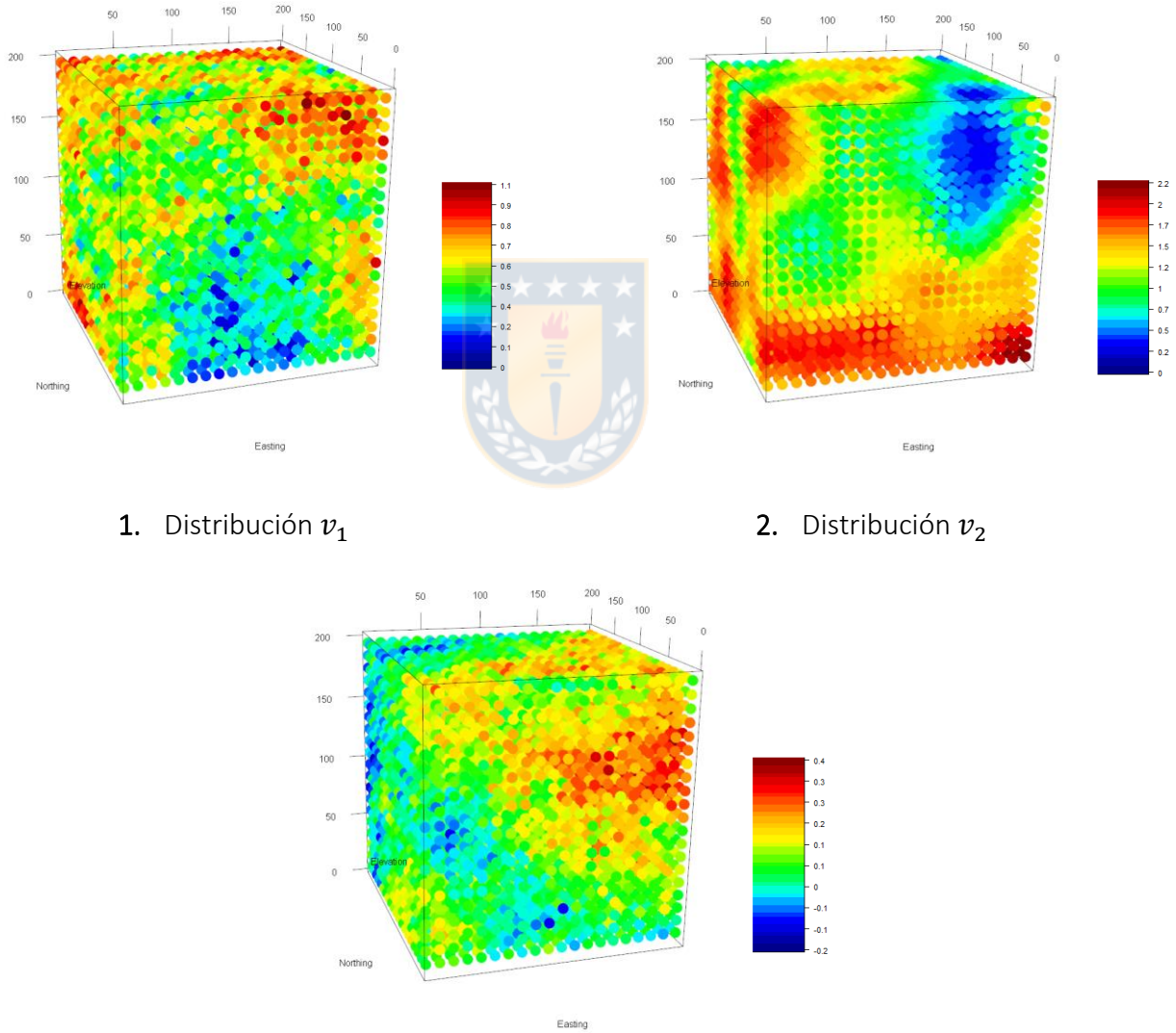
Figura 12: Variogramas simples para la construcción de Leyes simuladas v_1, v_2 y error e.

Teniendo construidos los variogramas se procede a establecer la ley media deseada $(\overline{Ley\ v}\ )$ para v_1 y v_2. En este caso, se escoge como ley media para v_1 0.55 y para v_2 1.3 y para llevarlo a la escala correspondiente y la ley sea segada positivamente se utiliza la ecuación Ec. 4.5.

$$Ley\ v^* = \frac{Ley\ v - min(Ley\ v)}{mean(Ley\ v - min(Ley\ v))} * \overline{Ley\ v} \qquad \text{Ec. 4.5}$$

Dado que el error puede tomar valores tanto positivos como negativos no se utiliza la estandarización anterior (Ec. 4.5) para transformar el rango.

En la **Figura 13** se muestra la visualización del resultado de la distribución de las leyes simuladas v_1 y v_2 y el error e para la simulación 100.



1. Distribución v_1

2. Distribución v_2

3. Distribución e

Figura 13: Distribuciones de leyes v_1 y v_2 y el error e para la simulación 100.

Para la construcción de la ley y se utiliza la Ec. 4.1 antes nombrada, donde se utilizan los ponderadores $a_0 = 0.8$, $a_1 = 1.5$, $a_2 = -0.5$, Además, se multiplica el resultado del error e por 0.125 para estandarizarlo en el rango correspondiente. Lo anterior se resumen en Ec 4.6. Por último, se utiliza Ec. 4.5 para establecer una ley media de 0.55.

$$Y_i = 0.8 + 1.5 * v_1 - 0.5 * v_2 + 0.125 * e, \qquad con\ i = 1, \dots, 100 \qquad \text{Ec. 4.6}$$

En la **Figura 14** se muestra la visualización del resultado de la distribución de la ley y para la simulación 100. Además, en el anexo 8.17 se encuentra una tabla resumen de las estadísticas de la ley y en las 100 simulaciones.

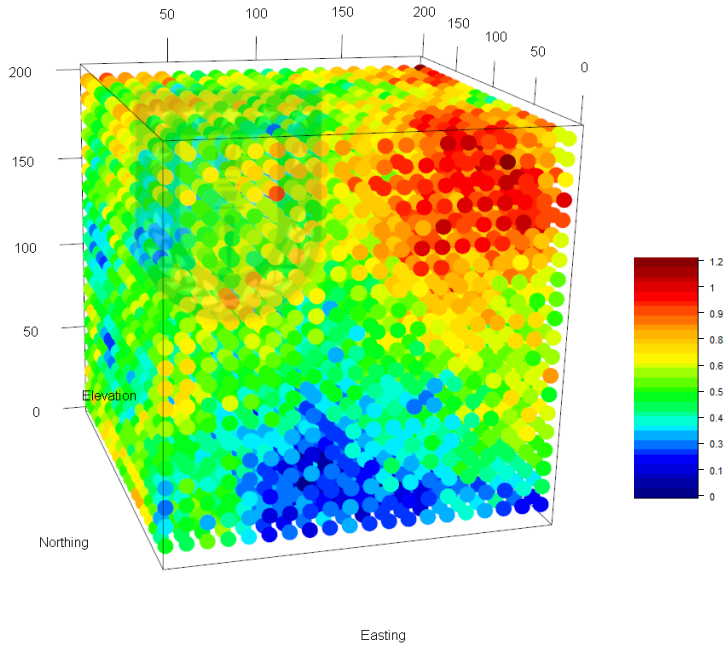
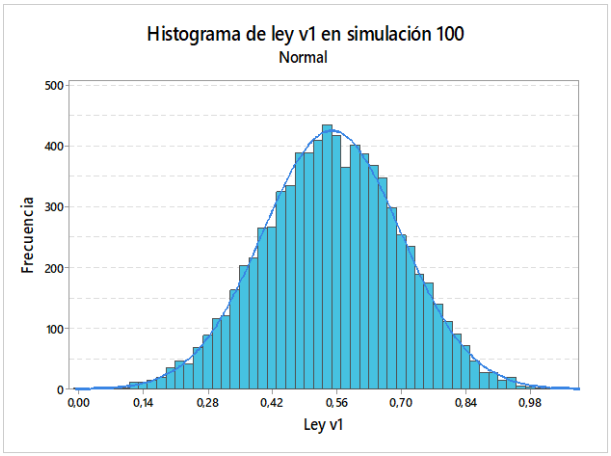


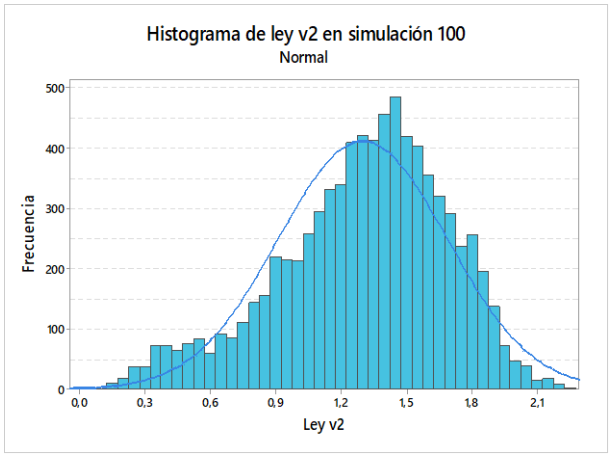
Figura 14: Distribución de ley y en simulación 100.

En la **Figura 15** se muestras las distribuciones de las leyes y, v_1 y v_2 y del error e para la simulación 100, se aprecia que todas las leyes son sesgadas positivamente, con un mínimo de 0 que es lo que se deseaba construir, para asemejarse a un yacimiento real, además, la ley media de y en la simulación es 0.55 con un ajuste de distribución normal.

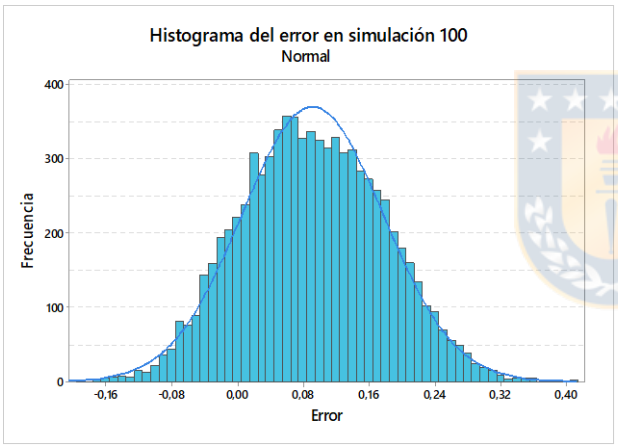
El histograma de ley y es utilizado para comparar los resultados de distribución de ley y obtenidos al aplicar métodos de regresión con la distribución real.



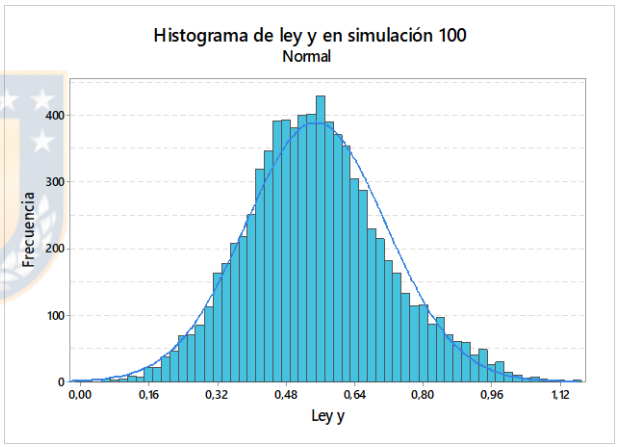
(a) Histograma de ley v_1.



(b) Histograma de ley v_2.



(c) Histogramam de e.



(d) Histograma de ley y.

Figura 15: Histogramas de distribución de leyes y erros en simulación 100.

4.1.3 PARTICIÓN DE LA BASE SIMULADA

Es necesario particionar las bases simuladas en un conjunto de entrenamiento, que se construye a partir de la extracción de sondajes aleatorios, y un conjunto de validación, la cual se utiliza posteriormente para hacer el cálculo de incertidumbre entre el resultado obtenido por los modelos de regresión y la base real.

Para la extracción aleatoria de los sondajes se toma como consideración los siguientes puntos:

- Ángulo de inclinación: $90º$.
- Profundidad mínima del sondaje: $30\ [m]$ (15% de la profundidad total del yacimiento).
- Profundidad máxima del sondaje: $150\ [m]$ (75% de la profundidad total del yacimiento).
- Cantidad de sondajes a extraer: 28 (representa 7% de la información de la superficie del yacimiento).

Luego de ejecutar el código para la extracción de los sondajes se obtienen como resultado los sondajes mostrados en las **Figura 16** y **Figura 17**. Para información más detallada de los sondajes extraídos ver anexo 8.18.

Por otro lado, los 28 sondajes obtenidos representan el 3.76% de la información del yacimiento mineral, correspondientes a 301 puntos, es decir la base de entrenamiento tendrá el 3.76% de la información y la base de validación tendrá el 96.24% de la información restante. Dado que se requiere hacer una comparación de los modelos de regresión se extraerán en las 100 simulaciones los mismos sondajes.

Figura 16: Vista planta Sondajes simulación 100.

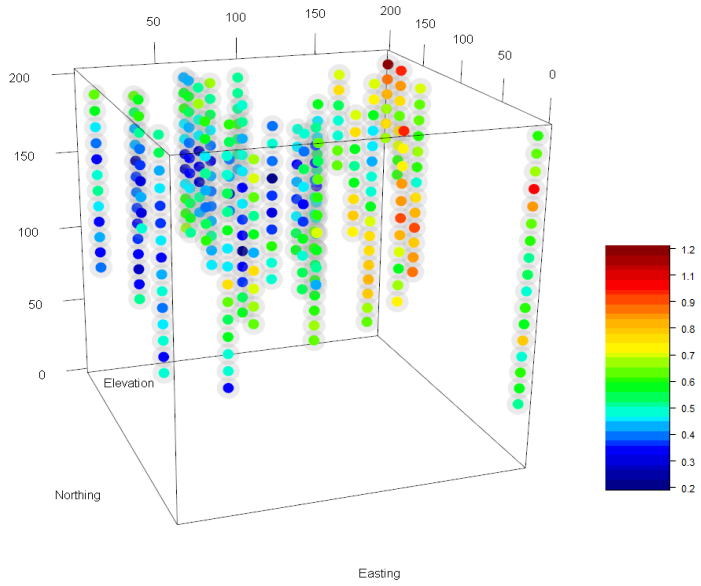


Figura 17: Visualización de sondajes en simulación 100.

4.2 ESTIMACIÓN DE LEYES v_1 Y v_2

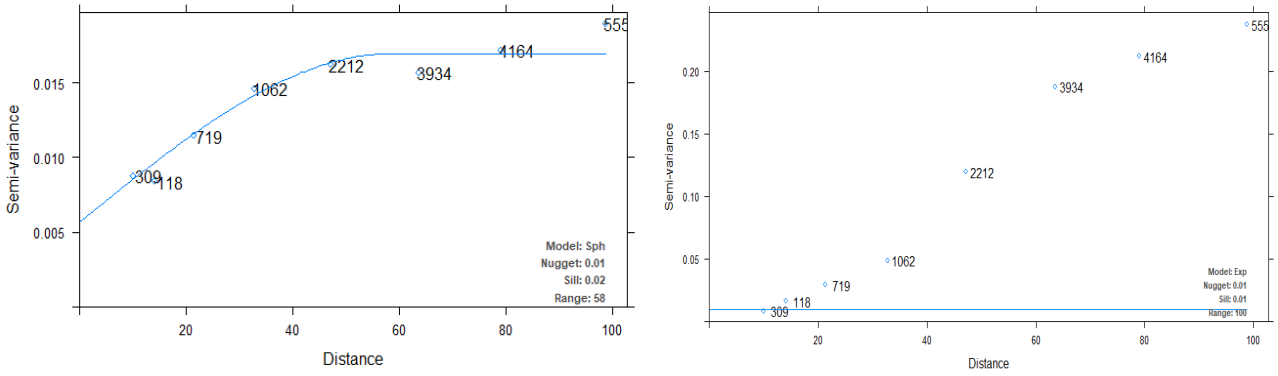
Se estiman las leyes v_1 y v_2 a través de Kriging Ordinario para luego hacer la estimación de la ley y a través de los métodos de regresión propuestos, de esta manera poder establecer la eficiencia de estos métodos teniendo en consideración el conocimiento de los atributos v_1 y v_2.

Para realizar la simulación de construyen variogramas omnidireccionales para cada simulación utilizando la función de *autofitVariogram* en R. La **Tabla 5** muestra los resultados obtenidos en la simulación 100, además en la **Figura 18** se encuentra la visualización de estos variogramas.

Tabla 5: Parámetros de variogramas experimentales para leyes v_1 y v_2, simulación 100.

Atributo continuo	Modelo	Pepa	Meseta	Rango
Ley v_1	Esférico	0.01	0.02	58
Ley v_2	Exponencial	0.01	0.01	100

Por otro lado, en la **Figura 19** se muestra un gráfico de caja del número de observaciones máximas que debieron utilizarse para una realizar la predicción a través de Kriging Ordinario en las 100 simulaciones y en la **Figura 20** y la **Figura 21** se visualiza la distribución de las leyes en la simulación 100 luego de aplicarse el método de regresión.



a) Variograma experimental Ley v_1 b) Variograma experimental Ley v_2

Figura 18: Variogramas experimentales de simulación de leyes simuladas v_1 y v_2, simulación 100.

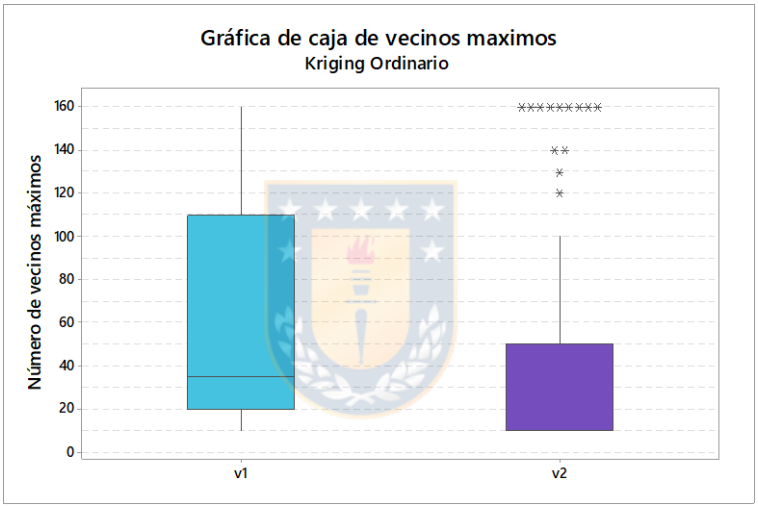


Figura 19: Gráfica de caja de vecinos máximos utilizados en la estimación de leyes v_1 y v_2.

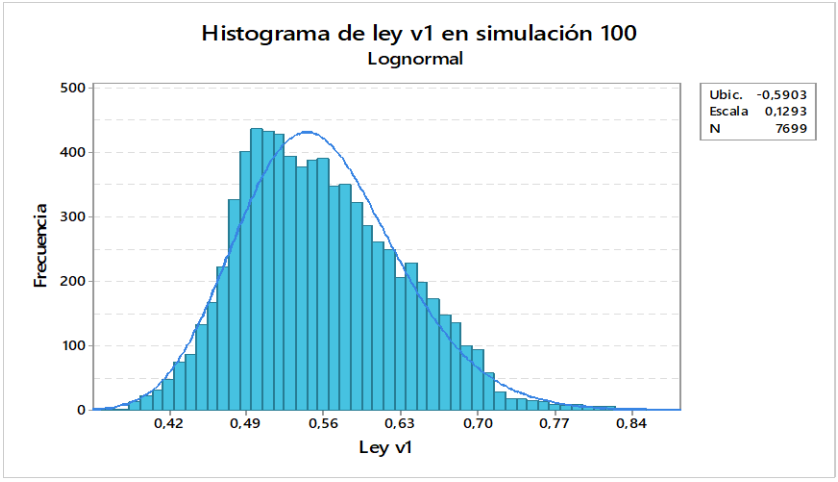


Figura 20: Histograma de estimación ley v_1 en simulación 100.

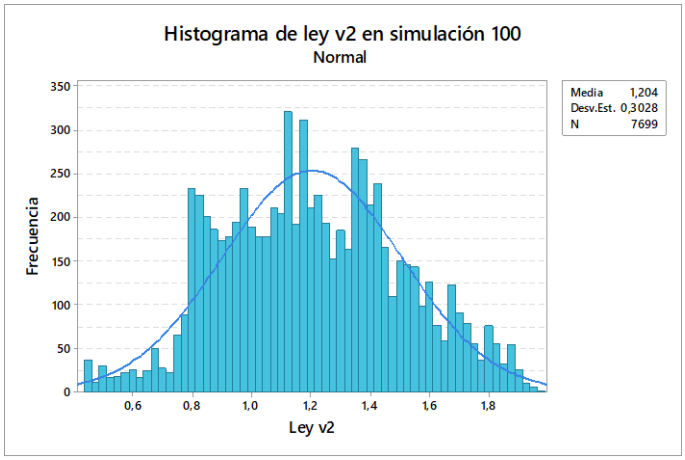


Figura 21: Histograma de estimación ley v_2 en simulación 100.

A través del mismo procedimiento se estima la ley y con el fin de obtener las varianzas de Kriging de los puntos estimados, los cuales son utilizados para realizar la categorización de recursos (ver título 4.5).

4.3 APLICACIÓN MODELOS DE REGRESIÓN PARA ESTIMAR LEY y

Con las 100 bases de entrenamiento ya construidas se procede a aplicar los métodos de regresión propuestos para la variable y. A continuación, se muestra la metodología y los resultados obtenidos en las simulaciones.

4.3.1 REGRESIÓN BETA

Como lo descrito en el capítulo 2 para poder hacer uso de regresión Beta se debe cumplir que los parámetros deben asumir valores en el intervalo estándar $(0, 1)$, producto que las bases de entrenamiento no cumplen con esta condición se debe utilizar Ec 2.3 para estandarizar.

Obteniendo los máximos y mínimos de la variable y y utilizando Ec. 2.3 se crean las nuevas bases llamadas Bases de entrenamiento normalizadas. En la **Figura 22** se muestra un gráfico de caja de la transformación de la ley y en la simulación 100.

Para la construcción del código presentado en el anexo 8.19 las funciones que optimizan el código para el caso simulado son la función enlace *"LogLog"* y los caracteres de la función de enlace *"Identity"* esto último producto que el modelo es de la forma $y \sim Northing + Easting + Elevation + v_1 + v_2$.

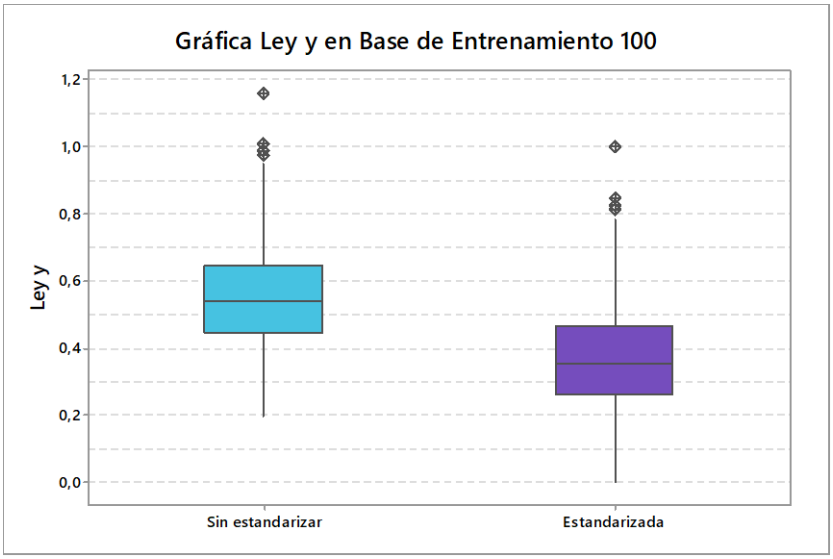


Figura 22: Gráfica de Caja para la estandarización de la ley y en simulación 100.

Cabe destacar que los resultados entregados por el código de la Regresión Beta estarán dentro del intervalo $(0\ ,1)$ por lo que se aplicar a los datos el inverso de la ecuación Ec. 2.3 y así llevarlos a la escala original.

Se muestra la **Figura 23** la visualización del modelo obtenido utilizando Regresión Beta de la simulación 100 y en la **Figura 24** se muestra un histograma de la ley y de la misma simulación.

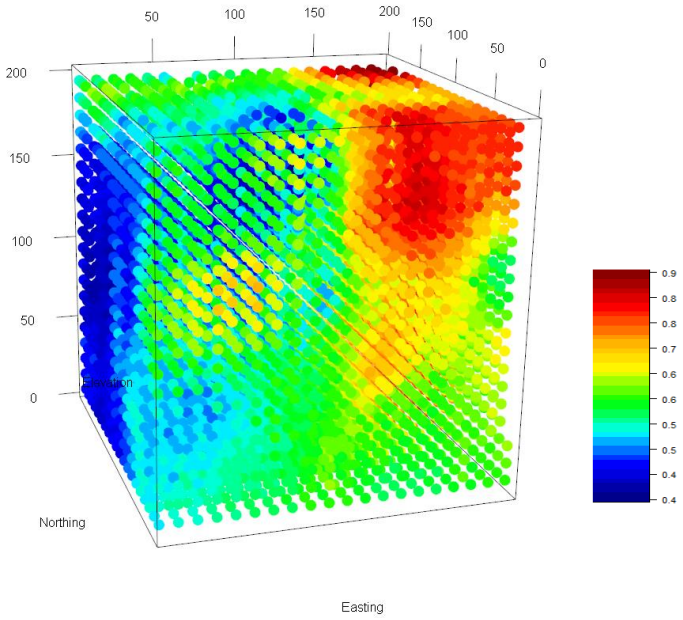


Figura 23: Base simulada 100 con BREG.

