Estimación Tridimensional de la Concentración de un contaminante en napas subterráneas

Project [MAT-288: Modeling Laboratory II]

Profesor : Pablo Aguirre
Tutor : Jason Gerhard

Gabriel Molina H.

Universidad Técnica Federico Santa María

Jueves 1 de Diciembre 2016

Contenidos

Contenidos

Contaminación en Napas Subterráneas

- Las napas subterráneas son capas de suelo con alto contenido de agua en sus poros o fisuras
- En la actualidad, principalmente en EEUU, existen filtraciones de contaminante en los suelos por acciones de las masas de agua subterráneas. Estos contaminantes pueden llegar a zonas pobladas por acción de dichas napas, y esto es un especial problema en el caso de contaminantes fósiles.

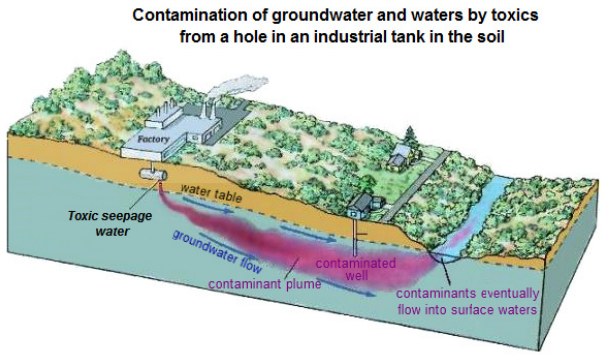


Figure: Ejemplo de contaminación por acción de napas

Contaminación en Napas Subterráneas

- Las napas subterráneas son capas de suelo con alto contenido de agua en sus poros o fisuras
- En la actualidad, principalmente en EEUU, existen filtraciones de contaminante en los suelos por acciones de las masas de agua subterráneas. Estos contaminantes pueden llegar a zonas pobladas por acción de dichas napas, y esto es un especial problema en el caso de contaminantes fósiles.

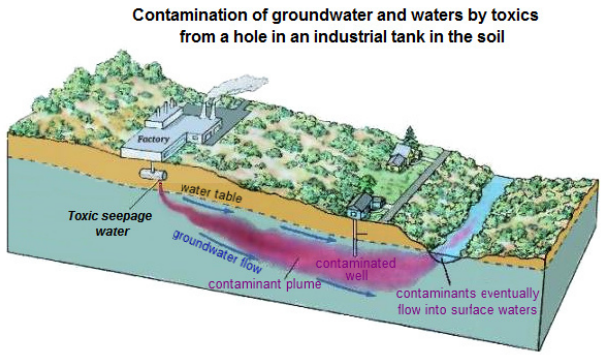


Figure: Ejemplo de contaminación por acción de napas

El Problema

- C_w es una medida de la concentración de un contaminante.
  El objetivo principal es que a partir de N muestras espaciales de C_w, podamos estimar C_w para un dominio completo.
- Para esto contamos con 625000 valores de C_w en un dominio tridimensional, estos valores fueron obtenidos a partir de una simulación validada.

El Problema

- C_w es una medida de la concentración de un contaminante.
  El objetivo principal es que a partir de N muestras espaciales de C_w, podamos estimar C_w para un dominio completo.
- Para esto contamos con 625000 valores de C_w en un dominio tridimensional, estos valores fueron obtenidos a partir de una simulación validada.

Cómo lucen los datos?: Dominio

- Se cuenta con 625000 valores de C_w asociados a un domino tridimensional $[0, 1000] \times [0, 500] \times [0, 10]$ como se ilustra en la figura