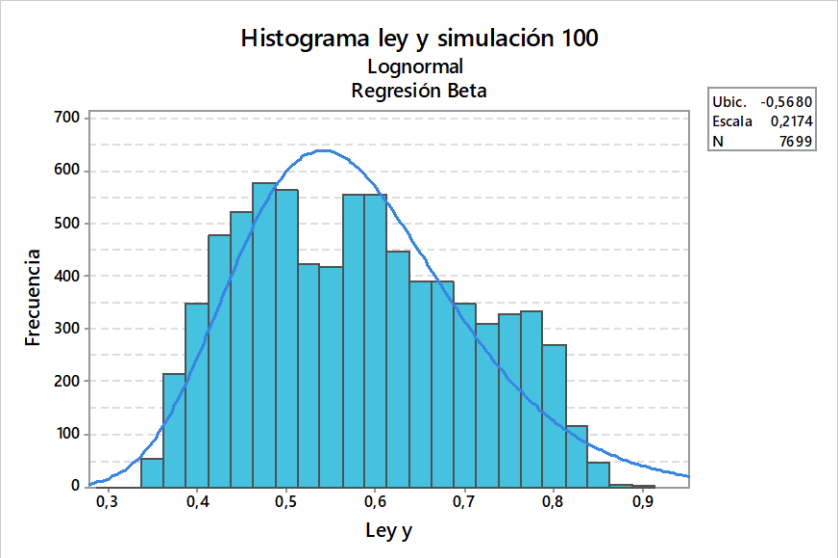


Figura 24: Histograma de ley y con BREG en simulación 100

4.3.2 REGRESIÓN GEOGRÁFICAMENTE PONDERADA

Para GWR producto de la misma construcción del modelo (busca valores en el mismo plano) lo primero que se realiza es la división de la simulación por niveles dejando un total de 20 niveles diferentes que corresponden a cada cota del yacimiento simulado.

Optimizando el código (ver anexo 8.20) se establece que se utiliza GWR con un ancho de banda variable, para cada iteración se prueban 50 tipos de ancho de bandas diferentes, la cantidad de vecinos a buscar para hacer la estimación varía entre 7% y el 50% del total de los datos presentes en la base de entrenamiento y el tipo de kernel escogido es gaussiano.

En la **Tabla 6** se muestra el resumen de los parámetros con el cual se optimiza el código GWR para las simulaciones.

Tabla 6: parámetros de optimización para GWR.

Parámetros	Valor
Cotas totales	20
Cantidad de Anchos de bandas	50
Cantidad mínima de vecinos	21
Cantidad máxima de vecinos	150
Kernel	Gaussiano

En la **Figura 25** se muestra un histograma de los resultados de ancho de bandas obtenidos para las 100 simulaciones y la **Tabla 7** se detalla la estadística resultante, donde se aprecia que la cantidad de vecinos que optimizan el modelo tienen a concentrarse en los valores extremos escogidos, resultando que la moda de 150 vecinos y una segunda moda de 21 vecinos, con frecuencia de 18 y 16 respectivamente.

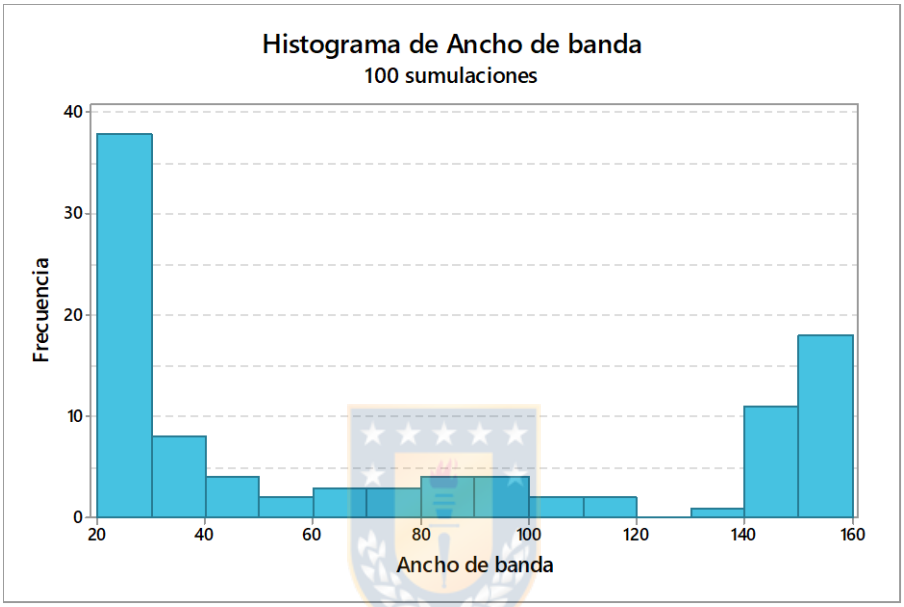


Figura 25: Histograma del ancho de banda obtenidos en las 100 simulaciones.

Tabla 7: Estadística descriptiva de los anchos de banda en las 100 simulaciones.

Variable	Valor	Frecuencia
N	100	---
Media	74.7	---
Mínimo	21	16
Máximo/Moda	150	18

Por último, luego de ejecutar el código y de obtener los resultados de los puntos a estimados se deben unen los niveles que anteriormente habían sido separados y se construye el yacimiento simulado.

En la **Figura 26** se visualiza el modelo obtenido utilizando GWR en la simulación 100, logrado con un ancho de banda de 63 y en la **Figura 27** se muestra un histograma de la ley y de la misma simulación.

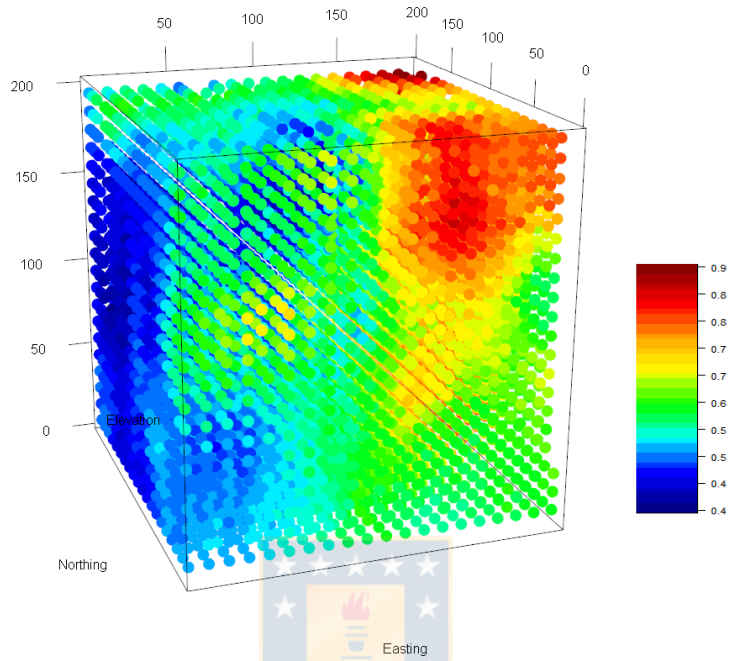


Figura 26: Base simulada 100 con GWR.

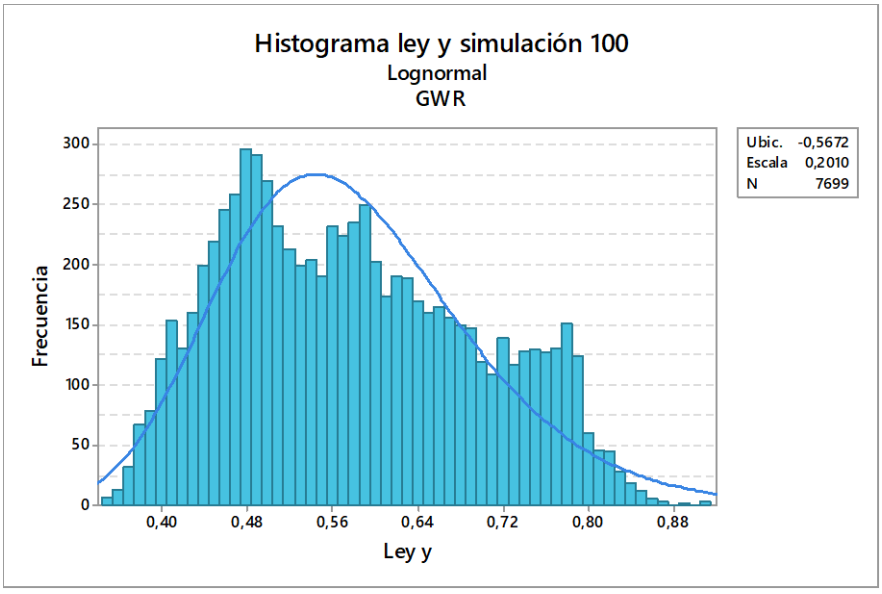


Figura 27: Histograma de ley y con GWR en simulación 100.

4.3.3 KRIGING BAYESIANO

Al igual que el caso de GWR, para utilizar el código de BKR es necesario separar la base por niveles, esto producto que el código utilizado en R es 2D, en consecuencia, se separa la base simulada, obteniendo planos cada $10\ [m]$ de distancia.

Para la construcción del código es necesario establecer los *inputs*, es decir, el *"model"*, que define los componentes del modelo y el *"prior"* que es la especificación del variograma para los parámetros del modelo (ver anexo 8.21).

Para el caso del *"model"* se especifican los valores de tendencia (covariables) en las ubicaciones de los datos, para esto se utiliza una tendencia (*trend*) de la forma $trend.spatial \sim Elevation + v_1 + v_2$.

Por otro lado, para el caso del *"prior"* se establecen los parámetros del variograma, en donde los valores fijos están dados por el variograma autoajustado de la base de sondajes de la forma $y \sim v_1 + v_2$ y los *priors* para los variogramas, que corresponden a las distribuciones de la meseta, rango y pepa están dados por "reciprocal", "uniform" y "uniform" respectivamente.

En la **Figura 29** se visualiza el modelo obtenido utilizando BKR en la simulación 100 y en la **Figura 28** se muestra un histograma de la ley y de la misma simulación.

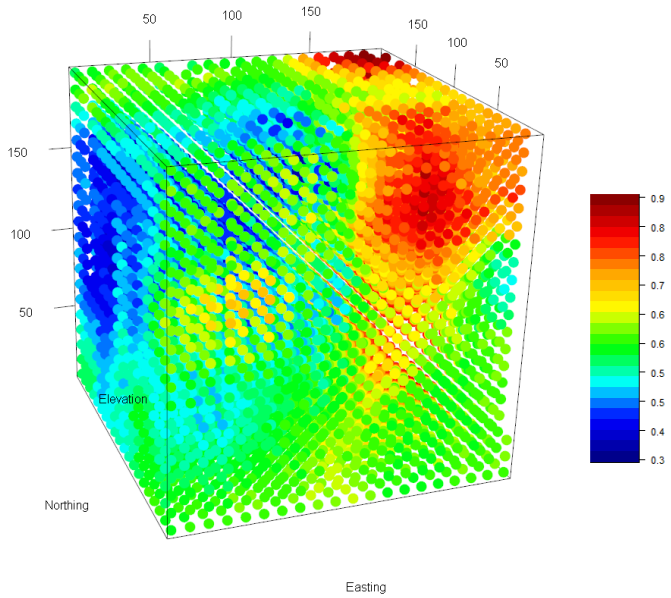


Figura 28: Base simulada 100 con BKR.

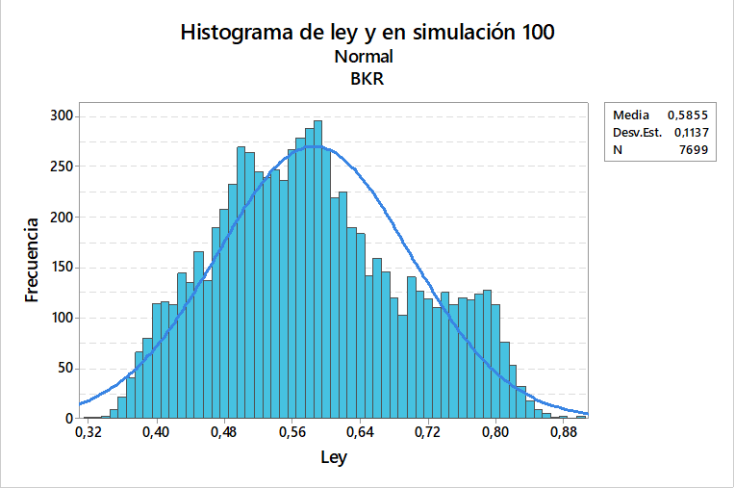


Figura 29: Histograma de ley y con BKR en simulación 100.

4.3.4 INVERSO DE LA DISTANCIA

Para la construcción de código de IDW se necesita establecer el coeficiente de ponderación (β) y la cantidad máxima de vecinos ($nmax$) que se pueden utilizar para realizar la estimación de y, en consecuencia, se iteran (β) entre 3 y 5 con un aumento de 0.1 y para los máximos vecinos se prueba valores entre 70 y 250 con un aumento de 10.

Para las combinaciones mencionadas anteriormente se rescatan para cada simulación las que obtienen los menores valores de RMSE. En la **Figura 30** y **Figura 31** se muestran los histogramas de los resultados para los β y $nmax$ en las 100 simulaciones.

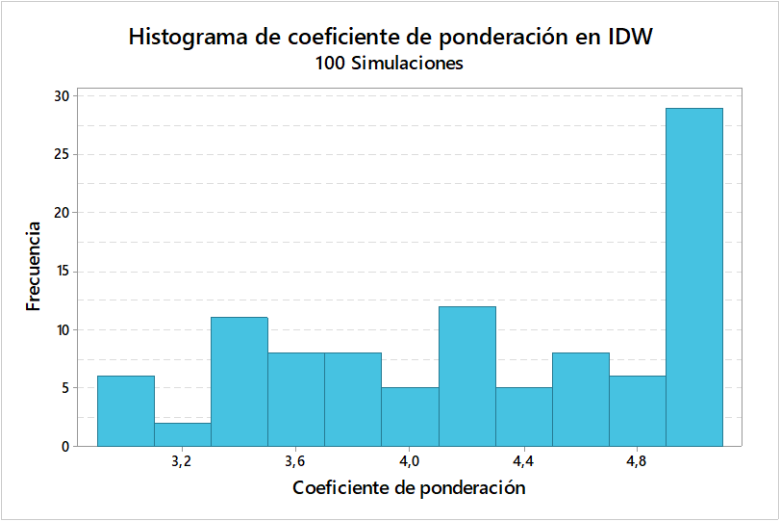


Figura 30: Histograma de coeficiente de ponderación en estimación de y en las 100 simulaciones.

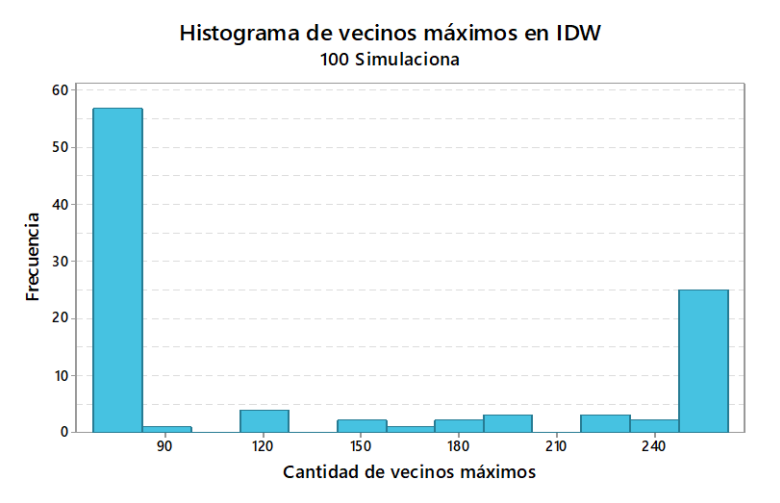


Figura 31: Histograma de vecinos máximos en estimación de y en las 100 simulaciones.

En la **Figura 32** se visualiza del modelo obtenido utilizando IDW en la simulación 100, logrado con un $\beta = 3.3$ y un $nmax = 250$ y en la **Figura 33** se muestra un histograma de la ley y de la misma simulación.

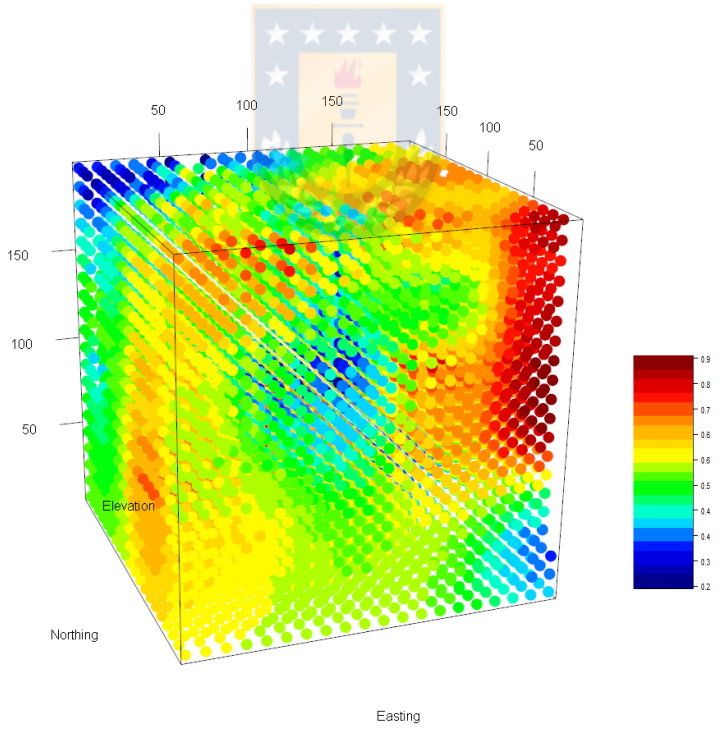


Figura 32: Base simulada 100 con IDW.

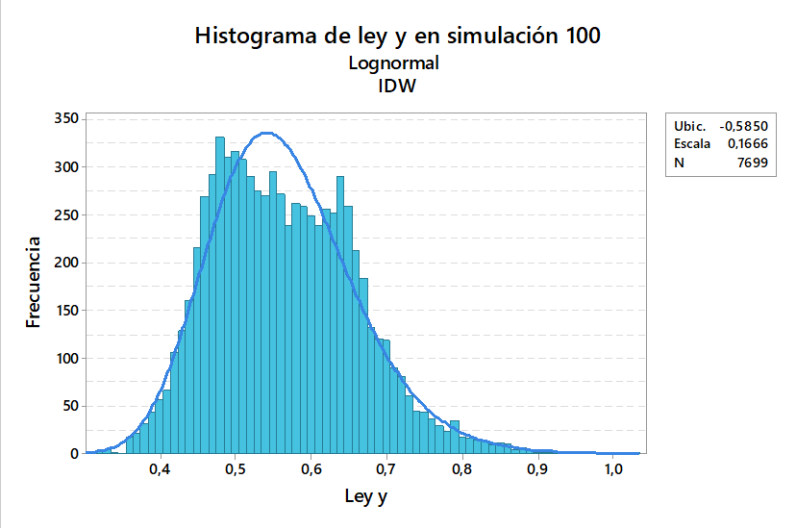


Figura 33: Histograma de ley y con IDW en simulación 100.

4.3.5 KRIGING ORDINARIO

Para la construcción de OK se comienza construyendo un variograma omnidireccional para la ley y utilizando la función *autofitVariogram*, para el caso de la simulación 100 se obtiene el variograma mostrado en la **Figura 34**.

Además, en la **Figura 35** se visualiza un histograma con los resultados de vecinos máximos utilizados en las 100 simulaciones.

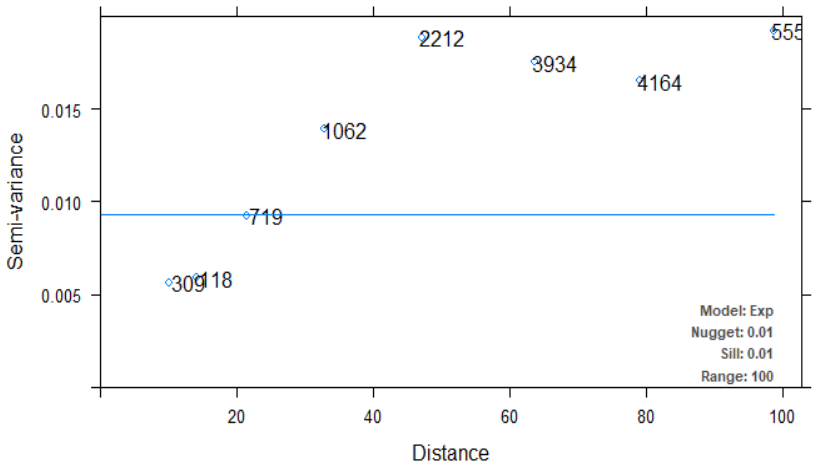


Figura 34: Variograma experimental de simulación de ley y en simulación 100.

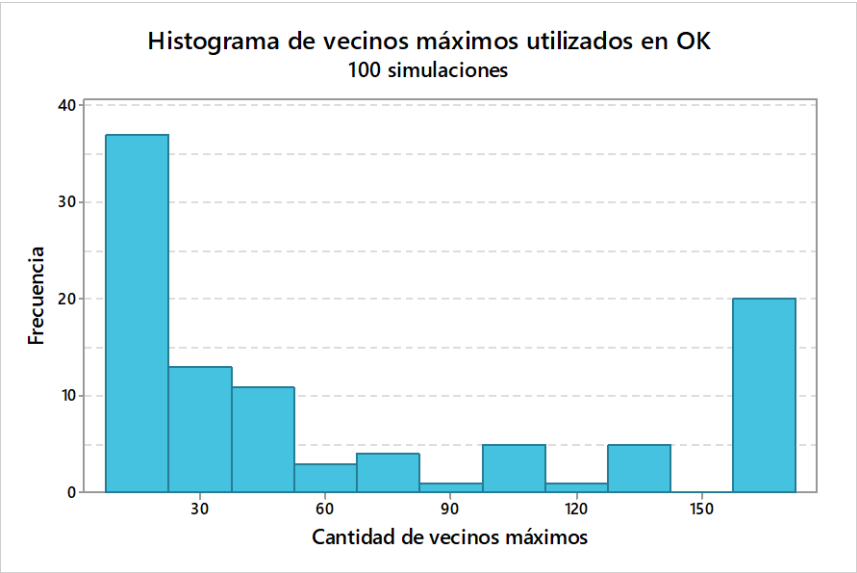


Figura 35: Histograma de vecinos máximos utilizados en la estimación de ley y.

En la **Figura 36** se visualiza el modelo obtenido utilizando OK en la simulación 100, logrado con una cantidad de vecinos máximos de 20 y en la **Figura 37** se muestra un histograma de la ley y de la misma simulación.

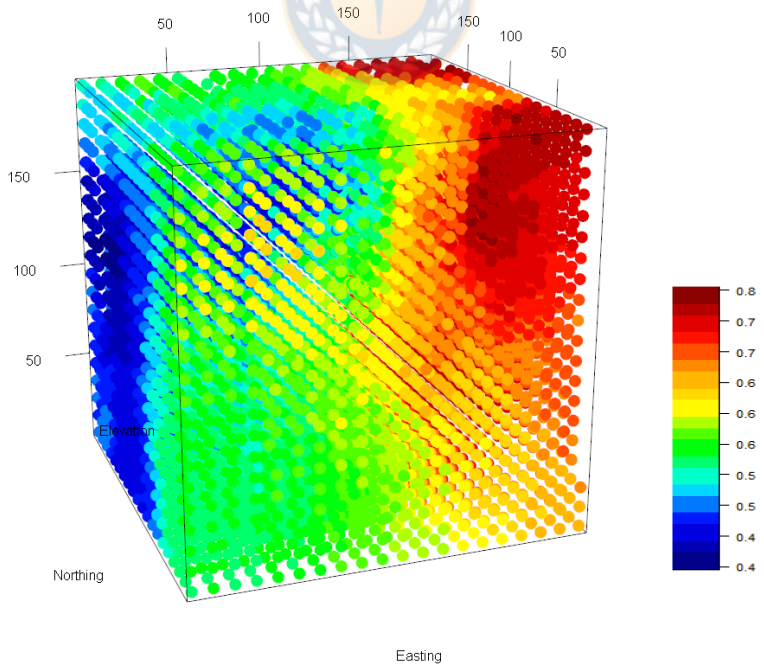


Figura 36: Base simulada 100 con OK.

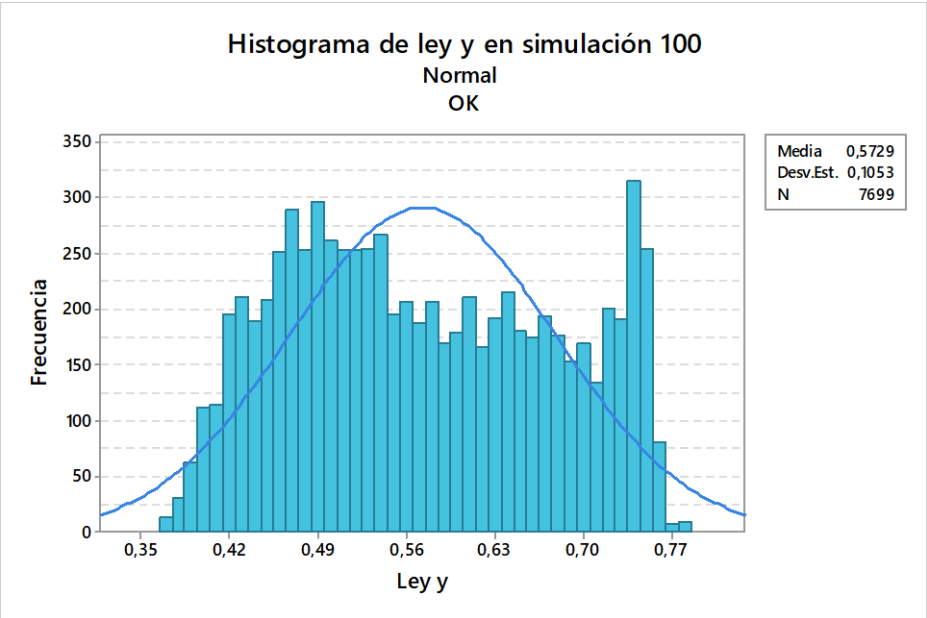


Figura 37: Histograma de ley y con OK en simulación 100.

4.4 ANÁLISIS DE ERROR DE LOS MODELOS

Para hacer el análisis de la calidad de las estimaciones realizadas se utiliza la raíz del error cuadrático medio (RMSE) (ver título 2.1.6). De esta manera utilizando la ecuación Ec. 2.2 se obtienen los resultados mostrados en el anexo 8.22, además, en la **Figura 38** y la **Figura 39** se muestran las gráficas de caja de los resultados de RMSE para las leyes v_1 y v_2 y la ley y respectivamente, para las 100 simulaciones.

Por otro lado, se determina para cada método cual es el mejor, el peor y el más probable caso, concluyendo que la simulación 94 es el mejor caso con un RMSE de 0.073 a través del método de regresión IDW, el peor caso en la simulación 40 con un RMSE de 0.183 a través del método GWR y BREG y el caso más probable se establece, a través de ponderación simple que se encuentra en la simulación 16.

En la **Tabla 8** se encuentra un resumen de la mejor, peor y el más probable caso para cada método y se muestra la media y la desviación estándar (σ) para las leyes estimadas y, v_1 y v_2 para cada uno de los casos señalados.

Tabla 8: Análisis de RMSE para el mejor, peor y más probable caso.

				y		v_1		v_2	
	Método	Sim	RMSE	Media	σ	Media	σ	Media	σ
Mejor caso	BREG	94	0.079	0.564	0.064	0.568	0.064	1.246	0.241
	GWR	94	0.078	0.561	0.059	0.568	0.064	1.246	0.241
	BKR	94	0.079	0.563	0.06	0.568	0.064	1.246	0.241
	IDW	94	0.073	0.56	0.069	0.568	0.064	1.246	0.241
	OK	94	0.076	0.558	0.071	0.568	0.064	1.246	0.241
Peor caso	BREG	40	0.183	0.489	0.124	0.526	0.087	1.374	0.381
	GWR	40	0.183	0.488	0.115	0.526	0.087	1.374	0.381
	BKR	53	0.173	0.547	0.165	0.559	0.138	1.318	0.268
	IDW	40	0.174	0.492	0.118	0.526	0.087	1.374	0.381
	OK	53	0.159	0.558	0.147	0.559	0.138	1.318	0.268
Caso más probable	BREG	31	0.124	0.586	0.094	0.555	0.105	1.252	0.291
	GWR	16	0.121	0.570	0.093	0.568	0.079	1.278	0.24
	BKR	16	0.121	0.566	0.090	0.568	0.079	1.278	0.24
	IDW	51	0.123	0.523	0.114	0.526	0.102	1.319	0.361
	OK	49	0.118	0.554	0.134	0.56	0.115	1.312	0.366

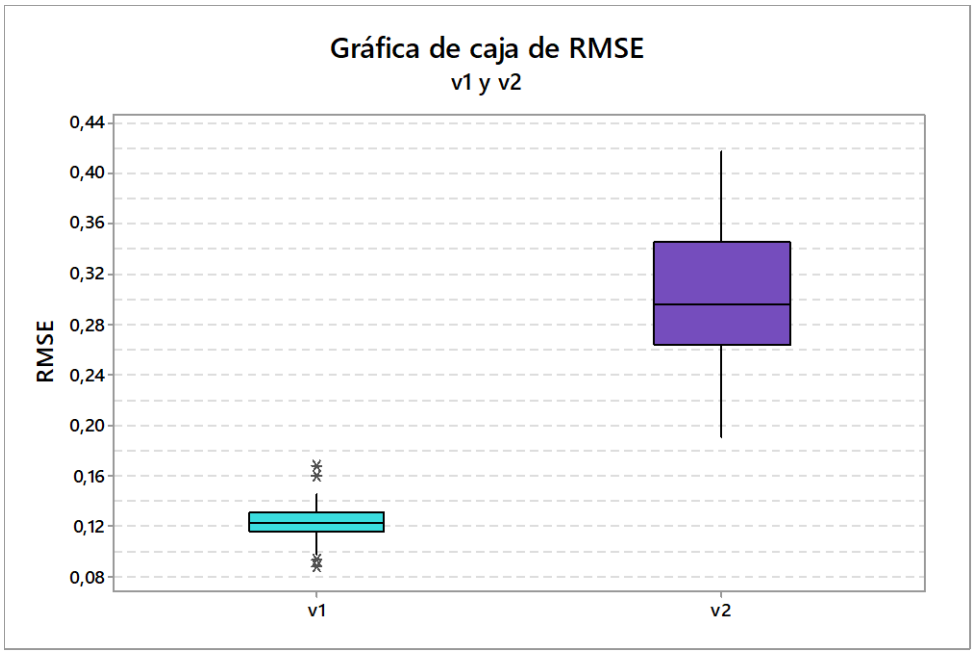


Figura 38: Gráfica de caja de RMSE para leyes v_1 y v_2 con KO.

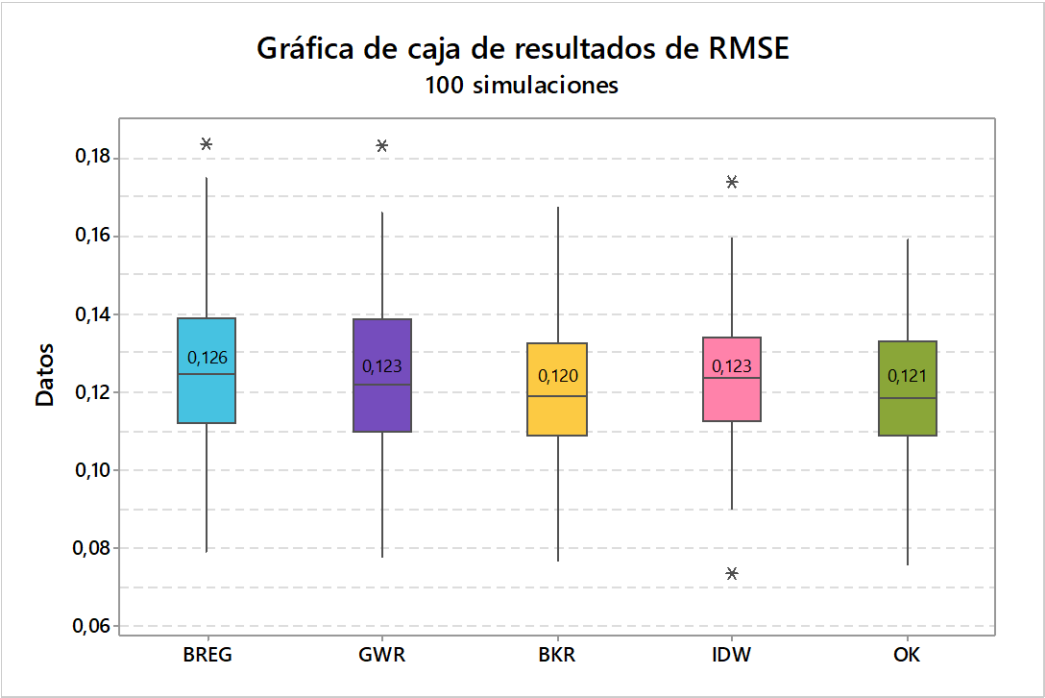


Figura 39: Gráfica de caja de RMSE de ley y para modelos de regresión.

De la **Figura 39** se concluye que los métodos de regresión utilizados tienen RMSE promedios parecidos, variando en milésimas, además, se concluye que para el caso simulado la mejor estimación de y se logra con BKR con un RMSE promedio de 0.120 seguido por OK con un RMSE promedio de 0.121, cabe destacar que GWR y IDW poseen el mismo RMSE promedio pero GWR tiene una mayor desviación estándar. D

Por otro lado, el método de regresión donde se obtienen el mayor RMSE es BREG con un RMSE promedio de 0.126.**Figura 39:** Gráfica de caja de RMSE de ley y para modelos de regresión.

4.5 CATEGORIZACIÓN DE RECURSOS MINERALES

El Kriging permite obtener, además de la estimación del valor de un bloque, una indicación de la precisión local a través de la varianza Kriging, producto de lo anterior, para poder hacer la categorización de recursos minerales se utiliza la metodología de sugerida por Diehl & David.

En la **Tabla 9** se resume las categorías de la clasificación, basadas en la cuantificación del error utilizando la desviación estándar Kriging.

Tabla 9: metodología de Diehl y David para clasificación de recursos minerales.

	Inferido	Indicado	Medido
Error (precisión)	± 60	± 40	± 20
confianza	$20 - 40\%$	$40 - 60\%$	$> 60\%$

El método propuesto por Diehl & David se basa en definir niveles de confianza y de precisión (error). La precisión se expresa en función de la desviación estándar Kriging y el valor estimado Kriging (ver Ec. 4.7).

$$Precision = \frac{\sigma_k * 100 * Z_{1-a}}{Z^*} \quad \text{Ec. 4.7}$$

Donde,

- σ_k: Desviación estándar de Kriging.
- Z^*: Valor del bloque estimado por Kriging.
- Z_{1-a}: Valor de la variable distribuida normalmente con un nivel de confianza $1 - a$.

Si se fija una precisión en 20% entonces se puede determinar la razón $\frac{\sigma_k}{Z^*}$ que divide las reservas medidas e indicadas y si se fija la precisión en 40% se puede obtener valor que divide las reservas indicadas con las inferidas. En la **Tabla 10** se muestran los límites establecidos según la metodología.

Tabla 10: Limites para la categorización de recursos.

	Inferido	Indicado	Medido
Error (precisión)	± 60	± 40	± 20
Límites	> 0.763	$0.763 - 0.238$	< 0.238

Producto que lo métodos de regresión empleados no cuentan con la varianza de Kriging, se asume que las varianzas son iguales a varianzas de Kriging ordinario, esto con el fin se establecer un criterio unificado para los métodos y así poder compararlos.

Se determina la categorización de recursos para las simulaciones 94, 40, 16 puesto que corresponden al mejor, peor y el más probable caso respectivamente, obteniendo los resultados resumidos en **Tabla 11** y visualizados en **Figura 40**, **Figura 41** y **Figura 42**.

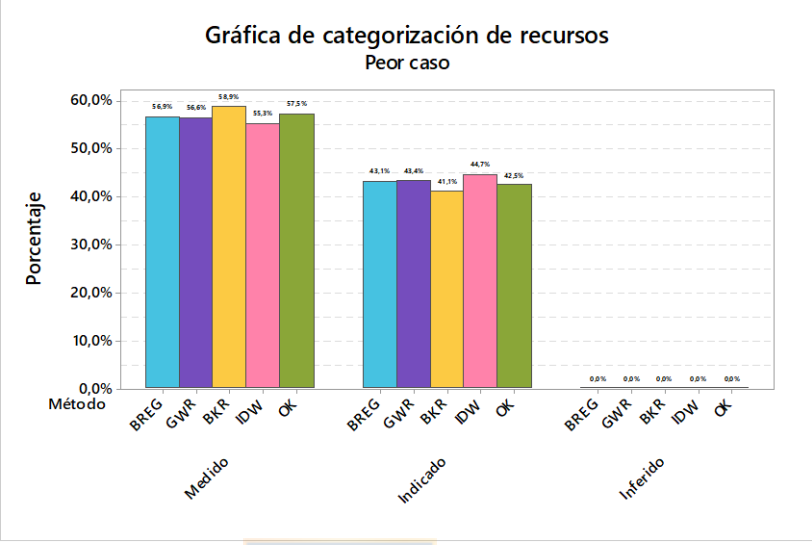


Figura 40: Categorización de recursos peor caso.

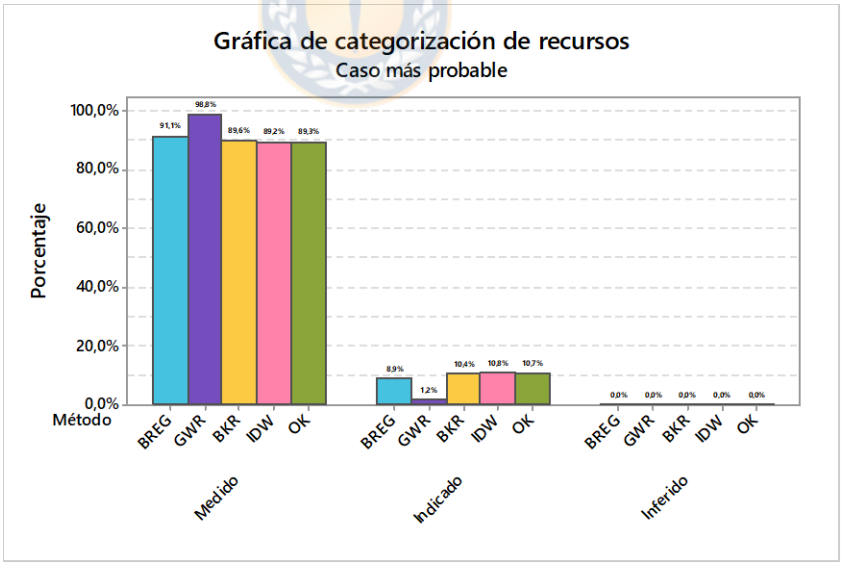


Figura 41: Categorización de recursos caso más probable.

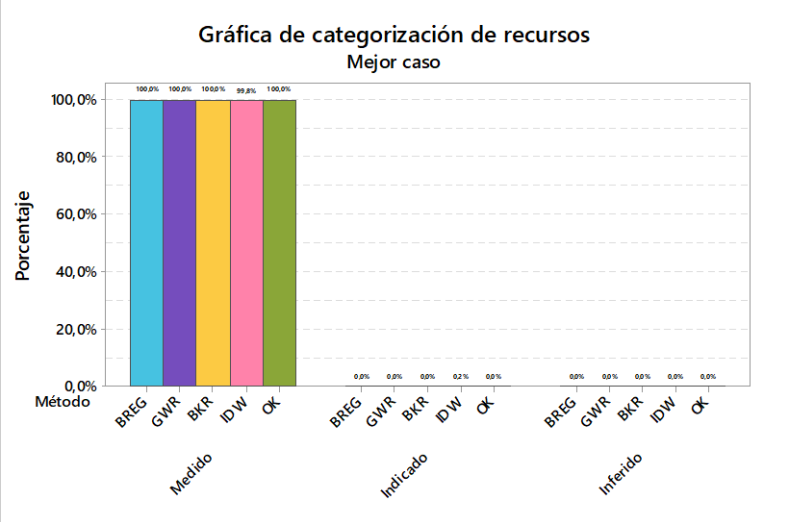


Figura 42: Categorización de recursos mejor caso.

Tabla 11: Categorización de recursos para el mejor, peor y más probable caso.

		Inferidos			Indicados			Medidos		
		[%]	[MTon]	$\bar{y}$	[%]	[MTon]	$\bar{y}$	[%]	[MTon]	$\bar{y}$
Peor	BREG	0	...	...	43.40	3.429	0.395	56.60	6.359	0.561
	GWR	0	...	...	43.09	3.485	0.404	56.91	6.278	0.551
	BKR	0	...	...	41.12	3.402	0.413	58.88	6.548	0.555
	IDW	0	...	...	44.66	3.723	0.427	55.34	6.118	0.552
	OK	0	...	...	42.50	3.662	0.430	57.50	6.405	0.556
Más Probable	BREG	0	...	...	1.60	0.142	0.445	98.4	11.580	0.589
	GWR	0	...	...	8.90	0.847	0.475	91.1	10.560	0.580
	BKR	0	...	...	10.35	1.015	0.488	89.65	10.323	0.572
	IDW	0	...	...	10.78	1.023	0.473	89.22	10.289	0.576
	OK	0	...	...	10.65	0.988	0.464	89.35	10.414	0.583
Mejor	BREG	0	...	...	0	...	...	100	11.302	0.565
	GWR	0	...	...	0	...	...	100	11.123	0.561
	BKR	0	...	...	0	...	...	100	11.281	0.564
	IDW	0	...	...	0.22	0.016	0.354	99.78	11.211	0.561
	OK	0	...	...	0	...	...	100	11.172	0.558

De la **Tabla 11** se concluye que en el peor caso en promedio se posee un 42% de recursos indicados y 58% de recursos medidos para todos los métodos; en el caso más probable en promedio se posee un 10% de recursos indicados y 90% de recursos medidos para todos los métodos, exceptuando BREG que posee 1.6% de recursos indicados y un 98.4% recursos medidos; en el mejor caso solo el método IDW posee recursos indicados con un 0.2% siendo el resto 79.8% recursos medidos, mientras que todos los demás poseen el 100% de recursos medidos.

4.6 ANÁLISIS DE RESULTADOS

Al desarrollar 100 simulaciones para realizar la estimación de la ley y para cada método de regresión permite ver tendencias de los resultados obtenidos, tanto en calidad de estimaciones como en costo computacional (tiempo de programación[6]), logrando obtener conclusiones tanto generales como particulares del comportamiento de los algoritmos.

En cuanto a las leyes estimadas v_1 y v_2 se escogió estimarlas con un método de regresión único, en este caso se escoge KO, para luego utilizar los mismos resultados de v_1 y v_2 para realizar la estimación de la ley y en cada punto, esta elección permitió, en un principio, tener un homogéneo arrastre del error de estimación por lo que facilita la comparación de los resultados de la estimación de la ley y por cada método.

En cuando a las leyes estimadas y cabe notar que:

- BREG en general tiene la tendencia a estimar leyes mayores en comparación al resto de los métodos, teniendo a una ley media de 0.586 en el caso más probable difiriendo de la ley media de la simulación que es 0.550, esto influye directamente en el RMSE causando ser el método de regresión menos efectivo de los cinco modelos probados. Además, al comparar los histogramas de distribución de la ley y en la simulación 100 (**Figura 15** y **Figura 24**) BREG es el que entrega la mayor diferencia de la distribución no pudiéndose acercar a la distribución normal del caso simulado ni logrando acercarse a la ley media.

[6] Tiempo en que se demora el algoritmo al entregar el resultado de ley y en un punto

- GWR se destaca por ser un método de regresión 2D que fue necesario adaptarlo para lograr hacer las estimaciones (ley y se distribuye en una malla 3D), esta adaptación se logra al separar el yacimiento en niveles, tanto verticales como horizontales. Entonces, al trabajar en el nivel horizontal se traen los valores de los niveles verticales respetando la distancia. Esto último afecta la calidad de la estimación de la ley y dado que el algoritmo diseñado no considera las distancias diagonales. Esto último producto del drástico aumento de costo computacional que conlleva considéralas.

  Además, se destaca que los anchos de banda que entregan los mejores resultados tienen a ser los que consideran un valor menor, lo cual hace sentido dado que se espera que el radio de búsqueda sea lo más acotado posible y tenga la mayor cantidad de vecinos, para lograr la estimación.

  Por último, en cuanto al histograma de distribución en la simulación 100 se aprecia que no resulta ser simétrica como lo es en al caso simulado y posee una cola derecha más alargada desplazando la media a valores más altos.
- BKR al igual que para el caso GWR es un algoritmo que está disponible en R solo en 2D, por lo que fue necesario hacer la misma adaptación por niveles que se describió en el punto anterior, esta misma adaptación puede afectar la calidad de la estimación en cada punto, a pesar de lo anterior BKR resulta ser uno de los mejores modelos de regresión para la estimación, superado de manera significativa los demás métodos propuestos y teniendo el menor RMSE promedio de 0.120 de los cinco métodos aunque por lado BKR es el métodos de regresión que ocupa el mayor costo computacional entre todos los métodos.

  En cuanto al histograma de la simulación 100 BKR es el método que más tiene a acercarse a la distribución normal de la simulación 100.
- IDW por su lado entregó resultados de estimación parecidos a GWR con la diferencia que IDW se tuvieron que probar una mayor cantidad de combinaciones posibles entre el coeficiente de ponderación y la cantidad de vecinos máximos utilizados para realizar la estimación, teniendo tiempos computacionales mayores que GWR. además, IDW al contrario que BREG tiene a entregar las menores leyes teniendo una ley promedio de 0.523 en el caso mas probable difiriendo de la ley de la simulación que es 0.550.

- OK en este caso es el método que entrega las mejores estimaciones para el caso simulado, ganándole al resto de los métodos en un 39% y además tiene menor costo computacional en comparación al BKR, esto producto que para OK solo se ajusta el variograma con la función *autofitVariogram* mientras que BKR además de ajustar el variograma hace variar la meseta el rango y la pepa a través de las funciones del código generando funciones de probabilidad para cada una de ellas.

Al realizar un análisis por cada método frente al resto se concluye que:

- BREG no logra ganarle a ningún método más del 50% de las veces y la mayor cantidad de veces que gana es frente al OK ganándole 41% y la menor cantidad de veces que gana es frente al BKR con un 12%.
- GWR logra ganarle al BREG un 85% de las veces y el caso en que menos gana es frente al BKR con un 28%.
- BKR logra ganarles a todos los métodos con un mínimo del 50% de las veces, el caso en que más gana es frente al BREG con un 88% de las veces y en la que menos ganas es frente al OK con un 50%.
- IDW las veces que más gana es frente al BREG con un 64% de las veces y en la que menos gana es 36% frente al BKR.
- OK al igual que BKR logra ganarles a todos los métodos con un mínimo de 50% y el caso en el que más ganas es frente al BREG con un 59% de las veces y en caos que menos gana es contra BKR con un 50%.

En la **Tabla 12** el detalle de lo descrito anteriormente, donde muestra la matriz de comparación para cada método de regresión frente al resto.

Tabla 12: Matriz de comparación entre métodos de regresión.

	Porcentaje de veces en que es mejor				
Método	BREG	GWR	BKR	IDW	OK
BREG	---	85%	88%	64%	59%
GWR	15%	---	72%	52%	54%

BKR	12%	28%	---	36%	50%
IDW	36%	48%	64%	---	56%
OK	41%	46%	50%	44%	---

Al analizar los resultados de RMSE para todos caso, se aprecia que en general el método de regresión que posee los mejores resultados es OK, teniendo un porcentaje de veces que supera a todo el resto de los métodos de 39%, siguiendo el método BKR con un porcentaje de 31%, luego IDW y GWR con porcentajes 15% y 13% respectivamente, y por último el peor método para el caso simulado es BREG con solo un 2% de veces que supera al resto de los métodos. Lo anterior se muestra en la **Tabla 13**.

Tabla 13: Porcentajes de veces en que cada método de regresión supera al resto.

Método	[%] de veces que es mejor	RMSE promedio
BREG	2	0.126
GWR	13	0.123
BKR	31	0.120
IDW	15	0.123
OK	39	0.121

Además, se destaca que para ningún caso se encuentran recursos inferidos en las estimaciones realizadas, lo cual tiene sentido dado que al construir el caso simulado y al escoger los sondajes esto presentan un distanciamiento homogéneo y bajo con un distanciamiento de $10\ [m]$ y apresar que la base con sondajes representa el 3.76% de toda la base, esta cantidad de información es lo suficiente para poder realizar un ERM de buena calidad.

Por último, cabe destacar la disminución de la calidad de la estimación a mayores profundidades, lo que implica tener mayores radios de búsqueda de vecinos, esto producto de la menor cantidad de información disponibles de los sondajes para hacer la estimación.

CAPÍTULO 5

CASO REAL

5.1 DESCRIPCIÓN DE LA BASE DE DATOS

La base datos real corresponde a una campaña de sondaje para un yacimiento de hierro, la cual está compuesta de 660 muestras recogidas de un total de 11 sondajes; para cada muestra se rescatan el atributo continuo ley de Fe y los atributos discretos tipo de roca y textura de roca, los cuales se muestran en la Tabla 14.

Tabla 14: Tipos de rocas y texturas en base real.

	Nombre	Código
Tipo de textura	Porfídica	POR
	Afanítica	AFA
	Maciza	MAC
	Brechosa	BRE
Tipo de roca	Metandesitas	MET
	Hierro de mena	HIE
	Brechas	BRH
	Andesita	AND
	Dioritas	DIO

Los sondajes desplegados son de diferentes largos y poseen un manteo mayoritariamente hacia el sur por lo que se escoge el área de trabajo con las dimensiones $1360[m]$ hacia el norte, 740 $[m]$ al este y una profundidad de 550 $[m]$, las que corresponden a las distancias máximas con información, abarcando toda el área de la campaña de sondaje.

5.1.1 DISTRIBUCIÓN DE LEY DE HIERRO

En la **Figura 43** se encuentra una visualización del despliegue de los sondajes respecto al atributo continuo ley de Fe, donde se puede apreciar que la mayoría de la campaña de sondajes se distribuye en leyes bajas de hierro, esto se confirma al analizar el histograma de la **Figura 44** que tiene una distribución asimétrica sesgada positivamente, ajustándose a una distribución lognormal, en la cual la moda se encuentra en la clase 24% de ley de Fe.

Por otro lado, se observa que los sondajes más externos tienen distribución de leyes de hierro más bien homogéneas con leyes bajas, caso contrario a los sondajes del centro.

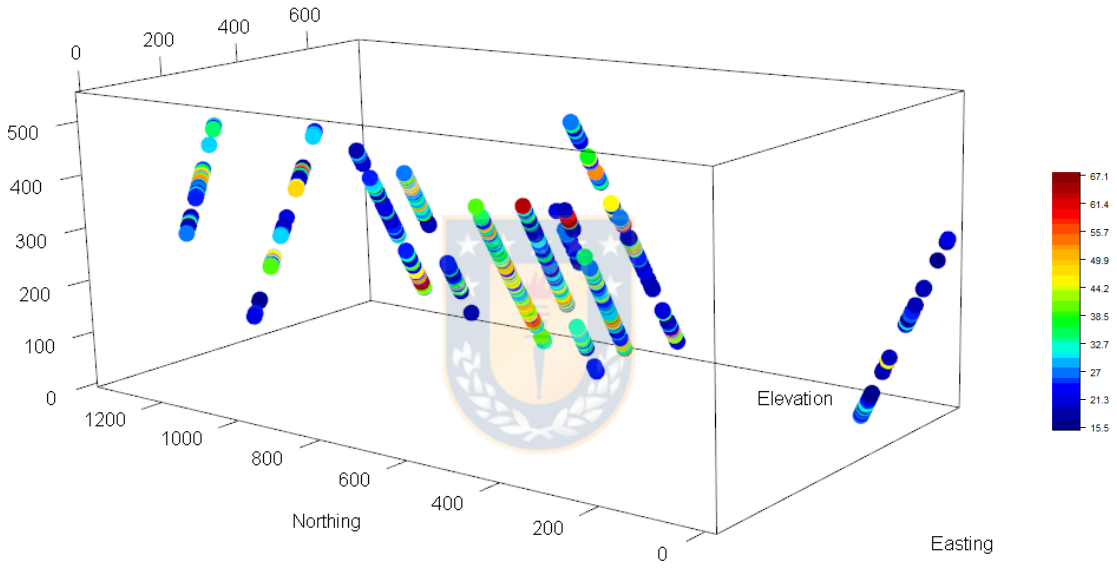


Figura 43: Visualización de Fe en sondajes.

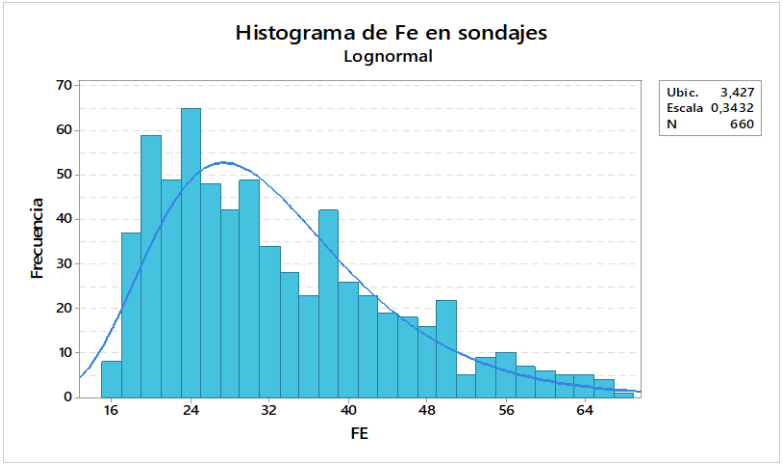


Figura 44: Histograma de distribución de ley de Fe en sondajes.

En la **Tabla 15** se entrega la estadística descriptiva para la distribución de la ley de Fe en los sondajes, la cual muestra que la ley media de Fe para los sondajes es de 32.7% y los valores extremos son 15.5% y 67.1%, mínimo y máximo respectivamente.

Tabla 15: Estadística descriptiva de ley de Fe en sondajes.

Característica	Valor
Distribución	Lognormal
Cuenta	660
Mínimo	15.53
Máximo	67.13
Media	32.70
Moda	38.66
Desviación estándar	11.66
Coeficiente de asimetría	0.816

5.1.2 DISTRIBUCIÓN DE TIPO DE ROCA

En la **Figura 45** se visualiza el despliegue de los sondajes respecto al atributo discreto tipo de roca (MET, HIE, BRH, AND, DIO), en donde se aprecia el amplio domino del tipo de roca BRH y HIE, además se aprecian dos casos particulares de sondajes, uno con una fuerte presencia del tipo de roca MET, cabe decir que exceptuando este sondaje en particular el tipo de roca MET no se encuentra en otro lugar de la campaña de sondajes, y otro sondaje que se encuentra ubicado de manera más externa que tiene una fuerte presencia del tipo de toca DIO, este último tipo de roca solo se encuentra en el sondaje ya señalado y en el otro sondaje que se encuentra en el extremo opuesto. Por último, el tipo de roca AND se encuentra distribuido mayoritariamente de manera homogénea en toda la campaña de sondajes.

Todo lo anterior se confirma con el gráfico circular presentado en la **Figura 46** en la cual se concluye que el 75.2% de la campaña de sondaje lo constituyen los tipos de roca BRH y HIE, mientras que solo 7.1% se compone de los tipos de roca MET y DIO.

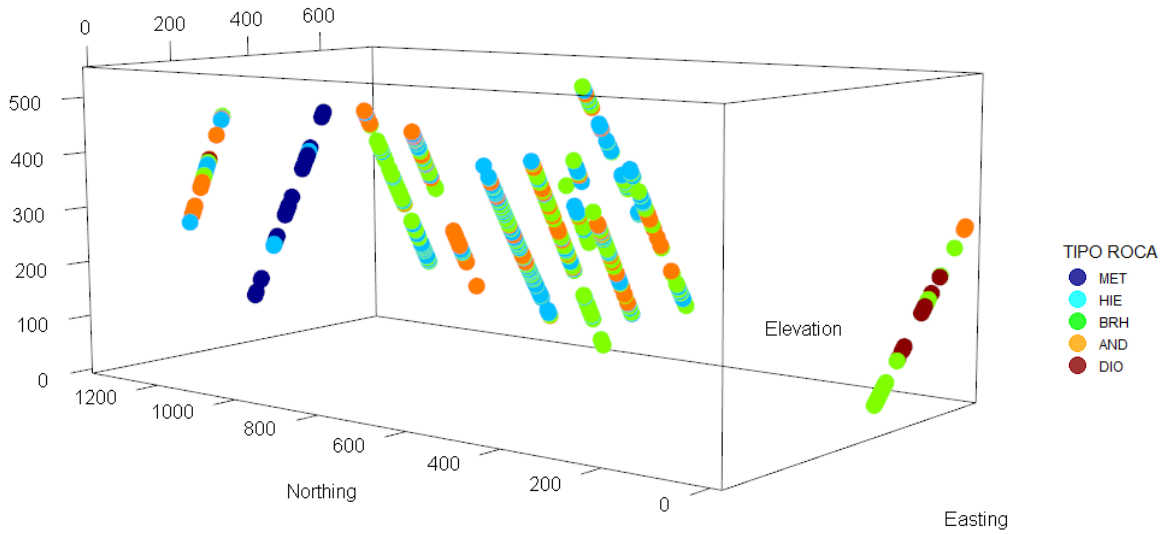


Figura 45: Visualización de tipo de roca en sondajes.

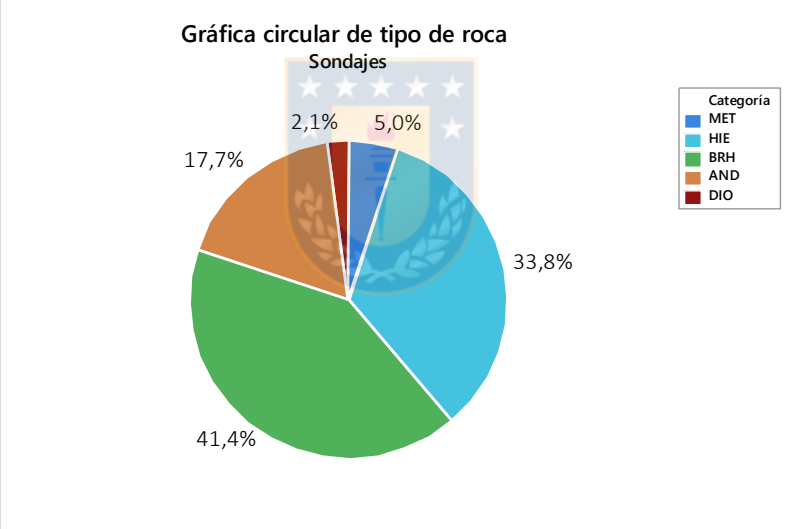


Figura 46: Gráfico circular de tipo de roca en sondajes.

A continuacion en la **Tabla 16** se presentan los resultados del analisis de estadistica descriptiva que tiene la ley de hierro según el tipo de roca, y en la **Figura 47** se muestran los resultados anteriores en un grafico de caja para mejor vizualización.

Como es de esperarse las mayores concentraciones de hierro se encuentran en el tipo de roca HIE, mena de hierro, con una ley promedio de 43.36%, mientras que el resto de los tipo de roca su ley varia entorno a un promedio de 26%.

Tabla 16: Ley de Fe según tipo de roca en sondajes.

		Tipo de roca				
		MET	HIE	BRH	AND	DIO
Ley de Fe [%]	Cuenta	33	223	273	117	14
	Media	26.52	43.36	28.49	24.99	23.91
	Moda	24.98	50.17	23.62	20.24	---
	Mínimo	18.00	18.38	15.53	15.68	15.53
	Máximo	47.81	67.13	50.17	49.17	45.88
	σ	8.48	11.09	7.45	6.18	9.01

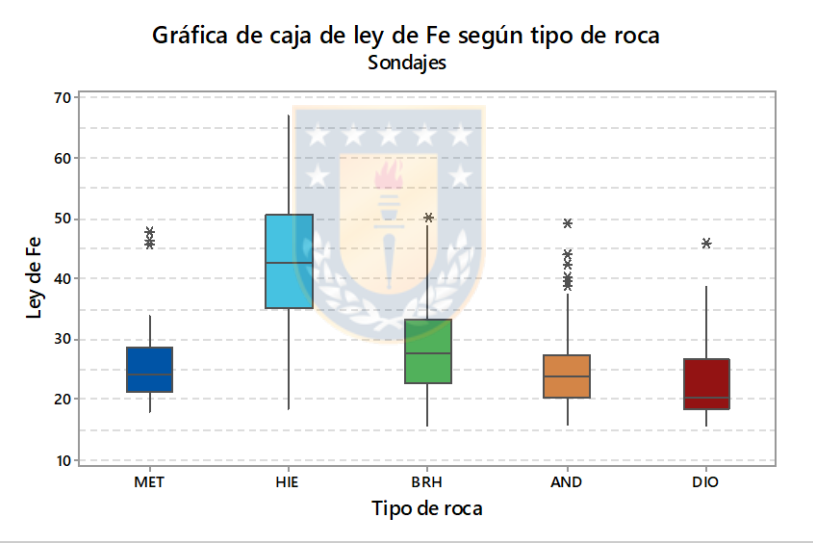


Figura 47: Gráfico de caja de ley de Fe según tipo de roca en sondajes.

5.1.3 DISTRIBUCIÓN DE TEXTURA DE ROCA

En la **Figura 48** se visualiza el despliegue de los sondajes respecto al atributo discreto textura de roca (POR, AFA MAC, BRE), en donde se aprecia el amplio domino de la textura de roca BRE en toda la campaña, además se destaca el caso de la textura MAC producto que solo está presente de manera muy aislada en dos sondajes; en cuanto a la textura AFA se visualiza una fuerte presencia en solo un sondaje mientras que la textura AFA se distingue una sutil presencia en casi todos los sondajes.

Lo descrito anteriormente se reafirma con el grafico circular presentado en la **Figura 49** en el cual se concluye que el 78% de la campaña de sondaje la compone la textura BRE, mientras que solo 5.5% lo compone las texturas AFA y MAC.

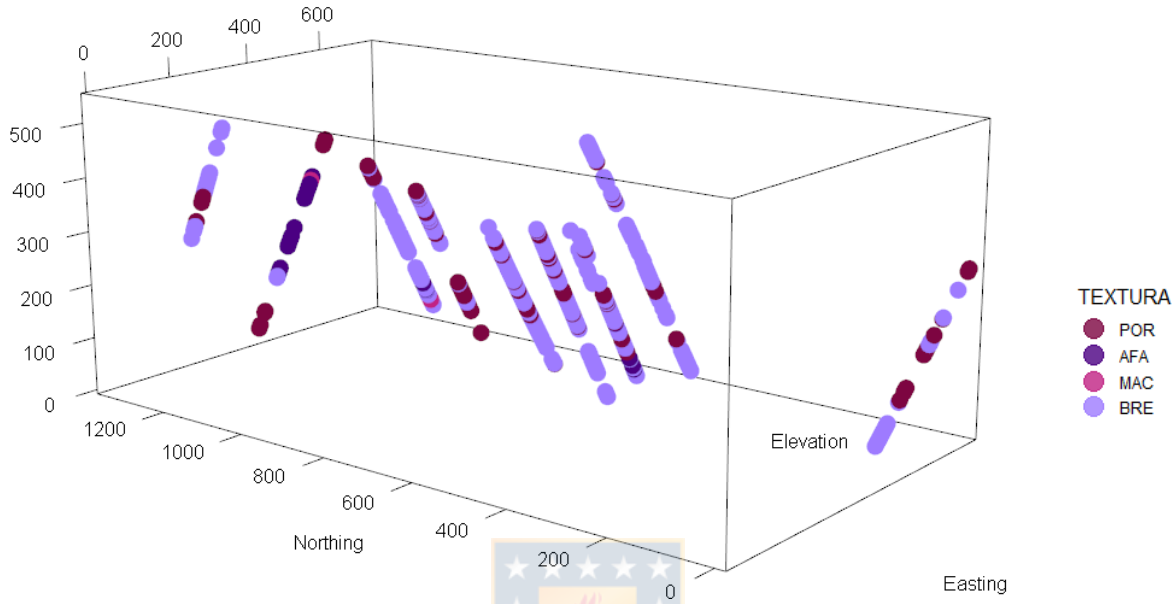


Figura 48: Visvisualización de textura en sondajes.

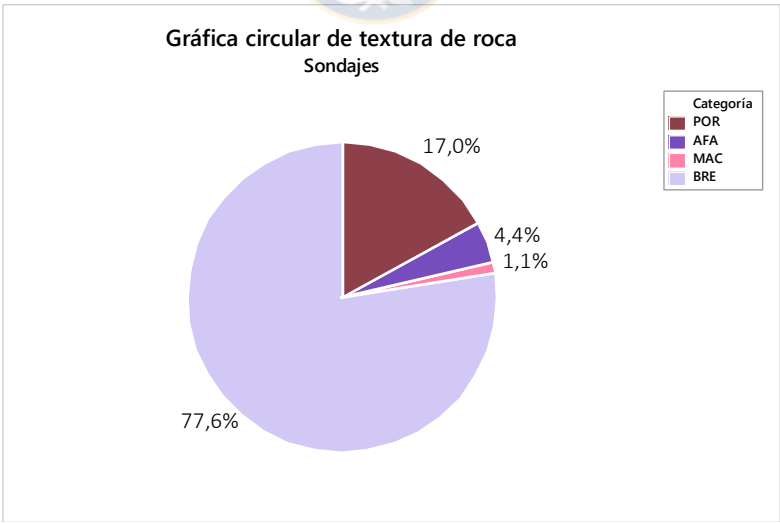


Figura 49: Gráfico circular de la textura de roca en sondajes.

A continuacion en la **Tabla 17** se presentan los resultados del analisis de la estadistica descriptiva que tiene la ley de hierro según textura de roca, y en la **Figura 49** se muestran los resultados anteriores en un grafico de caja para mejor vizualización.

Tiene sentido pensar que la textura MAC tendría asociadas las mayores leyes de hierro producto de la relación directa que posee la ley de hierro y la densidad de la roca, en este caso en particular se refleja lo anterior dado que para la textura MAC posee una ley media de hierro de 59.5%, pero con poca cantidad de muestras.

Por otro lado, la textura BRE tiene un promedio de ley de 34% una mayor desviación estándar con respecto al de las texturas POR y AFA, estas últimas poseen un promedio total de ley de hierro del 26.1%.

Tabla 17: Ley de Fe según textura de roca en sondajes.

		Textura de roca			
		POR	AFA	MAC	BRE
Ley de Fe [%]	Cuenta	112	29	7	512
	Media	24.90	27.34	59.49	34.34
	Moda	23.36	24.98	---	38.66
	Mínimo	15.53	18.00	35.27	15.53
	Máximo	49.17	47.81	67.13	66.04
	σ	6.50	9.37	11.21	11.51

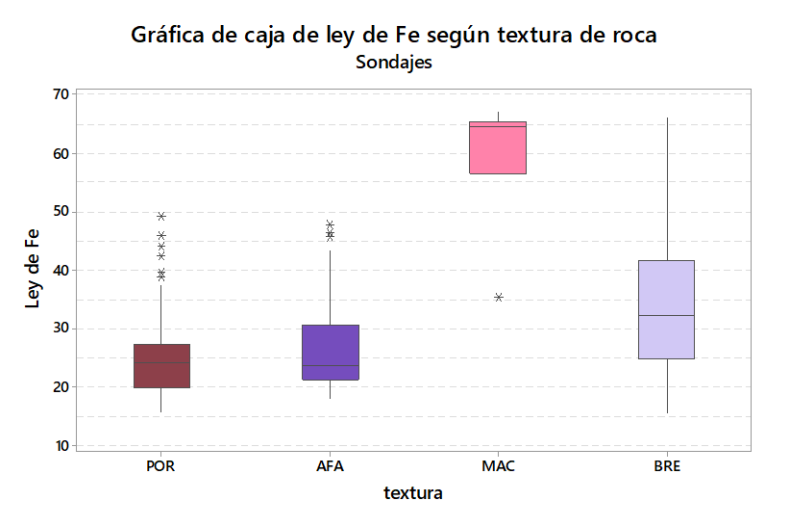


Figura 50: Gráfico de caja de ley de Fe según textura de roca en sondajes.

5.2 ESTIMACIÓN DE ATRIBUTOS DISCRETOS PARA YACIMIENTO

El objetivo de la memoria es determinar la eficiencia de ciertos métodos de regresión a través de la simulación de un atributo continúo condicionado de atributos conocidos, en este caso el atributo continuo corresponde a la ley de hierro y los atributos condicionados con el tipo de roca y la textura de roca.

Producto que la textura y el tipo de roca son atributos discretos y los métodos de regresión utilizados en esta Memoria de título son para la estimación de atributos continuos se utilizan los resultados obtenidos de la Memoria de título de Valentina Neira [40], quien utiliza el método co-Kriging indicador para hacer la estimación de textura y tipo de roca para el mismo yacimiento de hierro, obteniendo así los siguientes resultados.

5.2.1 CO-KRIGING INDICADOR PARA ESTIMACIÓN DE TIPO DE ROCA

En la **Figura 51** se visualiza el resultado obtenido al estimar el tipo de roca con co-Kriging indicador, el cual entrega un resultado de 63313 puntos, estimado un total de 62653 puntos, mientras que el resto de los puntos corresponden a los 600 puntos obtenidos de la campaña de sondaje. Cada uno de estos puntos representa el centroide de los bloques del modelo de bloques a construir, y que tienen dimensiones de $20 \times 20 \times 20\ [m^3]$ cada uno.

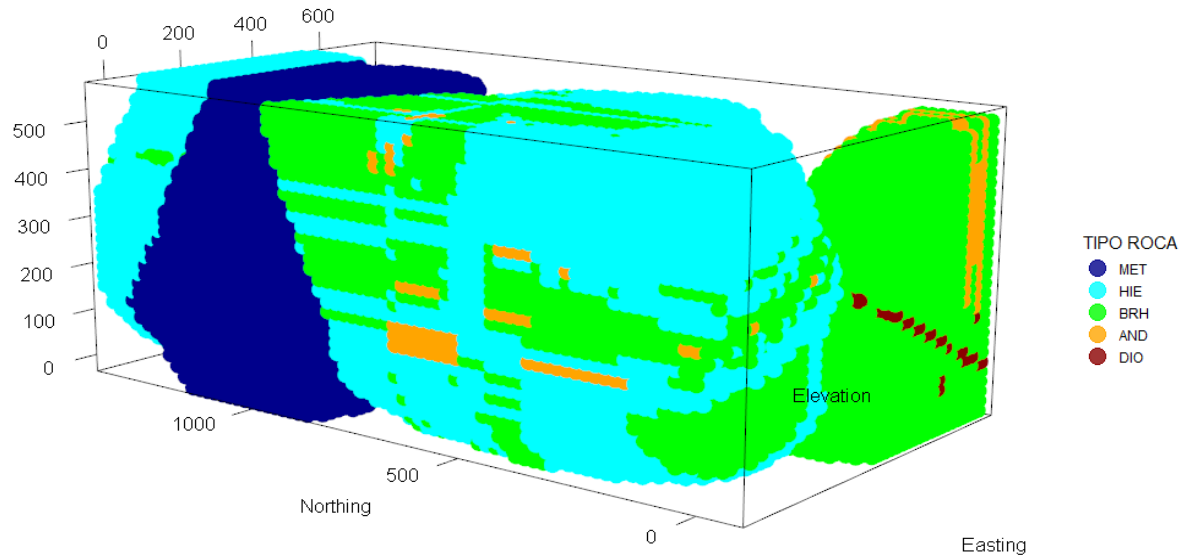


Figura 51: Estimación por co-Kriging indicador para el tipo de roca en yacimiento.

En la **Figura 52** se encuentra la gráfica circular de los resultados donde se concluye que el yacimiento se compone mayoritariamente del tipo de roca HIE y BRH con un 75.6%, que son valores similares a los obtenidos de los sondajes.

Una gran diferencia entre los valores obtenidos de los sondajes y el obtenido del yacimiento se presenta en el tipo de roca MET que de un 5% aumenta a un 21.1%, también en el tipo de roca AND que disminuye de un 17.7% a un 2.4%, por último, la presencia del tipo de roca DIO se mantiene parecido con respecto a los sondajes.

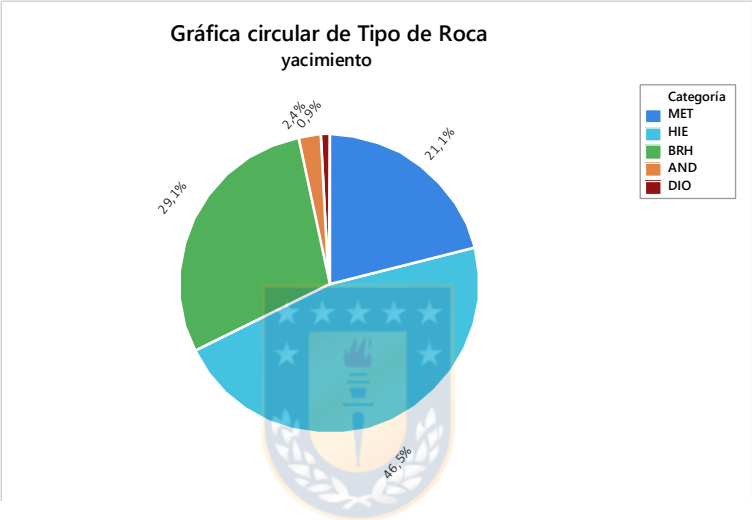


Figura 52: Gráfico circular de tipo de roca en estimación de yacimiento por co-kriging.

5.2.2 Co-Kriging indicador para estimación de textura de roca

En la **Figura 53** se visualiza el resultado obtenido al estimar la textura de roca con co-Kriging indicador, entregando la misma cantidad de estimaciones que el caso anterior.

En la **Figura 54** se encuentra la gráfica circular de los resultados donde se concluye que el yacimiento se compone en gran medida de las texturas BRE y POR con un total de 90.2% entre ambas texturas, aumentando de un 17.0% a 41.8% la textura POR y disminuyendo de un 77.6% a un 41.8% la textura BRE respecto a la distribución entregada por los sondajes.

Por otro lado, las texturas AFA y MAC componen al yacimiento en un 9.9% manteniéndose parecido en relación con lo entregado por los sondajes.

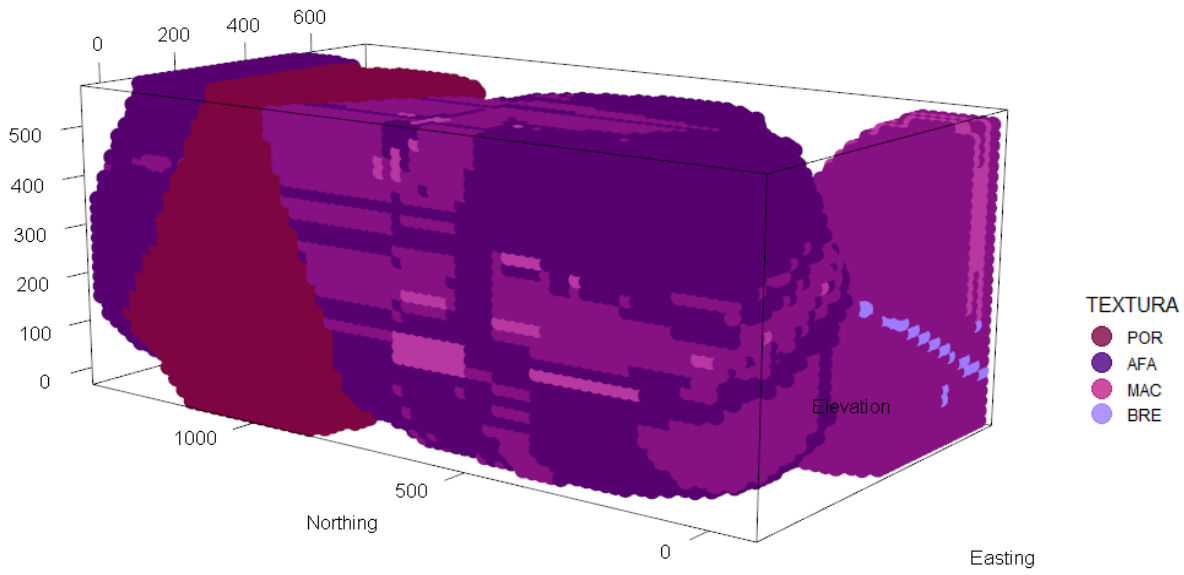


Figura 53: Estimación por co-Kriging indicador para textura de roca en yacimiento.

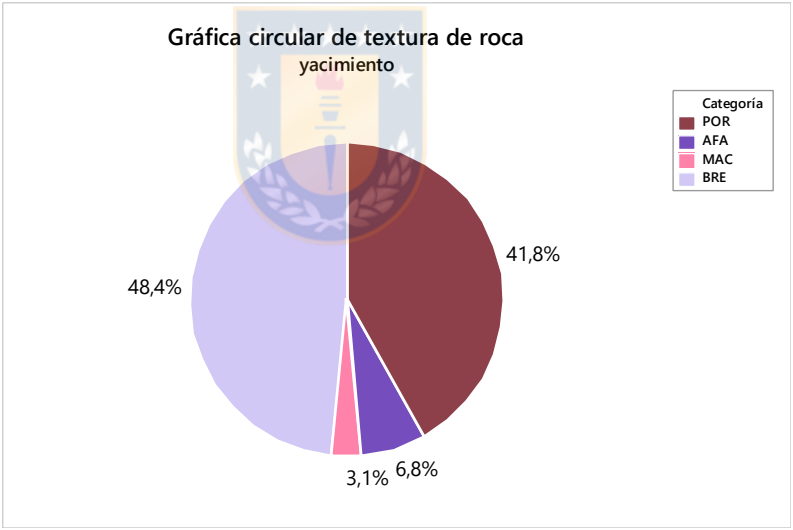


Figura 54: Gráfico circular de textura de roca en estimación de yacimiento por co-kriging.

5.3 ESTIMACIÓN DE ATRIBUTO CONTINUO PARA YACIMIENTO

Tomando los resultados antes descritos para el tipo de roca, textura del yacimiento y la información proveniente de la campaña de sondajes, se procede a hacer la estimación de la ley de Fe, desde ahora y_{Fe}, con los métodos propuestos, BREG, BK, GWR, además se hace una comparación con los métodos tradicionales, en este caso KO y IDW.

Cabe destacar que antes de aplicar el método de regresión en la estimación de y_{Fe} en el yacimiento, primero es necesario obtener las variables que optimizan los algoritmos a través de la optimización de la base con sondajes, esto se logra utilizando el método de validación cruzada uno a uno, es decir, a través de los 660 puntos muestreados de la campaña de sondajes se extraerá uno y a través de los 559 puntos restantes se estima el valor del punto extraído, todo esto con el fin de probar todas las combinaciones posibles de variables y elegir la más optima.

5.3.1 Regresión Beta

Como se nombra anteriormente para hacer uso del método de regresión para la estimación de la y_{Fe} primero se deben encontrar las variables que optimizan el algoritmo, en este caso se utiliza el método validación cruzada uno a uno sobre base de sondajes obtenidas del muestreo para encontrar estas variables.

En consecuencia a lo anterior, se parte estandarizando la base a trabajar dado que es requisito para el método de regresión que y_{Fe} se distribuya en el intervalo estándar $(0, 1)$ y producto que la base de sondajes no cumple con esta condición se utiliza Ec. 2.3 para estandarizar, obteniéndose los resultados mostrados en la Tabla 18.

Tabla 18: Estadística descriptiva de estandarización de y_{Fe} en base con sondajes.

		Base con sondajes	
	Característica	original	estandarizada
Ley de Fe	Cuenta	660	660
	Media	32.70	0.33
	Moda	38.66	0.45
	Mínimo	15.53	0.00
	Máximo	67.13	1.00
	Σ	11.66	0.23

Para la optimización del algoritmo se analizan combinaciones entre *Link* y *Link.phi* (ver anexo 8.19). Para la función *Link* se prueban las funciones *"logit", "probit", "cloglog", "cauchit", "log", "loglog"* y para los caracteres *link.phi* se prueban *"identity", "log" y "sqrt"*.

Lo anterior se traduce en analizar un total de 18 combinaciones, la elección de estas variables se realiza por medio del análisis del $RMSE$ (ver **Figura 55**), en el cual se concluye que la función *Link* y el carácter *Link.phi* que optimizan el código son *"cauchit",* y *"identity"* respectivamente.

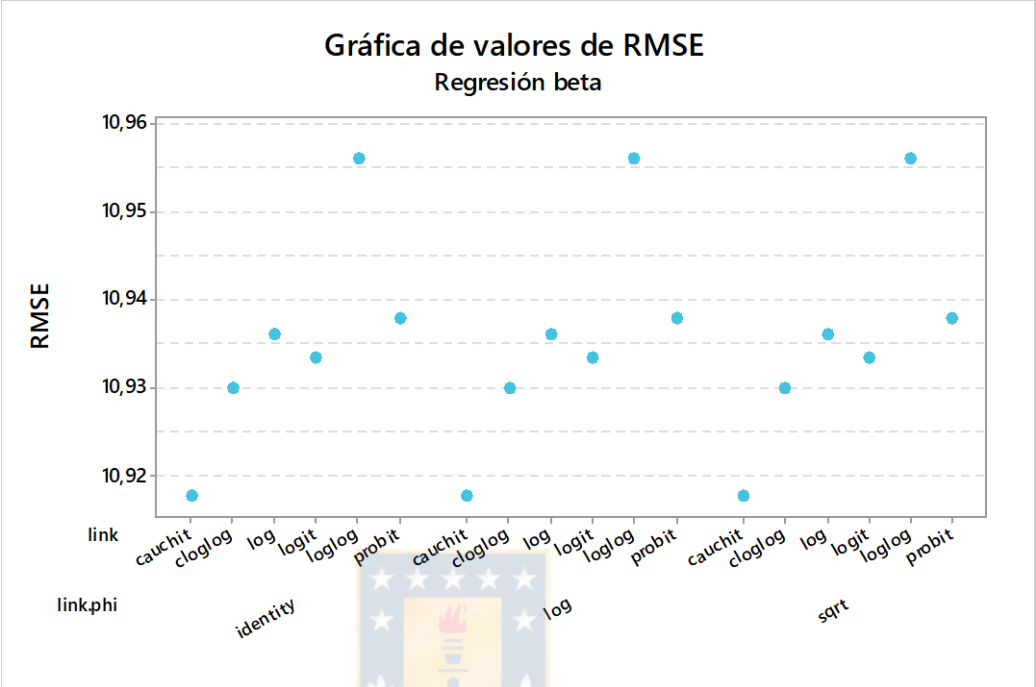


Figura 55: Resultados de RMSE para combinaciones de *Link* y *link.phi.*

Utilizando las variables escogidas y la base de sondajes estandarizados se procede a realizar la estimación de y_{Fe} con el modelo de la forma $y \sim Northing + Easting + Elevation + Tipo_{Roca} + Textura$. Por último, dado que los resultados entregados por el código de la Regresión Beta estarán dentro del intervalo $(0\ ,1)$ se aplican a los datos el inverso de la ecuación Ec 2.3 y así llevarlos a la escala original.

En la **Figura 56** se visualiza el resultado de la estimación, donde se aprecia que la distribución de y_{Fe} se condiciona fuertemente a la distribución que posee el tipo y la textura de roca siguiendo los patrones, también se aprecia la gran cantidad de leyes más bien bajas en todo el yacimiento, lo anterior se confirma con el histograma de y_{Fe} presentado en la **Figura 57**, que muestra un ajuste a la distribución lognormal con dos modas, una en la clase $y_{Fe} = 28\%$ y una segunda clase en $y_{Fe} = 21\%$.

Las y_{Fe} máximas y mínimas del yacimiento son 55.0% y 19.7% respectivamente, una media de 32.1% y una desviación estándar de 6.56, que comparado con la distribución de y_{Fe} de los sondajes la media es parecida pero las leyes máximas difieren, siendo mucho menores las estimadas por BREG que las entregadas por los sondajes.

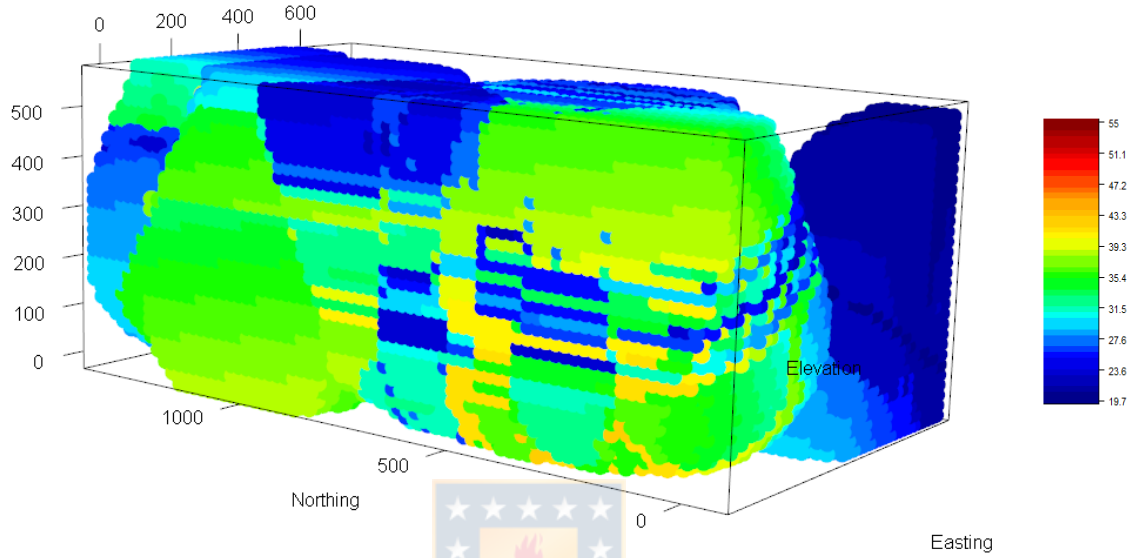


Figura 56: Visualización de estimación y_{Fe} por Regresión Beta en yacimiento.

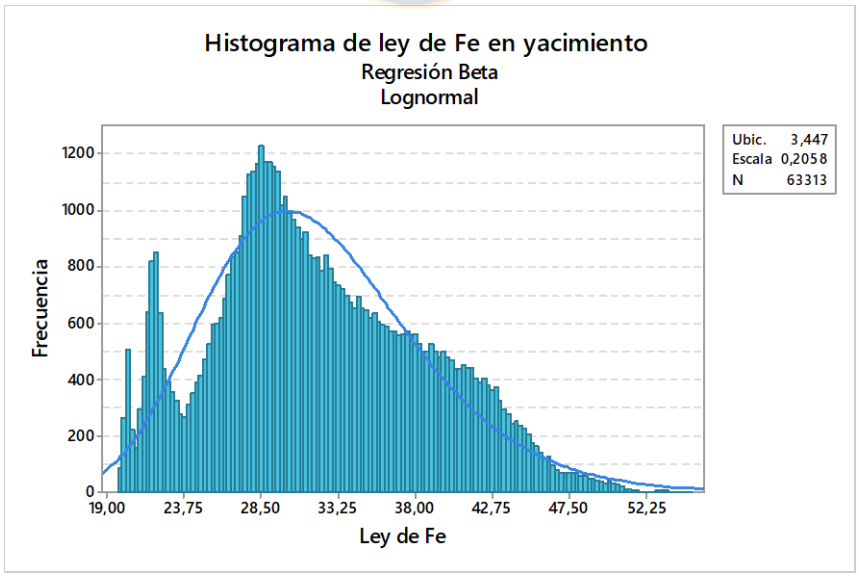


Figura 57: Histograma de estimación y_{Fe} por Regresión Beta en yacimiento.

5.3.2 REGRESIÓN GEOGRÁFICAMENTE PONDERADA

La optimización de parámetros con validación cruzada se comienza, al Igual que el procedimiento del caso simulado, realizando la división de los sondajes por niveles, dejando un total de 31 niveles diferentes que corresponden a cotas de los sondajes.

Para optimizar el código (ver anexo 8.20) se establece que se utiliza GWR con un ancho de banda variable, para cada iteración se prueban 50 tipos de ancho de bandas diferentes, la cantidad de vecinos a buscar para hacer la estimación varía entre 5% y el 50% del total de los datos presentes en la base de sondajes; los anchos de banda tienen como mínimo 33 vecinos y máximo 330 vecinos, aumentando en promedio 6 vecinos. Por último, se prueban 4 tipos de *kernels* diferentes, que son: *"gaussian", "exponential", "bisquare", "tricube", "boxcar"*. Probando un total de 250 combinaciones para el modelo.

En la **Figura 58** se muestran los resultados obtenidos en las 250 combinaciones para la optimización, donde se concluye que el menor $\mathbf{RMSE}$, con un valor 10.4, resulta con un ancho de banda 87 y un *Kernel* exponencial.

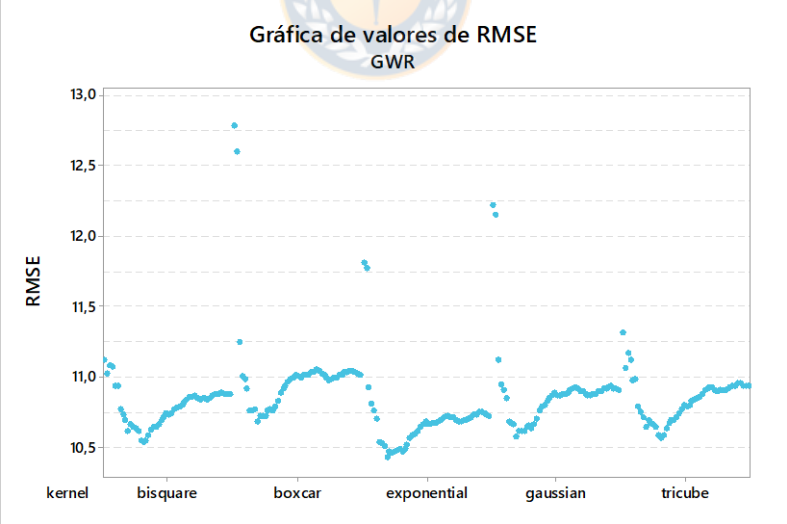


Figura 58: Resultados de RMSE para combinaciones de ancho de banda y *Kernel*.

Utilizando las variables escogidas y la base de sondajes separada por niveles se procede a realizar la estimación de y_{Fe} con el modelo de la forma $y \sim Northing + Easting + Elevation + Tipo_{Roca} + Textura$. Luego de hacer la estimación se vuelven a juntar los niveles y se construye el modelo de bloques.

En la **Figura 59** se visualiza el resultado de la estimación, donde, al igual que el caso con regresión beta, la distribución de y_{Fe} se condiciona fuertemente a la distribución que posee el tipo y la textura de roca siguiendo los patrones, también se aprecia la gran cantidad de leyes bajas en todo el yacimiento, lo anterior se confirma con el histograma de y_{Fe} presentado en la **Figura 60**, que muestra un ajuste a la distribución log-logística que a diferencia del log-normal, su función de distribución acumulativa se puede escribir en forma cerrada, con una menor desviación estándar respecto a Regresión Beta.

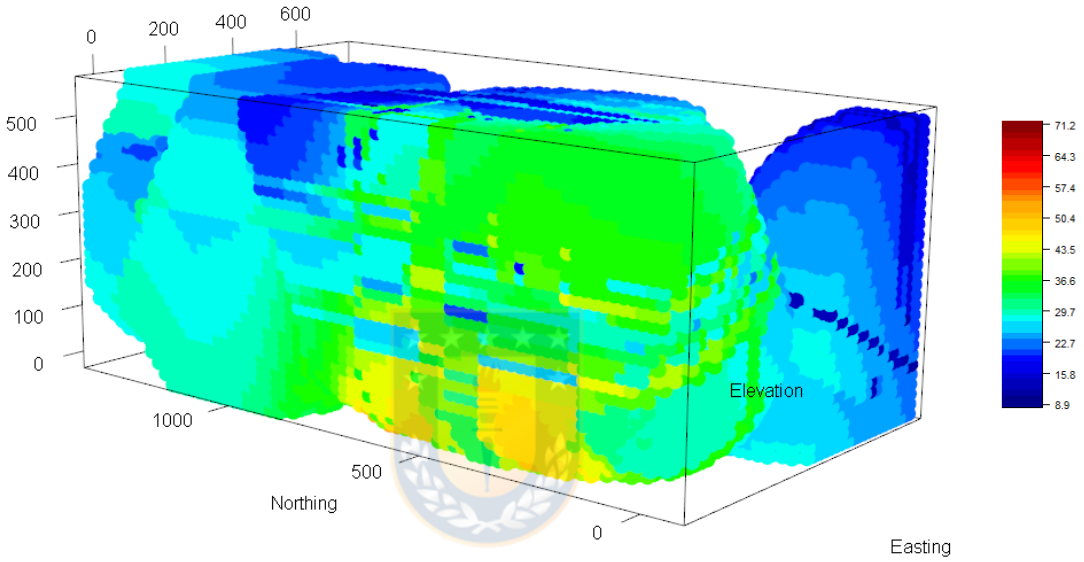


Figura 59: Visualización de estimación y_{Fe} por GWR en yacimiento.

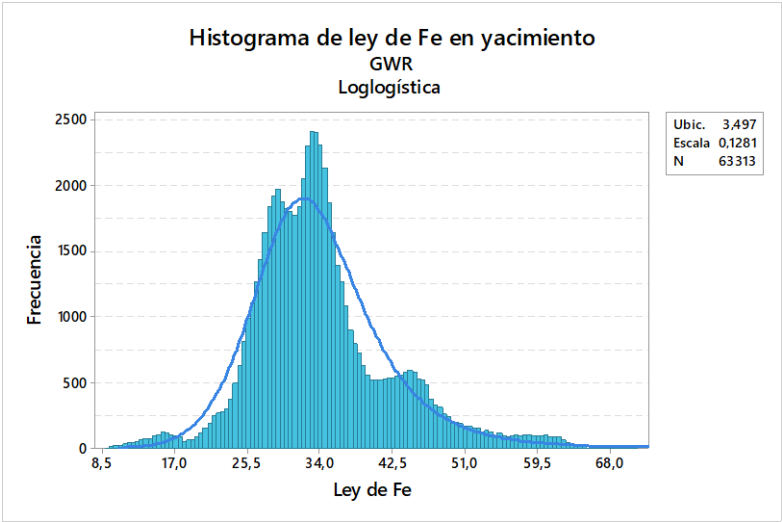


Figura 60: Histograma de estimación y_{Fe} por GWR en yacimiento.

Las y_{Fe} máximas y mínimas del yacimiento son 71.2% y 8.8% respectivamente, una media de 34.0% y una desviación estándar de 8.2, teniendo un parecido mayor a la distribución de y_{Fe} en los sondajes.

5.3.3 KRIGING BAYESIANO

Al igual que el caso GWR para hacer la validación cruzada es necesario hacer una separación por niveles de la base con sondajes dejando un total de 31 niveles diferentes.

Para optimizar el código se utilizan una cantidad de vecinos de 70, para cada nivel el *"model"* se utiliza una tendencia (*trend*) de la forma $trend.spatial \sim Elevation + Tipo\ de\ roca + textura$.

Para el caso del *"prior"* se establecen los parámetros del variograma, en donde los valores fijos están dados por el variograma autoajustado de la base con sondajes de la forma $Fe \sim Tipo\ de\ roca + Textura$ en el cual la meseta alcanza la meseta obtenida del variograma ajustado, el rango varía desde 0 con un largo 12 y aumentado a 12.2 del valor del variograma autoajustado y la pepa varía desde 0 con un largo 12 y aumentando a 0.06 del valor del variograma autoajustado y los *priors* para los variogramas, que corresponden a las distribuciones de la meseta, rango y pepa están dados por "reciprocal", "uniform" y "uniform" respectivamente. Con estos resultados se obtiene un RMSE de 10.10.

En la Figura 61 y Figura 62 se muestran el resultado de lo anterior.

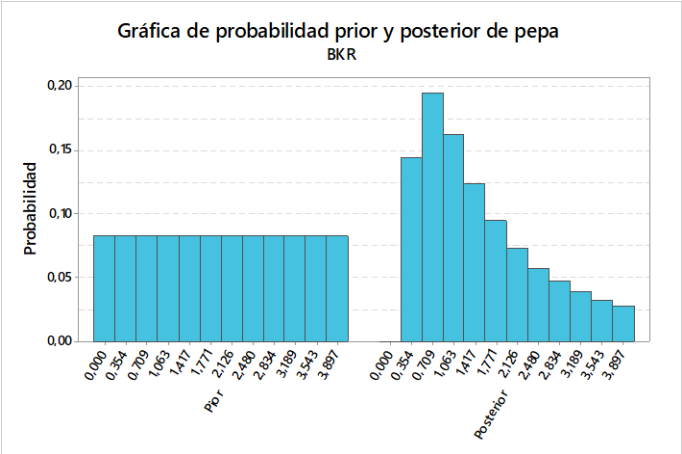


Figura 61: Gráfica de probabilidades prior y posterior de la pepa.

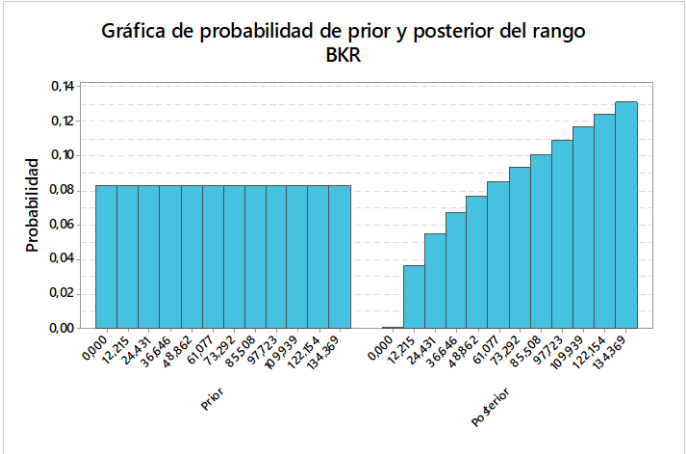


Figura 62: Gráfica de probabilidades prior y posterior del rango.

En la **Figura 63** se visualiza el resultado de la estimación, donde, al igual que el casos anteriores, la distribución de y_{Fe} se condiciona fuertemente a la distribución que posee el tipo y la textura de roca siguiendo los patrones, por otro lado, se aprecia la gran cantidad de leyes bajas en todo el yacimiento, lo anterior se confirma con el histograma de y_{Fe} presentado en la **Figura 64**, que muestra un ajuste a la distribución normal, una y_{Fe} máxima y mínima de 57.4% y 8.5% respectivamente y una ley promedio 31.9%.

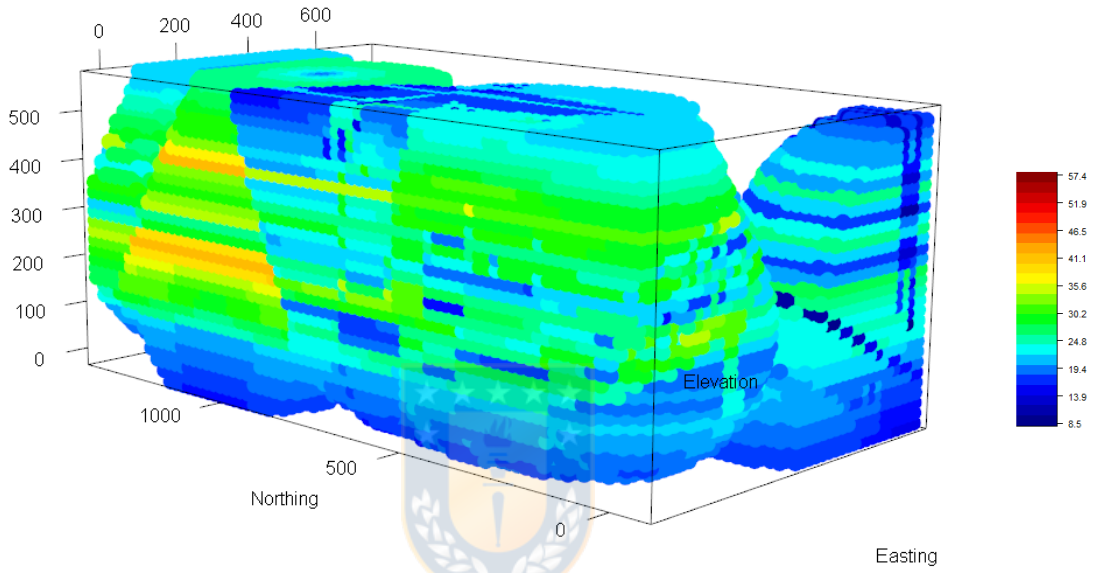


Figura 63: Visualización de estimación y_{Fe} por BKR en yacimiento.

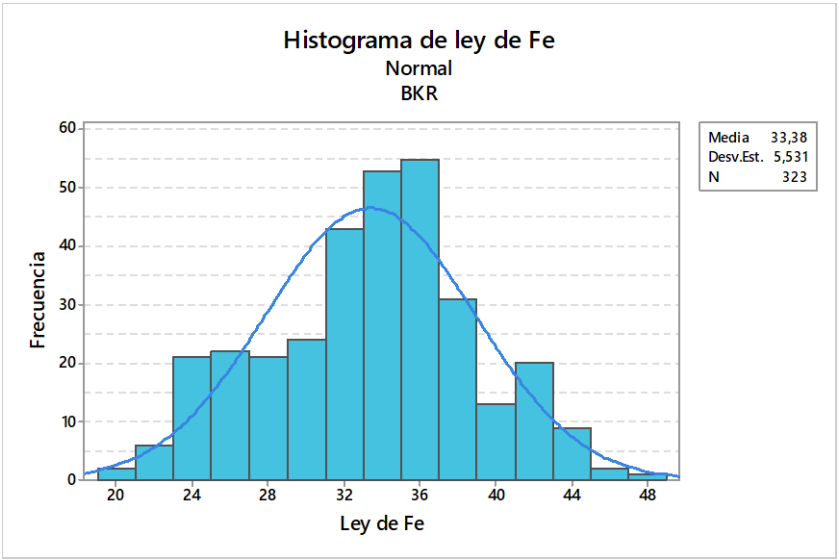


Figura 64: Histograma de estimación y_{Fe} por BKR en yacimiento.

5.3.4 INVERSO DE LA DISTANCIA

Para la optimización del código en la base con sondajes se prueban el coeficiente de ponderación (β), entre 0.5 y 3.0 aumentado a una razón de 0.1, y la cantidad máxima de vecinos utilizados para realizar la estimación entre 10 y 200 con un aumento de 10 creando así un total de 520 combinaciones.

En la **Figura 65** se muestran los resultados obtenidos en las 520 combinaciones para la optimización, donde se concluye que el menor $\mathbf{RMSE}$, con un valor 9.52, resulta con un $\beta = 1.2$ y una cantidad de máximo 70 vecinos.

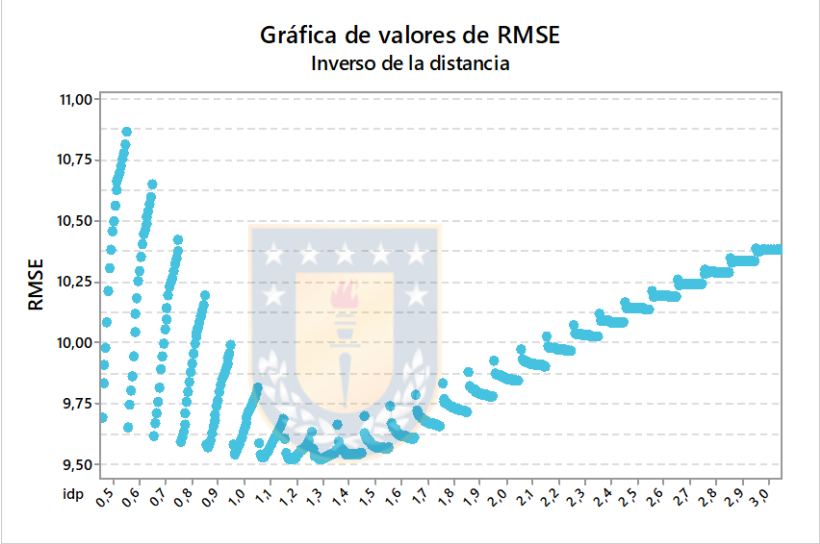


Figura 65: Resultados de RMSE para combinaciones de β y número de vecinos máximos.

Utilizando las variables escogidas se realizar la estimación de y_{Fe} con el modelo de la forma $y \sim Northing + Easting + Elevation$, dado que no se puede crear dependencia de las variables tipo y textura de roca por la misma construcción del algoritmo.

En la **Figura 66** se visualiza el resultado de la estimación, donde se aprecia la gran cantidad de leyes bajas en todo el yacimiento, lo anterior se confirma con el histograma de y_{Fe} presentado en la **Figura 67**, que muestra un ajuste a la distribución log-logística con dos modas, una en la clase $y_{Fe} = 32\%$ y una segunda en $y_{Fe} = 24\%$.

Las y_{Fe} máximas y mínimas del yacimiento son 49.1% y 19.5% respectivamente, una media de 31.4% y una desviación estándar de 3.70, que comparado con la distribución de y_{Fe} de los sondajes la media es parecida pero las leyes máximas son menores.

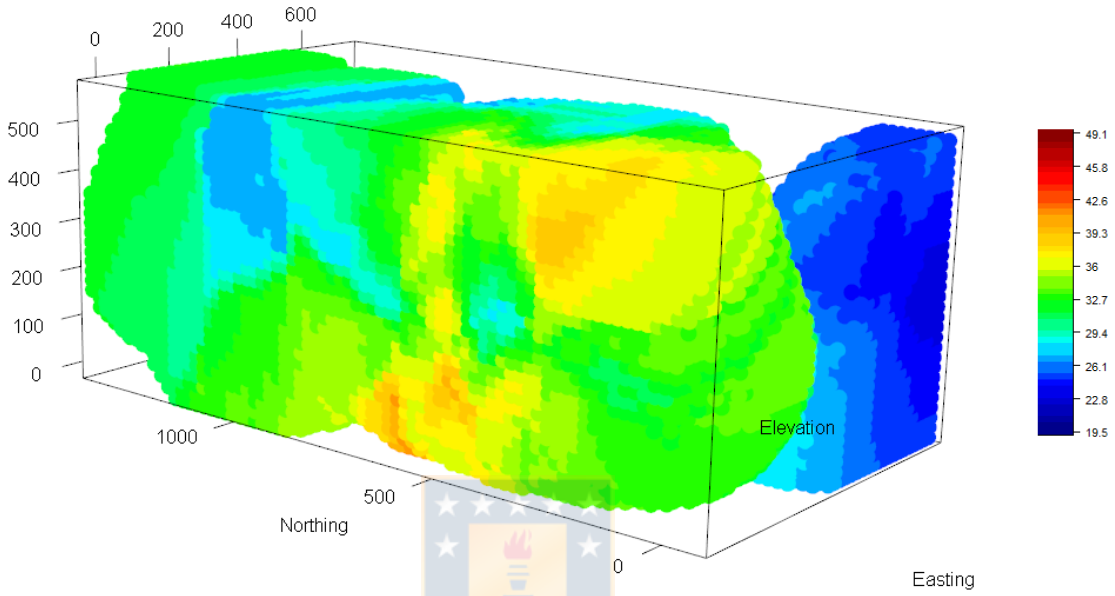


Figura 66: Visualización de estimación y_{Fe} por IDW en yacimiento.

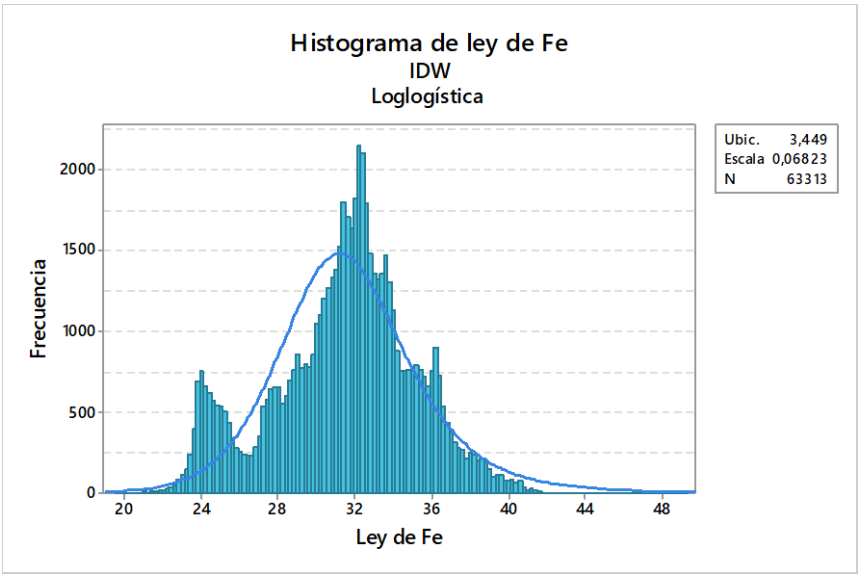


Figura 67: Histograma de estimación y_{Fe} por IDW en yacimiento.

5.3.5 KRIGING ORDINARIO

Para optimizar el código con los sondajes se comienza construyendo un variograma omnidireccional para y_{Fe} con la función *autofitVariogram,* además, es necesario determinar la cantidad de vecinos máximos que se utilizarán, para este caso se estudia el comportamiento de RMSE según la cantidad de vecinos utilizados, comenzando desde 10 y teniendo como limite el total de la base, es decir 660, probando así 60 tipos de cantidad de vecinos.

En la **Figura 68** se aprecia el comportamiento que va teniendo RMSE, donde se evidencia que a medida que la cantidad de vecinos aumenta el RMSE tiene a converger.

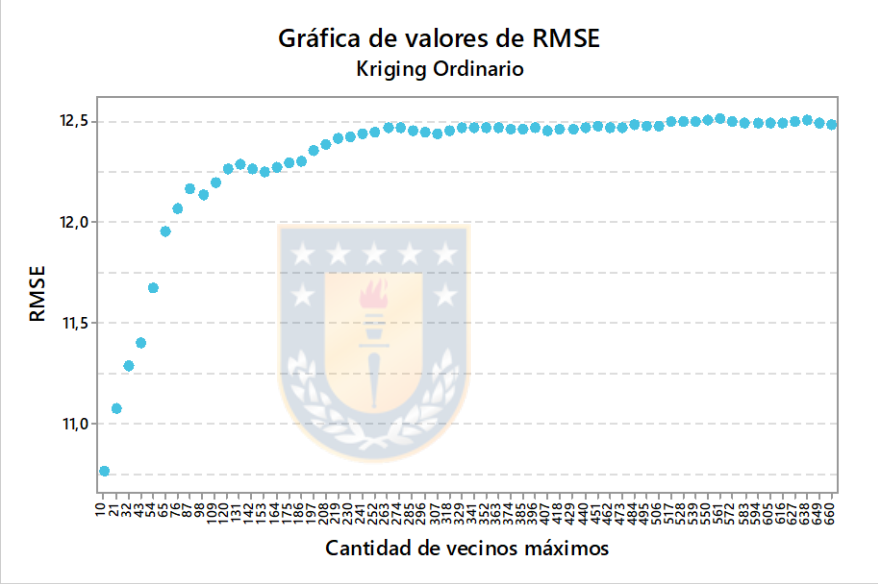


Figura 68: Resultados de RMSE para número de vecinos máximos en OK.

Luego de los resultados obtenido a través del análisis de vecinos máximos, estos se utilizan para la estimación del yacimiento.

En la **Figura 69** se muestra que a pesar de que OK es un algoritmo que no permite dependencia de variables, es decir no se toma en consideración el tipo de roca ni la textura, el algoritmo no logra obtener una continuidad espacial como en el caso de IDW , por otro lado, se aprecia la gran cantidad de leyes bajas en todo el yacimiento, lo anterior se confirma con el histograma de y_{Fe} presentado en la **Figura 70**, donde la mayor concentración de datos se encuentran en $y_{Fe} = 26.7\%$, la cual a su vez es la moda, además posee una y_{Fe} máxima y mínima de 57.4% y 19.9% respectivamente y una ley promedio 29.7%.

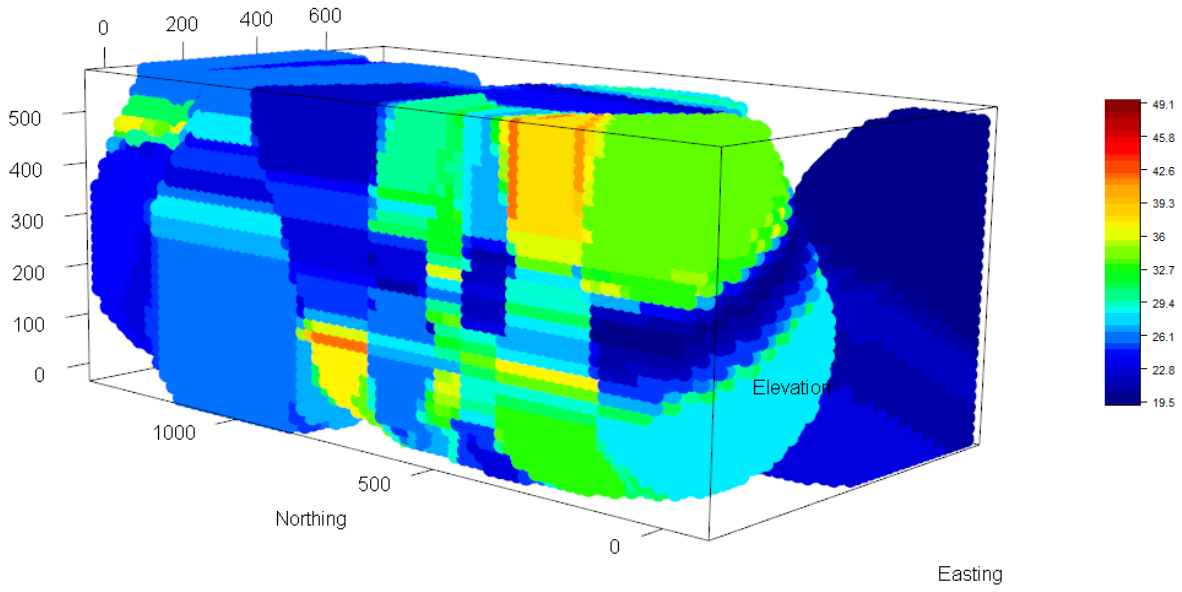


Figura 69: Visualización de estimación y_{Fe} por OK en yacimiento.

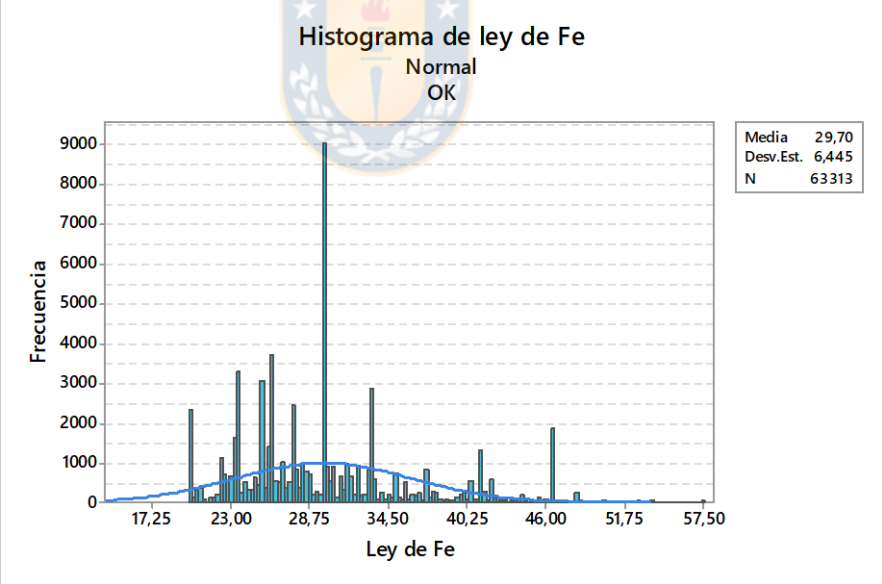


Figura 70: Histograma de estimación y_{Fe} por OK en yacimiento.

5.4 CATEGORIZACIÓN DE RECURSOS MINERALES

Para la categorizacion de recursos se utiliza la misma metodología descrita en el título 4.6, obteniendo los resultados mostrados en la **Figura 71** y detallados en la **Tabla 19.**

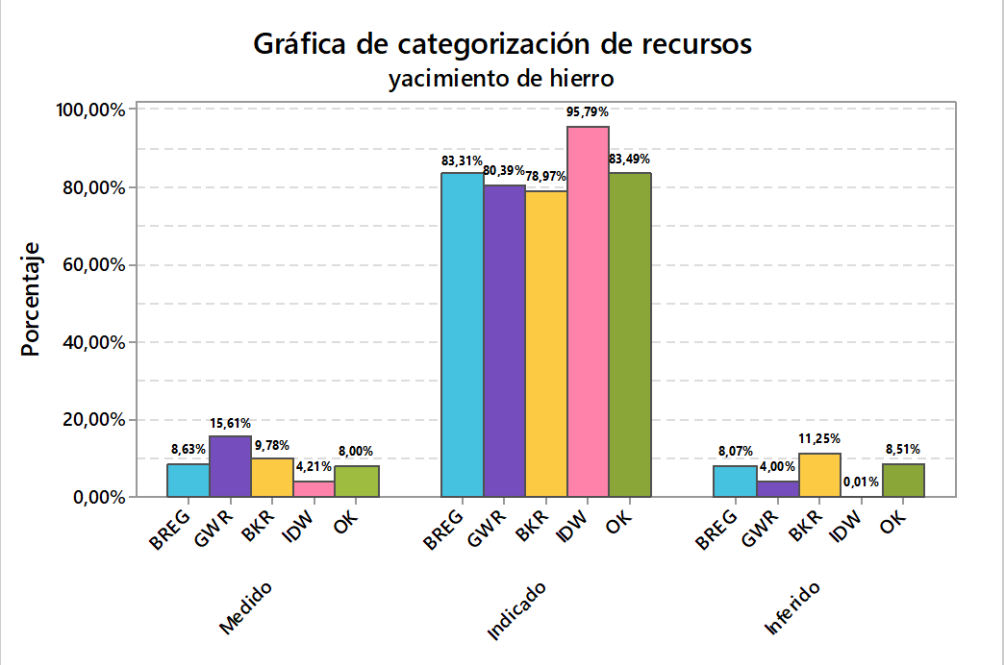


Figura 71: Categorización de recursos en yacimiento de hierro.

En la **Figura 71** se aprecia la gran cantidad de recursos indicados para los cinco métodos y la poca presencia de recursos inferidos y medidos, esto producto de la poca información proveniente de la campaña de sondajes.

Se destaca en los recursos medidos el método GWR con la mayor presencia de estos con un 15.61%, en los recursos indicados IDW con un 95.79% y en los recursos inferidos el método BKR con un 11.25%.

En la **Tabla 19** se encuentra el detalle de la categorización de recursos junto con la ley media y el tonelaje asociado a cada categorización.

Para el cálculo del tonelaje asociado del yacimiento, se utiliza la Ec. 5.1 para el cálculo de la densidad, puesto que la densidad en un yacimiento de hierro se relaciona directamente con la ley asociada.

$$D\left(\frac{ton}{m^3}\right) = 2.2959 + 0.03793 \times y_{Fe} \qquad \text{Ec. 5.1}$$

Tabla 19: Categorización de recursos para yacimiento de hierro.

		Inferidos			Indicados			Medidos		
		[%]	[MTon]	$\bar{y}$	[%]	[MTon]	$\bar{y}$	[%]	[MTon]	$\bar{y}$
Yacimiento de hierro	BREG	8.07	19.98	21.89	83.31	517.40	32.19	8.63	68.60	41.16
	GWR	4.00	8.48	18.29	80.39	492.89	32.49	15.61	141.79	46.98
	BKR	11.25	40.60	20.86	78.97	477.39	32.31	9.78	81.68	43.36
	IDW	0.010	0.23	22.85	95.79	708.74	31.29	4.21	27.96	34.05
	OK	8.51	134.31	26.13	83.49	481.39	29.42	8.00	64.80	41.72

5.5 ANÁLISIS DE RESULTADOS

Para la estimación de y_{Fe} en el yacimiento y en la base con sondajes a través de validación cruzada uno a uno, se concluye que:

- BREG al realizar la optimización con validación cruzada, la calidad de la estimación no se ve afectada por los caracteres de la función enlace y lo que afecta la estimación es la elección de la función enlace. Por otro lado, en la estimación de y_{Fe} en el yacimiento la distribución de ley obtenida por el método es parecida (ley mínima y media) a la distribución de ley en los sondajes, pero difiere significativamente en el máximo, siendo mucho mejor la ley máxima entregada por la estimación.
- GWR al ser un método de regresión 2D fue necesario adaptarlo a estimación 3D, de la misma forma que en el caso simulado, pudiendo afectar la calidad de las estimaciones realizadas. En cuando a la optimización del código se destaca el efecto que tiene la determinación del ancho de banda respecto a la elección del Kernel, es decir, genera un mayor cambio la variación de ancho de banda que el tipo de kernel, efecto que se muestra en Figura 58, a medida que aumenta el ancho de banda para un mismo kernel el RMSE va decayendo hasta encontrar un mínimo y luego cambia la pendiente aumentando el RMSE, por último la distribución de ley resultante para el yacimiento se adapta fácilmente a un distribución lognormal.

- BKR al igual de GWR tuvo que adaptarse para realizar estimaciones 3D, en cuanto a la optimización de parámetros se decide variar el rango y la pepa del variograma mientras que la meseta se mantiene como predefinida, esto producto del aumento del costo computacional en la optimización. Por otro lado, la distribución de ley resultante de la estimación del yacimiento pudo ajustarse a una distribución normal teniendo una ley media parecida a la obtenida de los sondajes, por último, BKR fue el método que logro estimar las menores leyes en el yacimiento.
- BREG, GWR, BKR son algoritmos que su estimación se condicionaron a los resultados de tipo de Roca y textura obtenidos por co-Kriging indicador, dado que la estimación se modeló de manera $y_{Fe} \sim Elevation + Northing + Easting + Tipo\ de\ roca + Textura$, por lo cual la continuidad espacial de las leyes fueron limitadas según los atributos discretos.
- IDW fue el método en el cual probaron la mayor cantidad de combinaciones entre el coeficiente de ponderación (β) y la cantidad máxima de vecinos con una total de 520, en donde ambas variables afectaban notoriamente el resultado de RMSE, además se pudo visualizar el efecto combinado de ambas variables en la Figura 68 en donde el aumento de ambas causaba una disminución del RMSE hasta lograr un mínimo. Por otro lado, en la estimación del yacimiento IDW fue el mejor método respecto a la continuidad espacial.
- OK a pesar de ser un método el cual no fue condicionado según el tipo de roca y textura no logro tener la misma continuidad espacial que IDW y en cuanto a la distribución de ley obtenida al realizar la estimación esta distribución se destacó por la gran cantidad de modas obtenidas no pudiéndose adaptar a ningún ajuste.

Para el análisis de la efectividad de los métodos de regresión se realiza una comparación de los RMSE obtenidos en la base con sondajes, En la Tabla 20 se muestran los resultados de RMSE para cada caso, en el cual se visualiza lo cercano de los resultados de RMSE entre todos los métodos, siendo el método IDW el cual logró el menor valor y BREG el método con el mayor valor de RMSE.

Tabla 20: Resultados de RMSE en base con sondajes.

Método	RMSE
BREG	10.91
GWR	10.40
BKR	10.10
IDW	9.52
OK	10.76

Por último, en la categorización de recursos se visualiza en la **Figura 71** la fuerte presencia de recursos indicados y la escasa cantidad de recursos medidos, esto causado por la irregular y poca información proveniente de los sondajes en comparación del tamaño del yacimiento. Además, se destacan los métodos GWR por la mayor cantidad de recursos medidos obtenidos en comparación al resto de los métodos de regresión y el método IDW por ser el método que más recursos indicados obtuvo.

CONCLUSIONES Y DISCUSIONES

El objetivo de esta Memoria de Título apuntó a analizar la aplicación de modelos de regresión lineal no convencionales para la construcción de modelos de bloques, en este caso se estudió la eficacia de los métodos Regresión Beta (BREG), Regresión geográficamente ponderada (GWR) y Kriging Bayesiano (BKR) y la evaluación del desempeño de los modelos se realizó mediante RMSE.

Para lograr el objetivo se dividió el estudio en un caso simulado y un caso real en los cuales se pusieron a prueba los tres métodos de regresión mencionados más dos métodos de regresión que fueron Inverso de la distancia (IDW) y Kriging Ordinario (OK) con el fin de realizar comparaciones entre estos métodos en la creación de modelos de bloques en yacimientos mineros.

En el caso simulado se destacó que desarrollar 100 simulaciones para realizar la estimación de un atributo continuo permitió ver tendencias de los resultados obtenidos, tanto en calidad de estimaciones como en el costo computacional (tiempo de programación[7]), logrando obtener conclusiones tanto generales como particulares del comportamiento de los algoritmos. Aquí se concluyó principalmente:

- BREG, GWR, BKR tuvieron resultados de desempeño (RMSE) del mismo rango que los métodos IDW y OK demostrando que los métodos propuestos son eficaces para la construcción de modelos de bloques en un yacimiento.
- BKR se destacó notoriamente al presentar los mejores resultados de RMSE promedio, superando los resultados de los métodos BREG, GWR, IDW e igualando los resultados de OK.
- BREG si bien logró tener buenos resultados en la estimación del atributo continuo fue a su vez el método con la menor evaluación de desempeño, discrepando en gran medida a la distribución de ley que se deseaba obtener.
- GWR y BKR tienen a aumentar rápidamente el costo computacional al aumentar el rango de los parámetros de cada algoritmo.

[7] Tiempo en que se demora el algoritmo al entregar el resultado de ley y en un punto.

- Producto de la cantidad de información disponible y el distanciamiento homogéneo que se poseía la gran cantidad de las estimaciones, para todos los métodos de regresión, fueron categorizadas como recursos medidos.

En el caso real se ponen a prueba los algoritmos frente a un escenario con alto nivel incertidumbre y escasa información, la cual proviene de una campaña de sondajes, en este escenario se pudo analizar el comportamiento de los métodos de regresión y la variación de la calidad de las estimaciones al disminuir la cantidad y calidad de información disponible. Aquí se concluyó principalmente:

- BREG, GWR, BKR, al igual que el caso simulado, tuvieron resultados de desempeño (RMSE) del mismo rango que los métodos IDW y OK demostrando que los métodos propuestos son eficaces para la construcción de modelos de bloques donde se posee escaza información disponible.
- IDW fue el método de regresión con mejor resultado de desempeño (RMSE) seguido por BKR, ambos métodos a su vez fueron los algoritmos que tuvieron los costos computacionales más altos.
- Al agregar más información, en este caso los atributos discretos tipo de roca y textura, influyeron positivamente en la evaluación de desempeño de BREG, GWR, BKR respecto al no considerar esta información.
- GWR y BK son los métodos de regresión que lograron estimar los menores valores de leyes, y a su vez GWR fue el método que logro estimar los mayores valores de leyes.

A pesar de que se logra cumplir con todos los objetivos de la investigación, es fundamental continuar la investigación tomando las siguientes recomendaciones:

- Analizar el comportamiento de los algoritmos GWR y BKR en la construcción de modelos de bloques en un software que tenga adaptados los métodos en 3D, dado que la adaptación realizada en la Memoria de Título (estimación por niveles) no permite buscar vecinos que se encuentren a una distancia diagonal al punto a estimar, solo se consideraran vecinos en los ejes verticales y horizontales.

- Analizar el arrastre del error de estimación en los métodos BREG, GWR y BKR en la construcción de un modelo de bloques, producto que la utilización de estos métodos fue para estimar un atributo continuo el cual dependía de otros atributos que se estimaron por otros métodos de regresión.

7 REFERENCIAS

[1] WACKERNAGEL, Hans. *Multivariate geostatistics: an introduction with applications*. Springer Science & Business Media, 2013.

[2] GIRALDO, Ramón. Introducción a la Geoestadística. *Teoría y Aplicación. Universidad Nacional de Colombia, Facultad de Ciencias, Departamento de Estadística, Bogotá, DC*, 2002.

[3] KRIGE, D. G.; GUARASCIO, M.; CAMISANI-CALZOLARI, F. A. Early South African geostatistical techniques in today's perspective. En *Geostatistics*. Springer, Dordrecht, 1989. p. 1-19.

[4] CHARLTON, Martin; FOTHERINGHAM, Stewart; BRUNSDON, Chris. Geographically weighted regression. *White paper. National Centre for Geocomputation. National University of Ireland Maynooth*, 2009.

[5] CRIBARI-NETO, Francisco; ZEILEIS, Achim. Beta regression in R. 2009.

[6] ABAD, I. O.; CÉLLERI, R. Estimación de precipitación espacial mediante correlación con variables secundarias y la ayuda de tecnologías de información geográfica. *Maestría en Geomática Orientada al Ordenamiento Territorial. Universidad de Cuenca. Facultad de Ingeniería. Cuenca, Ecuador*, 2014.

[7] CORREA, Julián Vélez; FIGUEROA, Pedro Nieto. Validación de medidas de evaluación para el pronóstico de la tasa de cambio en Colombia. *Bogotá: Colegio de Estudios Superiores de Administración–CESA–Maestría en Finanzas Corporativas*, 2016.

[8] TORRES, Sara Lucía González, et al. Modelo de regresión beta para la actividad enzimática de la lipasa pancreática.

[9] SCHMID, Matthias, et al. Boosted beta regression. *PloS one*, 2013, vol. 8, no 4, p. e61623.

[10] FUSTOS, Roberto, Modelo Lineal Generalizado espacial con variale respuesta beta, Unversidad de Concepción, 2013.

[11] QUESADA, Manuel, Análisis de Series Temporales. Modelos Heterocedásticos*, Universidad de Granada*.

[12] BRETON, Carrie V., et al. Prenatal tobacco smoke exposure is associated with childhood DNA CpG methylation. *PloS one*, 2014, vol. 9, no 6, p. e99716.

[13] SWEARINGEN, Christopher J., et al. Application of beta regression to analyze ischemic stroke volume in NINDS rt-PA clinical trials. *Neuroepidemiology*, 2011, vol. 37, no 2, p. 73-82.

[14] WARTON, David I.; HUI, Francis KC. The arcsine is asinine: the analysis of proportions in ecology. *Ecology*, 2011, vol. 92, no 1, p. 3-10.

[15] PETERSON, Erin E.; URQUHART, N. Scott. Predicting water quality impaired stream segments using landscape-scale data and a regional geostatistical model: a case study in Maryland. *Environmental Monitoring and Assessment*, 2006, vol. 121, no 1-3, p. 615-638.

[16] FERRARI, Silvia; CRIBARI-NETO, Francisco. Beta regression for modelling rates and proportions. *Journal of applied statistics*, 2004, vol. 31, no 7, p. 799-815.

[17] SIMAS, Alexandre B.; BARRETO-SOUZA, Wagner; ROCHA, Andréa V. Improved estimators for a general class of beta regression models. *Computational Statistics & Data Analysis*, 2010, vol. 54, no 2, p. 348-366.

[18] CHARLTON, Martin; FOTHERINGHAM, Stewart; BRUNSDON, Chris. Geographically weighted regression. *White paper. National Centre for Geocomputation. National University of Ireland Maynooth*, 2009.

[19] GUTIÉRREZ, José; GARCÍA PALOMARES, J. C.; CARDOZO, O. Regresión Geográficamente Ponderada (GWR) y estimación de la demanda de las estaciones del Metro de Madrid. En *Article in Spanish]. In: Proceedings of the 15th National Congress of Technologies of Geographic Information, Madrid, Spain*. 2012.

[20] CONDADO, Sara Garcia. *Generalización de variables medioambientales mediante interpolación GIS*. 2016. Tesis Doctoral. Universidad Politécnica de Madrid.

[21] MEI, Chang-Lin, et al. Efficient estimation of heteroscedastic mixed geographically weighted regression models. *The Annals of Regional Science*, 2020, p. 1-22.

[22] CLARK, Stephen D. Estimating local car ownership models. *Journal of transport Geography*, 2007, vol. 15, no 3, p. 184-197.

[23] SONG, Xiao-Dong, et al. Mapping soil organic carbon content by geographically weighted regression: A case study in the Heihe River Basin, China. *Geoderma*, 2016, vol. 261, p. 11-22.

[24] SÁNCHEZ, Landy. Métodos para el análisis espacial: una aplicación al estudio de la geografía de la pobreza. En *Ponencia presentada en el II Congreso de la Asociación Latinoamericana de Población, Guadalajara*. 2006.

[25] DUQUE, Juan Carlos; CEBALLOS, Hermilson Velásquez; AGUDELO, Jorge. Infraestructura pública y precios de vivienda: una aplicación de regresión geográficamente ponderada en el contexto de precios hedónicos. *Ecos de Economía: A Latin American Journal of Applied Economics*, 2011, vol. 15, no 33, p. 99-122.

[26] GALLARDO, A. Geostadística. *Revista ecosistemas*, 2006, vol. 15, no 3.

[27] GIRALDO, Ramón. Introducción a la Geoestadística. Teoría y aplicación. *Bogotá, Facultad de Ciencias, Departamento de Estadística, Universidad Nacional de Colombia*, 2002.

[28] GIRALDO, R. Introducción a la geoestadística: teoría y aplicación. Bogotá, DC: Universidad Nacional de Colombia. Facultad de Ciencias. Departamento de Estadística. 94 p. 2002.

[29] EMERY, Xavier. Geoestadística . s.l. *: Facultad de ciencias físicas y matemíticas Universidad de Chile*, 2013.

[30] MANTILLA SALTOS, Gabriel Fernando, et al. *Comparación de metodologías estadísticas para interpolar la precipitación en el ecuador*. 2018. Tesis de Licenciatura. Espol.

[31] CRESSIE, Noel. *Statistics for spatial data*. John Wiley & Sons, 2015.

[32] NANNAVECCHIA, Antonella; SPAGNOLO, Stefano; POLLICE, Alessio. Air Quality in Taranto: Multivariate Bayesian Kriging.

[33] AXIS-ARROYO, J.; MATEU, J.; TORRUCO, D. Diferencias entre modelos geoestadísticos aplicados en el análisis de la distribución espaciotemporal de especies biológicas. 2003.

[34] NOBRE, M. M.; SYKES, J. F. Application of Bayesian kriging to subsurface characterization. *Canadian geotechnical journal*, 1992, vol. 29, no 4, p. 589-598.

[35] NIYIBIZI, Bart; BRORSEN, Wade; PARK, Eunchun. Using Bayesian Kriging for Spatial Smoothing of Trends in the Means and Variances of Crop Yield Densities. 2018.

[36] MOREANO VITERI, Ricardo José. *Sistema de información para la interpolación espacial y temporal de datos sobre el tiempo atmosférico y el clima del Ecuador*. 2007. Tesis de Licenciatura. QUITO/EPN/2008.

[37] VOLFOVÁ, Adéla; ŠMEJKAL, Martin. Geostatistical Methods in R. *Geoinformatics FCE CTU*, 2012, vol. 8, p. 29-54.

[38] GARCÍA, José Miguel Contreras; PORTILLO, Elena Molina; CEZÓN, Pedro Arteaga. *Introducción a la Programación Estadística con R para profesores*. Grupo de Educación Estadística, Universidad de Granada, 2010.

[39] MEI, Chang-Lin, et al. Efficient estimation of heteroscedastic mixed geographically weighted regression models. *The Annals of Regional Science*, 2020, p. 1-22.

[40] NEIRA, Valentina. Estimación de atributos categóricos del modelo geológico con herramientas de Data Mining. *Universidad de concepción,* Concepción 2020.

8 ANEXOS

8.1 FUNCIÓN DE ENLACE

Los modelos lineales generalizados incluyen una función de enlace que relaciona el valor esperado con los predictores lineales del modelo. Una función de enlace transforma las probabilidades de una variable respuesta categórica en una escala continua que no tiene límites. Una vez que se completa la transformación, la relación entre los predictores y la respuesta se puede modelar con regresión lineal. Por ejemplo, una variable de respuesta binaria puede tener dos valores únicos. La conversión de estos valores a probabilidades hace que la variable de respuesta varíe de 0 a 1. Cuando aplica una función de enlace apropiada a las probabilidades, los números que resultan varían de $-\infty$ a $+\infty$.

La forma general de la función de enlace es la siguiente:

$$g(\mu_i) = x_i^T \beta \qquad \text{Ec. 8.1}$$

8.2 DISTRIBUCIÓN DE GAUSS

La distribución de Gauss o normal fue reconocida por primera vez por el francés Abraham de Moivre (1667-1754). Posteriormente, Carl Friedrich Gauss (1777-1855) elaboró desarrollos más profundos y formuló la ecuación de la curva; de ahí que también se la conozca, más comúnmente, como la "Campana de Gauss". La distribución de una variable normal está completamente determinada por dos parámetros, su media y su desviación estándar, denotadas generalmente por μ y σ. Con esta notación, la densidad de la normal viene dada por la ecuación:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}, -\infty < x < \infty \qquad \text{Ec. 8.2}$$

Se determina la curva en forma de campana. Así, se dice que una característica X sigue una distribución normal de media μ y varianza σ, y se denota como $X \sim N(\mu, \sigma)$, si su función de densidad viene dada por la Ecuación mostrada.

8.3 Distribución Exponencial

La distribución exponencial es aquella que modela el tiempo transcurrido entre dos sucesos que se producen de forma independiente, separada y uniforme en el tiempo. Se dice que una variable aleatoria X sigue una distribución exponencial de parámetro λ, y se denota por $X \sim \exp(\lambda)$, si su función de densidad es:

$$f(x) = \lambda e^{-\lambda x}, x \geq 0 \qquad \text{Ec. 8.3}$$

8.4 Máxima verosimilitud

El método más común para estimar parámetros es el método de máxima verosimilitud. Sea $X_1, \dots, X_n$ independientes e idénticamente distribuidas con función de densidad de probabilidad $p(x; \theta)$ entonces la función de verosimilitud se define como:

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} p(x_i; \theta) \qquad \text{Ec. 8.4}$$

Y la Log-verosimilitud se define como:

$$l(\theta) = \log \mathcal{L}(\theta) = \sum_{i=1}^{n} \log p(x_i; \theta) \qquad \text{Ec. 8.5}$$

La función de verosimilitud no es más que la densidad conjunta de los datos, con la diferencia de que se trata como función del parámetro θ. Por lo tanto $\mathcal{L}: \Theta \to [0, \infty)$, en general $\mathcal{L}(\theta)$ no integra uno respecto a θ. El estimador de máxima verosimilitud es el valor de θ que maximiza $\mathcal{L}(\theta)$.

El máximo de $l(\theta)$ se alcanza en el mismo lugar que el máximo de $\mathcal{L}(\theta)$, por lo que maximizar la log-verosimilitud es equivalente a maximizar la verosimilitud.

8.5 Regresión Beta con dispersión variable

En este modelo, el parámetro de precisión no es constante para todas las observaciones, sino que se modela de manera similar al parámetro medio. Mas específicamente, $y_i \sim \mathcal{B}(\mu_i, \emptyset_i)$ independientemente, $i = 1, \dots, n$ y con:

$$g_1(\mu_i) = \eta_{1i} = x_i^T \beta \qquad \text{Ec. 8.6}$$

$$g_2(\emptyset_i) = \eta_{2i} = z_i^T \gamma \qquad \text{Ec. 8.7}$$

Donde, $\beta = \left(\beta_1, \dots, \beta_k\right)^T$, $\gamma = (\gamma_1, \dots, \gamma_h)^T, k + h < n$, son los conjuntos de regresión en las ecuaciones, η_{1i} y η_{2i} son los predictores lineales, y x_i y z_i son vectores regresores. Como antes, ambos vectores de coeficientes son estimados por ML, simplemente reemplazando $\emptyset$ por $\emptyset_i$.

Simas et. al (2010) amplían aún más el modelo anterior al permitir predictores no lineales en las ecuaciones Ec. 8.18 y Ec. 828. Además, han obtenido correcciones de sesgo analítico para los estimadores de ML de los parámetros, generalizando así los resultados de Ospina, Cribari Neto y Vasconcellos (2006), quienes derivaron correcciones del sesgo para las regresiones beta de dispersión.

8.6 RESIDUO PONDERADO ESTANDARIZADO 2

Espinheira, Ferrari y Cribari-Neto (2008) propusieron otros residuos, en particular uno residual con mejores propiedades, se define entonces:

$$r_{SW2,i} = \frac{y_i^* - \hat{\mu}_i^*}{\sqrt{\hat{v}_i(1 - h_{ii})}} \qquad \text{Ec. 8.8}$$

Donde $y_i^* = \log\left\{\frac{y_i}{1-y_i}\right\}$ y $\mu_i^* = \psi(\mu_i\phi) - \psi\big((1-\mu_i)\phi\big), \psi(\cdot)$denota a la función digamma. La estandarización es entonces $v_i = \{\psi'(\mu_i\phi) + \psi'\big((1-\mu_i)\phi\big)$ y h_{ii} el elemento diagonal i-ésimo de la matriz gorro (para más detalles ver Ferrari y Cribari-Neto 2004; Espinheira et al. 2008b). Los gorros denotan evaluación en las estimaciones de ML.

Es de destacar que el modelo de regresión beta descrito anteriormente se desarrolló para permitir a los profesionales modelar variables continuas que asumen valores en el intervalo de la unidad, como tasas, proporciones e índices de concentración o desigualdad (por ejemplo, la Desigualdad de Gini[8]).

[8] El coeficiente de Gini es una medida de la desigualdad ideada por el estadístico italiano Corrado Gini. Normalmente se utiliza para medir la desigualdad en los ingresos, dentro de un país, pero puede utilizarse para medir cualquier forma de distribución desigual.

8.7 HETEROGENEIDAD ESPACIAL

El término heterogeneidad espacial se refiere a la variación en las relaciones sobre el espacio. En el más general de los casos se podría considerar que una relación diferente se presente para cada punto en el espacio. La estimación de este tipo de modelos lleva implícitos problemas relacionados con los grados de libertad, simplemente no se tienen suficientes datos con los cuales producir estimadores para cada punto en el espacio.

Para proceder con el análisis, se debe proveer una especificación para la variación sobre el espacio. Por lo tanto, para llevar a cabo estimaciones e inferencias con soporte formal y asegurar la identificabilidad del modelo, es necesario imponer algunas restricciones a la expresión general.

8.8 MÉTODO DE MÍNIMOS CUADRADOS ORDINARIOS

El criterio de este método consiste en proporcionar estimadores de los parámetros que minimicen la suma de los cuadrados de los errores. Operativamente el proceso es construir una función objetivo en términos de la suma de los cuadrados de los errores y mediante optimización (condiciones de primer orden - C.P.O., y condiciones de segundo orden - C.S.O.) obtener las fórmulas de cálculo de los estimadores.

Debido a que la función de regresión poblacional no se puede observar directamente, los estimadores de mínimos cuadrados ordinarios se obtienen a partir de la función de regresión muestral (FRM). La FRM es:

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + e_1 \qquad \text{Ec. 8.9}$$

$$Y_i = \hat{Y}_i + e_1 \qquad \text{Ec. 8.10}$$

La suma del cuadrado de los errores puede expresarse como sigue:

$$\sum e_i^2 = \sum \left(Y_i - \hat{Y}_i\right)^2 = \sum \left(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i\right)^2 \qquad \text{Ec. 8.11}$$

De acuerdo con el principio de mínimos cuadrados ordinarios:

$$\min \sum e_i^2 = \ min \sum \left(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i\right)^2 \qquad \text{Ec. 8.12}$$

Derivando la anterior expresión con respecto a $\hat{\beta}_1$ y $\hat{\beta}_2$ e igualando a cero, respectivamente, y resolviendo las ecuaciones normales, se encuentran los estimadores de los parámetros de la regresión:

$$\hat{\beta}_2 = \frac{n\sum X_i Y_i - (\sum X_i)(\sum Y_i)}{n\sum X_i^2 - (\sum X_i)^2} = \frac{\widehat{Cov}(X,Y)}{\widehat{Var}(X)} \qquad \text{Ec. 8.13}$$

$$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{Y} \qquad \text{Ec. 8.14}$$

8.9 MÉTODO DE MÍNIMOS CUADRADOS PONDERADOS

Cuando la condición de normalidad de los datos no se cumple, una de las posibilidades para la estimación de los parámetros son métodos alternativos como el criterio de mínimos cuadrados ponderados (*Weighted Least Squares*), para cuya aplicación no es necesaria dicha condición. Una de las ventajas que presenta este método es que permite introducir en los análisis variables ordinales, variables dicotómicas y variables continuas que no se ajusten a criterios de normalidad. Este método consiste en estimar α minimizando la suma de cuadrados:

$$\sum_{i=0}^{n}\sum_{j=1}^{n_i} w_i\left(y_{ij} - x_{ij}^T\alpha\right)^2 \qquad \text{Ec. 8.15}$$

Donde w_i son pesos dados usando alguno de los siguientes esquemas de ponderación:

1. **Esquema 1:** En el que los pesos están dados por $w_i = \frac{1}{N}, i = 1, \ldots, n.$ Donde:

$$N = \sum_{i=1}^{n} n_i \qquad \text{Ec. 8.16}$$

2. **Esquema 2:** En el que los pesos están dados por $w_i = \frac{1}{nn_i}, i = 1, \ldots, n.$

$$y = X\alpha + e \qquad \text{Ec. 8.17}$$

El esquema 1 usa el mismo peso para todos los individuos y fue implementado por Hoover et al. (1998). El esquema 2 es considerado por Huang et al. (2002) y usa diferentes pesos para los individuos. Huang et al. (2002) demuestran que el esquema 1 puede llevar a estimaciones inconsistentes de α. Minimizando la función objetivo Ec. 8.15 trabajando con el modelo Ec. 8.17, se obtiene un estimador de α dado por:

$$\hat{\alpha}_{MVC} = (X^T W X)^{-1} X^T W y \qquad \text{Ec. 8.18}$$

Donde, W= $diag[W_1, \dots, W_n\ ]$, con $W_i = w_i \mathrm{I}_{\mathrm{ni}}$ la matriz de pesos del i-ésimo individuo, $i = 1, \dots, n$, e I_{ni} la matriz de identidad de $n_i \times n_i$.

8.10 Criterio de información de Akaike (AIC)

La idea básica del AIC, concebido para la estrategia de estimación que involucra el método de MCP, es encontrar la combinación de parámetros de suavizamiento, determinados por la cantidad de nodos, que minimicen la expresión:

$$AIC(\rho) = -2Loglik + 2df \qquad \text{Ec. 8.19}$$

Donde $\rho = [\rho_0, \rho_1, \dots, \rho_d]^T$ es el vector conformado por los parámetros de suavizamiento,

$$Loglik = -\frac{n}{2} log\left(\frac{2\pi e}{n} SCE_\rho\right) \qquad \text{Ec. 8.20}$$

Con,

$$SCE_\rho = \sum_{i=1}^{n} \sum_{j=1}^{n_i} \left(y_{ij} - \hat{y}_{ij}\right)^2 \qquad \text{Ec. 8.21}$$

$$df = tr(A) \qquad \text{Ec. 8.22}$$

Donde A es la matriz de suavizamiento asociada con el estimador, que en el caso del estimador Ec. 8.18 está dada por:

$$A = X(X^T W X)^{-1} X^T W \qquad \text{Ec. 8.23}$$

Se elige el modelo que minimice la cantidad $AIC(\rho)$. Esta medida permite encontrar un equilibrio entre la bondad de ajuste del modelo, representada por Loglik, y la complejidad del modelo, representada por df. Es decir, la bondad de ajuste del modelo es penalizada con la complejidad de este.

Una de las desventajas de este método es que no tiene en cuenta directamente la posible correlación de las medidas repetidas de cada individuo, una de las características más importantes de la estructura de los datos longitudinales. En la siguiente sección se propone una alternativa de estimación que considera la estructura de correlación de las medidas repetidas.

8.11 DIVERGENCIA DE KULLBACK-LEIBLER

En teoría de la probabilidad y teoría de la información, la a divergencia de Kullback-Leibler (KL) (también conocida como divergencia de la información, ganancia de la información, entropía relativa o KLIC por sus siglas en inglés) es una medida no simétrica de la similitud o diferencia entre dos funciones de distribución de probabilidad P y Q .

KL mide el número esperado de extra bits requeridos en muestras de código de P cuando se usa un código basado en Q, en lugar de un código basado en P. Generalmente P representa la "verdadera" distribución de los datos, observaciones, o cualquier distribución teórica.

Aunque a menudo se considera como una métrica o distancia, la divergencia KL no lo es en realidad, por ejemplo, no es simétrica ya que la divergencia KL de P a Q no necesariamente es la misma KL de Q a P.

La divergencia KL es un caso especial de una clase más amplia de divergencias llamadas divergencias f. Fue originalmente introducida por Solomon Kullback y Richard Leibler en 1951 como la divergencia direccionada entre dos distribuciones.

Para distribuciones de probabilidad P y Q de una variable aleatoria discreta su divergencia KL se defina como:

$$D_{KL}(P\|Q) = \sum_i P(i) ln \frac{P(i)}{Q(i)} \qquad \text{Ec. 8.24}$$

En otras palabras, es el promedio ponderado de la diferencia logarítmica entre las probabilidades P y Q, donde el promedio se toma usando las probabilidades P. La divergencia KL se define si P y Q suman 1 y si $Q(i) > 0$ para cualquier i tal que $P(i) > 0$. Si la cantidad $0 \ln 0$ aparece en la fórmula se interpreta como cero.

8.12 ESTIMACIÓN POR MEDIO DE LA FUNCIÓN DE SEMIVARIANZA

Los pesos λ pueden ser estimados a través de la función de semivarianza, para lo cual se requiere conocer la relación entre las funciones de covariograma y de semivarianza. Antes de esto conveniente tener en cuenta la siguiente notación:

$$\sigma^2 = V\big(Z(x)\big) \qquad \text{Ec. 8.25}$$

$$\gamma_{ij} = \gamma(h) \qquad \text{Ec. 8.26}$$

Donde,

- h: Es la distancia entre los puntos i y j

Análogamente,

$$\mathrm{C}_{ij} = C(h) \qquad \text{Ec. 8.27}$$

La relación entre las dos funciones es la siguiente:

$$\gamma_{ij} = E\big[\mathrm{Z}(x_j) - \mathrm{Z}(x_i)^2\big]$$

$$= \frac{1}{2} E\left[\mathrm{Z}(x_j)^2 - 2\, Z(x_j)\mathrm{Z}(x_i) + \mathrm{Z}(x_i)^2\right]$$

$$= \frac{1}{2}[V(Z(x)] + \frac{1}{2}[V(Z(x)] - COV\big[Z(x_j)\mathrm{Z}(x_i)\big]$$

$$= V[Z(x)] - COV\big[Z(x_j)\mathrm{Z}(x_i)\big]$$

$$= \sigma^2 - \mathrm{C}_{ij}$$

$$\mathrm{C}_{ij} = \sigma^2 - \gamma_{ij} \qquad \text{Ec. 8.28}$$

Reemplazado Ec. 8.28 en Ec 2.54 se determinan los pesos óptimos λ en términos de la función de semivarianza:

$$\frac{\partial \sigma_k^2}{\partial \lambda_n} = \sum_{j=1}^{n} \lambda_j \mathrm{C}_{nj} + \mu - \mathrm{C}_{n0} = \sum_{j=1}^{n} \lambda_j \big(\sigma^2 - \gamma_{nj}\big) + \mu - (\sigma^2 - \gamma_{n0}) \qquad \text{Ec. 8.29}$$

$$= \sigma^2 \sum_{j=1}^{n} \lambda_j - \sum_{j=1}^{n} \lambda_j \gamma_{nj} + \mu - (\sigma^2 - \gamma_{n0})$$

$$\Rightarrow \sum_{j=i}^{n} \lambda_j \gamma_{nj} - \mu = \gamma_{n0} \qquad \text{Ec. 8.30}$$

El sistema de ecuaciones se completa con Ec. 2.55. De acuerdo con lo anterior los pesos se obtienen en términos del semivariograma a través del sistema de ecuaciones:

$$\begin{pmatrix} \gamma_{11} & \cdots & \gamma_{1n} & 1 \\ \vdots & \ddots & \vdots & \vdots \\ \gamma_{11} & \cdots & \gamma_{nn} & 1 \\ 1 & \cdots & 1 & 0 \end{pmatrix} \begin{pmatrix} \lambda_i \\ \vdots \\ \lambda_n \\ -\mu \end{pmatrix} = \begin{pmatrix} \gamma_{10} \\ \vdots \\ \gamma_{n0} \\ 1 \end{pmatrix}$$

Para establecer la expresión de la correspondiente varianza del error de predicción en términos de la función de semivarianza se reemplaza Ec 8.28 en Ec. 2.61, de donde:

$$\sigma_k^2 = \sum_{i=1}^{n} \lambda_i \gamma_{i0} + \mu \qquad \text{Ec. 8.31}$$

Los pesos de Kriging ordinario también pueden ser estimados mediante el uso del correlograma aplicando la siguiente relación: $\rho_{ij} = \frac{c_{ij}}{\sigma^2}$, caso en el que la correspondiente varianza de predicción estaría dada por (Isaaks y Srivastava, 1989):

$$\sigma_k^2 = \sigma^2 \left(1 - \sum_{i=1}^{n} \lambda_i \gamma_{i0} + \mu \right) \qquad \text{Ec. 8.32}$$

8.13 Kriging Universal

En los supuestos hechos hasta ahora respecto al Kriging se ha asumido que la variable regionalizada es estacionaria, pero en muchos casos, la variable no satisface estas condiciones y se caracteriza por exhibir una tendencia. Para tratar este tipo de variables es frecuente descomponer la variable $z(x)$ como la suma de la tendencia, tratada como una función determinística $m(x)$, más una componente estocástica estacionaria $R(x)$:

$$z(x) = m(x) + R(x) \qquad \text{Ec. 8.33}$$

Donde $E[z(x)] = m(x)$ y $E[R(x)] = 0$

La tendencia puede expresarse como una suma ponderada de funciones conocidas $f_l(x)$ y coeficientes desconocidos a_l con $l = 0, \dots, L$, mediante:

$$m(x) = \sum_{l=0}^{L} a_l f_l(x), \qquad \text{Ec. 8.34}$$

Donde, por convenio $f_0(x) = 1$, para todo x.

El predictor de Kriging universal se defino como:

$$Z^*(x_0) \sum_{i=1}^{n} \lambda_i Z(x_i), \qquad \text{Ec. 8.35}$$

8.14 NUBE DE CORRELACIÓN DIFERIDA

Se trata de la nube de puntos $Z(x_\alpha), Z(x_\beta + h)$ donde h es un vector dado, mientras que x_α y $x_\beta + h$ son posiciones con datos. Los pares de datos con valores muy disímiles corresponderán a los puntos de la nube más alejados de la primera bisectriz.

Cuando los datos están ubicados en una grilla regular, se toma un vector h múltiplo del espaciamiento de esta grilla. En caso contrario, se debe introducir tolerancias en la longitud y la orientación de h, a falta de que la nube se reduciría a muy pocos puntos. La Figura 72 da una ilustración, para un vector h de longitud 0.25 km (con una tolerancia de 0.01 km) sin importar la orientación. Los puntos más alejados de la bisectriz han sido puestos en relieve y los pares de datos correspondientes han sido destacados en el mapa de ubicación: se trata de datos cercanos cuyos valores son muy diferentes.

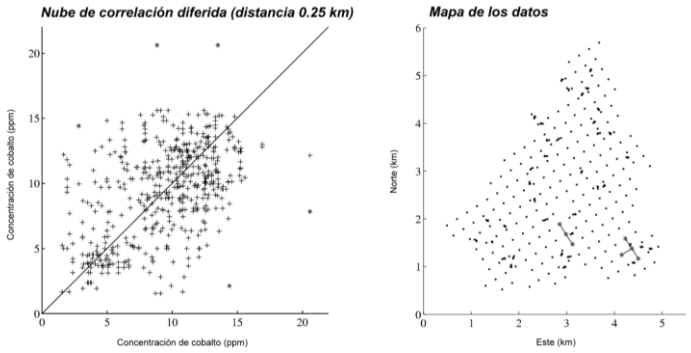


Figura 72: Nube de correlación diferida y mapa de ubicación de los datos.

8.15 TEOREMA DE BAYES

El teorema de Bayes es un procedimiento para obtener probabilidades condicionales (probabilidades de ocurrencia de acontecimientos condicionadas a la ocurrencia de otros acontecimientos). La expresión del teorema de Bayes para dos variables discretas es:

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{P(B)} = \frac{P(B|A_i)P(A_i)}{\sum p(B|A_i)P(A_i)} \quad \text{Ec. 8.36}$$

Donde,

- $P(A_i)$: Son las probabilidades a priori.
- $P(B|A_i)$: Es la probabilidad de B en la hipótesis A_i.
- $P(A_i|B)$: Son las probabilidades a posteriori.

Esto se cumple siempre que $\forall i = 1, \dots, n.$

El teorema de Bayes da respuesta a cuestiones de tipo causal, predictivas y de diagnóstico. En las cuestiones causales se quiere saber cuál es la probabilidad de acontecimientos que son la consecuencia de otros acontecimientos. En las cuestiones predictivas se quiere saber cuál es la probabilidad de acontecimientos dada información de la ocurrencia de los acontecimientos predictores. En las cuestiones de tipo diagnóstico se quiere saber cuál es la probabilidad del acontecimiento (o acontecimientos) causales o predictivos dado que tenemos información de las consecuencias.

8.16 PROBABILIDAD CONJUNTA

Si dos sucesos M y N son independientes, la probabilidad de ocurrencia de ambos sucesos simultáneamente será igual al producto de las probabilidades individuales.

$$P(M\ y\ N) = P(M\ \wedge N) = P(M)P(N) \quad \text{Ec. 8.37}$$

Si se tiene más de dos sucesos independientes, la probabilidad conjunta será igual al producto de las probabilidades de cada uno de los sucesos.

8.17 ESTADÍSTICAS DE LEY y EN LAS SIMULACIONES

Tabla 21: Estadísticas de la ley y en las simulaciones.

Simulación	Min.	1st Qu.	Mediana	Media	3rd Qu.	Max.
1	0.0000	0.4054	0.5434	0.5500	0.6891	1.267
2	0.0000	0.4164	0.5305	0.5500	0.6720	1.233
3	0.0000	0.4546	0.5463	0.5500	0.6430	1.255
4	0.0000	0.4112	0.5311	0.5500	0.6804	1.258
5	0.0000	0.4374	0.5485	0.5500	0.6579	1.158
6	0.0000	0.4370	0.5485	0.5500	0.6609	1.148
7	0.0000	0.4612	0.5552	0.5500	0.6442	1.019
8	0.0000	0.4180	0.5288	0.5500	0.6768	1.136
9	0.0000	0.4435	0.5452	0.5500	0.6544	1.013
10	0.0000	0.4437	0.5502	0.5500	0.6588	1.046
11	0.0000	0.4212	0.5490	0.5500	0.6750	1.123
12	0.0000	0.4356	0.5417	0.5500	0.6616	1.118
13	0.0000	0.4535	0.5505	0.5500	0.6478	1.097
14	0.0000	0.4441	0.5620	0.5500	0.6586	1.113
15	0.0000	0.4225	0.5325	0.5500	0.6571	1.212
16	0.0000	0.3996	0.5381	0.5500	0.6967	1.192
17	0.0000	0.4359	0.5570	0.5500	0.6698	1.083
18	0.0000	0.4488	0.5466	0.5500	0.6503	1.105
19	0.0000	0.4482	0.5533	0.5500	0.6587	1.155
20	0.0000	0.4330	0.5467	0.5500	0.6688	1.143
21	0.0000	0.4433	0.5507	0.5500	0.6584	1.317
22	0.0000	0.4559	0.5566	0.5500	0.6534	1.004
23	0.0000	0.4249	0.5489	0.5500	0.6762	1.163
24	0.0000	0.4534	0.5610	0.5500	0.6580	0.982
25	0.0000	0.4464	0.5456	0.5500	0.6540	1.111
26	0.0000	0.4413	0.5468	0.5500	0.6551	1.087
27	0.0000	0.4297	0.5376	0.5500	0.6606	1.335
28	0.0000	0.4147	0.5520	0.5500	0.6849	1.096
29	0.0000	0.4276	0.5441	0.5500	0.6648	1.248
30	0.0000	0.4664	0.5539	0.5500	0.6369	1.040
31	0.0000	0.4371	0.5514	0.5500	0.6626	1.107
32	0.0000	0.4455	0.5407	0.5500	0.6528	1.051
33	0.0000	0.4664	0.5505	0.5500	0.6368	1.026
34	0.0000	0.4290	0.5507	0.5500	0.6771	1.054
35	0.0000	0.4356	0.5464	0.5500	0.6605	1.154
36	0.0000	0.4322	0.5405	0.5500	0.6591	1.133
37	0.0000	0.4199	0.5462	0.5500	0.6712	1.353
38	0.0000	0.4435	0.5493	0.5500	0.6589	1.020
39	0.0000	0.4469	0.5573	0.5500	0.6580	1.098
40	0.0000	0.4315	0.5409	0.5500	0.6637	1.099
41	0.0000	0.4384	0.5473	0.5500	0.6575	1.079
42	0.0000	0.4270	0.5503	0.5500	0.6744	1.126
43	0.0000	0.4431	0.5520	0.5500	0.6563	1.208
44	0.0000	0.4561	0.5514	0.5500	0.6440	1.080
45	0.0000	0.4186	0.5424	0.5500	0.6736	1.190
46	0.0000	0.4332	0.5538	0.5500	0.6679	1.104
47	0.0000	0.4489	0.5549	0.5500	0.6511	1.034

48	0.0000	0.4196	0.5523	0.5500	0.6828	1.074
49	0.0000	0.4555	0.5527	0.5500	0.6451	1.033
50	0.0000	0.4389	0.5504	0.5500	0.6571	0.999
51	0.0000	0.4441	0.5468	0.5500	0.6541	1.208
52	0.0000	0.4166	0.5374	0.5500	0.6689	1.193
53	0.0000	0.4258	0.5440	0.5500	0.6664	1.209
54	0.0000	0.4290	0.5462	0.5500	0.6734	1.140
55	0.0000	0.4240	0.5406	0.5500	0.6659	1.172
56	0.0000	0.4136	0.5384	0.5500	0.6837	1.190
57	0.0000	0.4574	0.5530	0.5500	0.6454	1.040
58	0.0000	0.4519	0.5511	0.5500	0.6462	1.073
59	0.0000	0.4546	0.5476	0.5500	0.6463	1.080
60	0.0000	0.4508	0.5540	0.5500	0.6555	1.030
61	0.0000	0.4713	0.5600	0.5500	0.6416	1.017
62	0.0000	0.4519	0.5559	0.5500	0.6521	1.138
63	0.0000	0.4314	0.5529	0.5500	0.6690	1.099
64	0.0000	0.4417	0.5565	0.5500	0.6607	1.248
65	0.0000	0.4333	0.5435	0.5500	0.6641	1.154
66	0.0000	0.4496	0.5534	0.5500	0.6558	1.081
67	0.0000	0.4660	0.5580	0.5500	0.6475	0.991
68	0.0000	0.4077	0.5356	0.5500	0.6869	1.134
69	0.0000	0.4475	0.5495	0.5500	0.6526	1.095
70	0.0000	0.4211	0.5440	0.5500	0.6741	1.169
71	0.0000	0.4495	0.5468	0.5500	0.6500	1.018
72	0.0000	0.4159	0.5432	0.5500	0.6770	1.262
73	0.0000	0.4304	0.5447	0.5500	0.6612	1.267
74	0.0000	0.4525	0.5549	0.5500	0.6508	0.966
75	0.0000	0.4660	0.5503	0.5500	0.6361	1.006
76	0.0000	0.4524	0.5707	0.5500	0.6633	1.000
77	0.0000	0.4712	0.5553	0.5500	0.6334	0.974
78	0.0000	0.4460	0.5489	0.5500	0.6575	1.037
79	0.0000	0.4593	0.5563	0.5500	0.6450	0.959
80	0.0000	0.4456	0.5599	0.5500	0.6615	1.104
81	0.0000	0.4517	0.5387	0.5500	0.6350	1.097
82	0.0000	0.4537	0.5440	0.5500	0.6456	0.984
83	0.0000	0.4571	0.5547	0.5500	0.6509	0.989
84	0.0000	0.4389	0.5403	0.5500	0.6537	1.143
85	0.0000	0.4565	0.5516	0.5500	0.6448	1.013
86	0.0000	0.4323	0.5363	0.5500	0.6558	1.327
87	0.0000	0.4205	0.5454	0.5500	0.6679	1.226
88	0.0000	0.4431	0.5520	0.5500	0.6641	1.085
89	0.0000	0.4401	0.5554	0.5500	0.6636	1.125
90	0.0000	0.4520	0.5467	0.5500	0.6472	1.130
91	0.0000	0.4344	0.5431	0.5500	0.6620	1.094
92	0.0000	0.4326	0.5522	0.5500	0.6644	1.186
93	0.0000	0.4160	0.5393	0.5500	0.6707	1.256
94	0.0000	0.4567	0.5538	0.5500	0.6514	1.040
95	0.0000	0.4246	0.5559	0.5500	0.6734	1.164
96	0.0000	0.4377	0.5483	0.5500	0.6601	1.130
97	0.0000	0.4360	0.5494	0.5500	0.6676	1.109
98	0.0000	0.4531	0.5470	0.5500	0.6431	1.097
99	0.0000	0.4278	0.5497	0.5500	0.6662	1.217
100	0.0000	0.4276	0.5462	0.5500	0.6731	1.176

8.18 SONDAJES EN BASE DE ENTRENAMIENTO

Tabla 22: Sondajes en dase de entrenamiento y datos simulación 100.

Sondaje	Coordenada $[m]$			Promedio simulación 100			
	Easting	Northing	Elevation	y	v_1	v_2	e
1	90	30	80	0.61	0.55	1.21	0.16
2	10	170	110	0.46	0.63	1.69	-0.01
3	130	80	70	0.70	0.58	0.84	0.14
4	140	40	100	0.80	0.54	0.47	0.21
5	80	120	130	0.44	0.50	1.53	0.04
6	60	80	50	0.46	0.62	1.86	0.08
7	20	70	140	0.45	0.47	1.37	0.03
8	200	200	50	1.34	0.82	0.87	0.57
9	80	70	90	0.42	0.51	1.77	0.13
10	30	160	130	0.41	0.57	1.70	-0.03
11	40	20	130	0.55	0.50	1.10	0.07
12	70	170	80	0.45	0.59	1.61	-0.01
13	160	180	60	0.62	0.72	1.40	0.04
14	90	50	90	0.49	0.51	1.47	0.12
15	50	90	70	0.53	0.66	1.66	0.06
16	100	60	70	0.50	0.50	1.46	0.15
17	200	10	160	0.65	0.63	1.29	0.18
18	100	190	60	0.54	0.61	1.42	0.04
19	70	190	100	0.48	0.56	1.47	0.02
20	170	150	160	0.66	0.61	1.08	0.12
21	70	200	100	0.51	0.59	1.42	0.02
22	30	150	80	0.45	0.64	1.72	-0.03
23	90	60	70	0.47	0.50	1.63	0.16
24	50	10	90	0.65	0.54	0.91	0.12
25	120	110	150	0.55	0.46	1.04	0.10
26	80	180	120	0.47	0.53	1.45	0.03
27	200	180	70	0.76	0.72	1.03	0.14
28	190	130	120	0.74	0.67	0.91	0.12

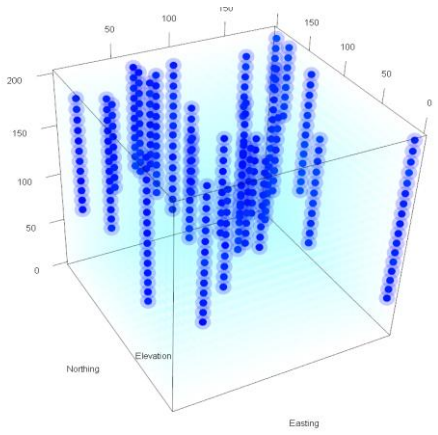


Figura 73: Vista Sondajes extraídos en simulaciones.

8.19 CÓDIGO REGRESIÓN BETA EN R

```
betareg (formula, data, subset, na.action, weights, offset,
        link = c ("logit", "probit", "cloglog", "cauchit", "log", "loglog"),
        link.phi = NULL, type = c ("ML", "BC", "BR"),
        control = betareg. control (…), model = TRUE,
        y = TRUE, x = FALSE, …)
betareg.fit(x, y, z = NULL, weights = NULL, offset = NULL,
        link = "logit", link.phi = "log", type = "ML", control=
      betareg.control())
```

https://www.rdocumentation.org/packages/betareg/versions/3.1-3/topics/betareg

1. *Formula:* : Descripción simbólica del modelo.
2. *data, subst, na.action* : Argumentos que controlan el procesamiento de fórmulas a través de `model.frame`.
3. *Weigths* : Vector numérico opcional de pesos de caja.
4. *Offset* : Vector numérico opcional con un componente conocido a priori que se incluirá en el predictor lineal de la media. En `betareg.fit`, `offset` también puede ser una lista de dos compensaciones para la ecuación de media y precisión, respectivamente.
5. *Link* : Especificación de caracteres de la función de enlace en el modelo medio (mu). Actualmente, se admiten `"logit"`, `"probit"`, `"cloglog"`, `"cauchit"`, `"log"`, `"loglog"` Alternativamente, se puede suministrar un objeto de clase `"link-glm"`.
6. *link.phi* : Especificación de caracteres de la función de enlace en el modelo de precisión (phi). Actualmente, se admiten `"identity"`, `"log"`, `"sqrt"`. El valor predeterminado es `"log"` a menos que la `formula` sea del tipo y ~ x donde el valor predeterminado es `"identity"` (para compatibilidad con versiones anteriores). Alternativamente, se puede suministrar un objeto de clase `"link-glm"`.
7. *type* : Especificación de caracteres del tipo de estimador. Actualmente, se admiten la máxima verosimilitud (`"ML"`), ML con corrección de sesgo (`"BC"`) y ML con reducción de sesgo (`"BR"`).
8. *control* : Una lista de argumentos de control especificados a través de `betareg.control`.
9. *model, y, x* : Lógicas. Si es `TRUE`, se devuelven los componentes correspondientes del ajuste (marco del modelo, respuesta, matriz del modelo). Para `betareg.fit`, `x` debería ser una matriz regresora numérica e `y` debería ser el vector de respuesta numérica (con valores en (0,1)).
10. *z* : Matriz numérica. Matriz regresora para el modelo de precisión, por defecto solo en una intersección.

8.20 CÓDIGO REGRESIÓN GEOGRÁFICAMENTE PONDERADA EN R

```
gwr.basic(formula, data, regression.points, bw, kernel="bisquare",
  adaptive=FALSE, p=2, theta=0, longlat=F, dMat, F123.test=F, cv=F, W.vect=NULL)
```

https://www.rdocumentation.org/packages/GWmodel/versions/2.1-4/topics/gwr.basic

1. *Formula* : Descripción simbólica del modelo.
2. *Data* : Argumentos que controlan el procesamiento de fórmulas a través de `model.frame`.
3. *Regression.points* : Objeto `Spatial.DataFrame`, es decir, `SpatialPointsDataFrame` o `SpatialPolygonsDataFrame` como se define en el paquete `sp`.
4. *bw* : Ancho de banda utilizado en la función de ponderación, posiblemente calculado por `bw.gwr`; ancho de banda fijo (distancia) o adaptativo (número de vecinos más cercanos).
5. *kernel* : Funcion nucleo escogida, que puede ser *gaussian, exponential, bisqyare, tricubeo* o *boxcar.*
6. *Adaptive* : Si es `TRUE`, calcula un kernel adaptativo donde el ancho de banda `bw` corresponde al número de vecinos más cercanos (es decir, distancia adaptativa); de lo contrario si el valor predeterminado es `FALSE`, donde se encuentra un kernel fijo (el ancho de banda es una distancia fija).
7. *p* : Valor de la distancia de *Minkowsk*i, el valor predeterminado es 2, es decir, la distancia euclidiana.
8. *Theta* : Ángulo en radianes para rotar el sistema de coordenadas, el valor predeterminado es 0.
9. *Longlat* : Si es `TRUE`, se calcularán en distancias circulares.
10. *dMat* : una matriz de distancia predefinida, se puede calcular mediante la función `gw.dist`.
11. *F123.test* : Si es `TRUE`, realice tres pruebas F separadas de acuerdo con Leung et al. (2000).
12. *Cv* : Si es `TRUE,` los datos de validación cruzada se calcularán y devolverán en la salida `Spatial. DataFrame`.
13. *W.vect* : `NULL` predeterminado, si se proporciona, se utiliza para ponderar la matriz de ponderación de distancia.
14. *X* : un objeto de clase `"gwrm"`, devuelto por la función `gwr.basic`.

8.21 CÓDIGO KRIGING BAYESIANO EN R

```
krige.bayes(geodata, coords = geodata$coords, data = geodata$data,
            locations = "no",
     model = model.control(trend.d = "cte", trend.l = "cte",
                  cov.model = "exponential", kappa = 0.5,
                  aniso.pars = NULL, lambda = 1),
     prior = prior.control(beta.prior = c("flat", "normal",
                                          "fixed"),
                  beta = NULL, beta.var = NULL,
                  sill.prior = c("reciprocal", "fixed"),
                  sill = NULL,
                  range.prior = c("uniform", "exponential",
                                  "fixed", "squared.reciprocal",
                                  "reciprocal"),
                  exponential.prior.par = 1, range = NULL,
                  range.discrete = NULL,
                  nugget.prior = c("fixed", "uniform"),
                  nugget = 0, nugget.discrete = NULL),
     output = output.control(n.posterior = 1000,
                  n.predictive = NULL, moments = TRUE,
                  simulations.predictive = TRUE,
                  keep.simulations = TRUE, mean.estimator = TRUE,
                  quantile.estimator =  NULL,
                  probability.estimator = NULL,
                  signal = FALSE, messages.screen = TRUE))
```

https://www.rdocumentation.org/packages/geoR/versions/1.0-0/topics/krige.bayes

1. *Geodata* : Lista que contiene elementos de `coords` y `data` como se describe a continuación. Normalmente, un objeto de la clase `geodata`: un conjunto de datos geográficos. Si no se proporcionan los argumentos, `coords` y `data` m.

2. *Coords* : una matriz $N \times 2$ donde cada fila tiene las coordenadas $2D$ de las n ubicaciones de datos. De forma predeterminada, toma las coordenadas de los componentes del argumento `geodata`, si se proporcionan.

3. *Data* : Un vector con N valores de datos. De forma predeterminada, toma los datos del componente de los geodatos del argumento, si se proporcionan.

4. *Locations* : Matriz o marco de datos $N \times 2$ con las coordenadas $2D$ de las N ubicaciones de predicción. El valor predeterminado es `"No"`, en cuyo caso la función solo devuelve resultados en las distribuciones posteriores de los parámetros del modelo.

5. *Model* : define los componentes del modelo

6. *Prior* : especificación de los `pior` para los parámetros del modelo

7. *Outputs* : Define las opciones de salida.

8.22 RESULTADOS RMSE PARA CASO SIMULADO

Tabla 23: Resultados RMES para caso simulado.

Simulación	Ley y				Ley v_1 y v_2	
	BREG	GWR	BY	OK	OK v_1	OK v_2
1	0.111	0.111	0.113	0.103	0.112	0.303
2	0.136	0.133	0.134	0.115	0.112	0.402
3	0.115	0.113	0.117	0.115	0.117	0.351
4	0.132	0.135	0.134	0.140	0.129	0.245
5	0.133	0.130	0.131	0.122	0.119	0.302
6	0.109	0.109	0.110	0.102	0.112	0.366
7	0.087	0.085	0.088	0.087	0.101	0.316
8	0.138	0.136	0.138	0.129	0.160	0.247
9	0.109	0.107	0.109	0.098	0.124	0.346
10	0.095	0.097	0.097	0.089	0.122	0.360
11	0.120	0.121	0.124	0.124	0.127	0.237
12	0.113	0.111	0.114	0.116	0.120	0.229
13	0.143	0.140	0.143	0.124	0.123	0.303
14	0.114	0.112	0.118	0.115	0.119	0.243
15	0.112	0.105	0.112	0.109	0.117	0.222
16	0.125	0.122	0.122	0.117	0.116	0.306
17	0.130	0.121	0.129	0.121	0.128	0.290
18	0.144	0.141	0.134	0.136	0.145	0.263
19	0.106	0.105	0.107	0.103	0.135	0.262
20	0.126	0.125	0.127	0.117	0.126	0.273
21	0.120	0.116	0.112	0.109	0.107	0.280
22	0.139	0.137	0.136	0.132	0.131	0.355
23	0.142	0.140	0.142	0.134	0.117	0.349
24	0.117	0.114	0.116	0.108	0.129	0.311
25	0.103	0.098	0.109	0.095	0.134	0.345
26	0.122	0.119	0.118	0.118	0.126	0.279
27	0.125	0.122	0.117	0.115	0.108	0.379
28	0.102	0.097	0.107	0.105	0.107	0.253
29	0.138	0.135	0.137	0.133	0.124	0.296
30	0.126	0.125	0.125	0.128	0.126	0.296
31	0.124	0.117	0.119	0.114	0.141	0.359
32	0.144	0.142	0.143	0.148	0.134	0.364
33	0.143	0.139	0.144	0.140	0.145	0.283
34	0.139	0.139	0.138	0.143	0.135	0.264
35	0.108	0.107	0.108	0.106	0.110	0.237
36	0.145	0.143	0.145	0.141	0.128	0.283
37	0.145	0.145	0.139	0.135	0.132	0.332
38	0.139	0.135	0.131	0.125	0.124	0.339
39	0.118	0.114	0.116	0.120	0.128	0.278
40	0.184	0.183	0.170	0.158	0.129	0.395
41	0.114	0.110	0.110	0.103	0.131	0.296
42	0.112	0.114	0.124	0.114	0.118	0.228
43	0.099	0.099	0.099	0.111	0.117	0.311
44	0.138	0.139	0.137	0.142	0.137	0.310
45	0.106	0.106	0.105	0.102	0.105	0.241
46	0.131	0.125	0.117	0.109	0.117	0.418

47	0.122	0.120	0.120	0.126	0.140	0.223
48	0.115	0.110	0.112	0.108	0.121	0.280
49	0.113	0.108	0.118	0.118	0.132	0.308
50	0.130	0.130	0.131	0.126	0.111	0.398
51	0.110	0.110	0.112	0.114	0.121	0.266
52	0.124	0.120	0.112	0.120	0.117	0.249
53	0.175	0.166	0.173	0.159	0.118	0.348
54	0.120	0.113	0.117	0.116	0.123	0.251
55	0.127	0.125	0.123	0.119	0.098	0.250
56	0.130	0.129	0.127	0.130	0.108	0.322
57	0.123	0.122	0.123	0.121	0.129	0.319
58	0.153	0.155	0.157	0.152	0.168	0.351
59	0.111	0.109	0.118	0.103	0.116	0.213
60	0.143	0.139	0.141	0.133	0.133	0.312
61	0.100	0.096	0.103	0.098	0.111	0.271
62	0.129	0.126	0.124	0.121	0.128	0.286
63	0.130	0.134	0.123	0.116	0.120	0.337
64	0.126	0.126	0.129	0.133	0.127	0.231
65	0.117	0.109	0.117	0.100	0.098	0.392
66	0.123	0.118	0.118	0.119	0.088	0.271
67	0.113	0.110	0.116	0.103	0.109	0.241
68	0.102	0.098	0.104	0.105	0.122	0.274
69	0.122	0.119	0.119	0.116	0.131	0.233
70	0.142	0.144	0.151	0.147	0.123	0.384
71	0.128	0.122	0.125	0.114	0.135	0.282
72	0.104	0.102	0.106	0.105	0.121	0.246
73	0.131	0.128	0.130	0.124	0.125	0.273
74	0.143	0.143	0.144	0.149	0.138	0.262
75	0.143	0.140	0.145	0.127	0.144	0.294
76	0.140	0.138	0.139	0.135	0.142	0.391
77	0.124	0.119	0.129	0.124	0.120	0.199
78	0.141	0.141	0.158	0.152	0.116	0.377
79	0.131	0.126	0.133	0.129	0.139	0.272
80	0.107	0.106	0.107	0.115	0.120	0.320
81	0.155	0.155	0.150	0.142	0.140	0.296
82	0.120	0.123	0.121	0.116	0.116	0.339
83	0.152	0.143	0.134	0.140	0.118	0.277
84	0.123	0.120	0.122	0.119	0.126	0.336
85	0.111	0.109	0.110	0.113	0.094	0.220
86	0.126	0.122	0.139	0.118	0.143	0.287
87	0.102	0.102	0.099	0.104	0.109	0.191
88	0.142	0.140	0.135	0.124	0.124	0.306
89	0.124	0.122	0.121	0.113	0.139	0.357
90	0.137	0.140	0.133	0.140	0.136	0.281
91	0.123	0.116	0.120	0.107	0.126	0.358
92	0.147	0.143	0.145	0.137	0.130	0.322
93	0.103	0.105	0.106	0.107	0.124	0.398
94	0.079	0.078	0.079	0.076	0.100	0.327
95	0.111	0.110	0.112	0.109	0.117	0.328
96	0.156	0.147	0.157	0.154	0.120	0.346
97	0.097	0.097	0.099	0.092	0.102	0.296
98	0.130	0.124	0.132	0.125	0.126	0.349
99	0.156	0.151	0.139	0.151	0.094	0.380
100	0.144	0.141	0.148	0.146	0.135	0.319

UNIVERSIDAD DE CONCEPCIÓN - FACULTAD DE INGENIERÍA
Departamento de Ingeniería Metalúrgica
Hoja Resumen Memoria de Título

Título: "Modelos de Regresión en evaluación de yacimientos"	
Nombre Memorista: Paula Francisca González Fariña	
Modalidad	**Investigación**
Concepto	Sobresaliente
Calificación	7.0
Fecha	16-11-2020
Profesor (es) Patrocinante (s)	
Prof. Roberto Fustos T. - Prof. Francisco Muñoz G.	
Prof. Fernando Parada	**Ingeniero Supervisor:**
	Institución:
Comisión (Nombre y Firma)	
Prof. Jean Navarrete C.	

Resumen

La estimación de recursos minerales (ERM) es un proceso indispensable para el desarrollo de un proyecto minero, esto producto que en esta etapa se puede determinar la continuidad o no continuidad del proyecto, basados en si el depósito es capaz de proporcionar un beneficio económico.

En consecuencia, es necesario que la ERM sea lo más precisa posible, pero en la realidad producto de la baja cantidad de información que se disponen provenientes de sondajes de exploración se hace imposible que el proceso de la adquisición, procesamiento e interpretación de estos datos este lleno de subjetividades por parte del personal que maneja la información lo que puede conllevar a tener una errónea o pobre interpretación de la distribución real del mineral del depósito estudiado.

La problemática que se aborda en esta Memoria de Título se relaciona directamente con los modelos de regresión lineal utilizados en la actualidad para la realización de la ERM analizando las principales limitaciones y complicaciones de estos modelos a la hora de realizar la ERM.

La metodología consiste en la construcción de dos escenarios, el posterior análisis de los resultados y una categorización de recursos; el primero de estos escenarios se basa en la construcción de un conjunto de datos simulado con el que se obtendrá un modelo de bloques, de este se selecciona un porcentaje datos y se aplican los modelos de regresión: Regresión Beta, Regresión Geográficamente Ponderada, Kriging Bayesiano, Inverso de la distancia y Kriging ordinario generando nuevos modelos de bloques, finalmente se comparan ambos modelos de bloques a través de la raíz del error cuadrático medio (RMSE); en el segundo escenario se realiza una ERM a un conjunto de datos reales de campo repitiendo el procedimiento anterior.

Para el caso simulado se destaca la gran eficacia de los cinco métodos para la estimación, presentado los menores resultados de RMSE el método Kriging Bayesiano, seguido de Kriging Ordinario utilizando un menor costo computacional, y por el caso contrario presentando los mayores resultados de RMSE se encuentra el método Regresión beta.

En cuanto al caso real fue necesario optimizar los parámetros de los algoritmos a través de validación cruzada teniendo el menor resultado de RMSE el método Inverso de la distancia y el mayor valor de RMSE el método Regresión Beta.