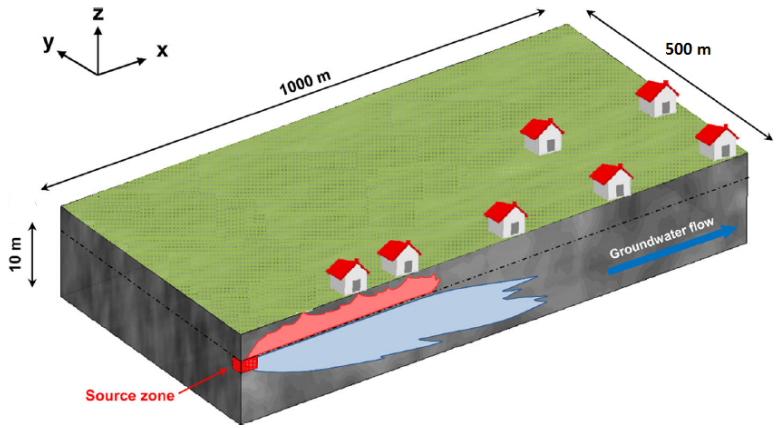


Figure: Sketch domain.

- Los valores de C_w fueron otorgados en los nodos de una grilla con $\Delta x = 4$, $\Delta y = 2$ y $\Delta z = 1$.

Cómo lucen los datos?: Dominio

- Se cuenta con 625000 valores de C_w asociados a un domino tridimensional $[0, 1000] \times [0, 500] \times [0, 10]$ como se ilustra en la figura

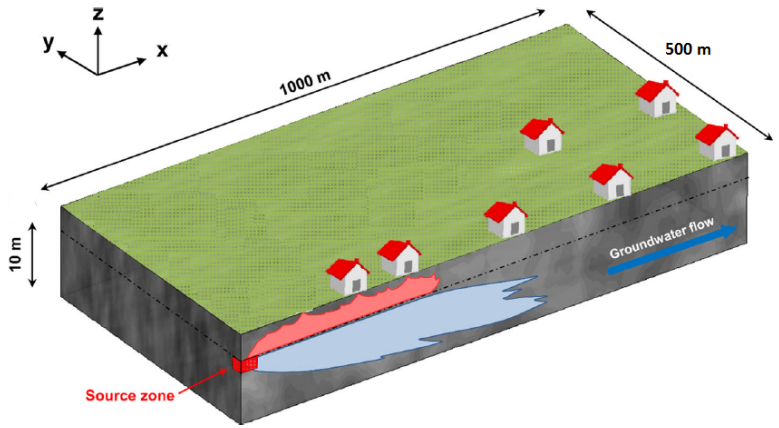


Figure: Sketch domain.

- Los valores de C_w fueron otorgados en los nodos de una grilla con $\Delta x = 4$, $\Delta y = 2$ y $\Delta z = 1$.

Cómo lucen los datos?: Dominio

- En la primera etapa (Laboratorio de Modelación I) se realizaron 10 predicciones bidimensionales, haciendo estimaciones para 10 valores (niveles) de z fijos, para distintos valores de N, intentando entregar un valor de N adecuado.

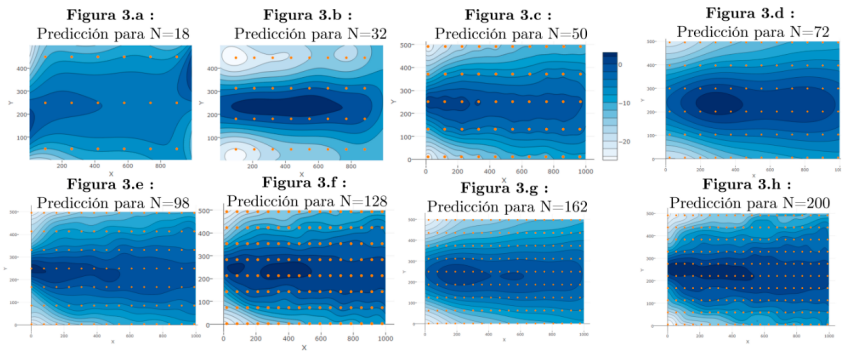


Figure: Sketch domain.

- El principal inconveniente del método anterior es que no se toma en cuenta relación entre capas, pudiendo omitir información relevante, por lo que propone desarrollar un modelo 3D para esta segunda parte del proyecto.

Cómo lucen los datos?: Dominio

- En la primera etapa (Laboratorio de Modelación I) se realizaron 10 predicciones bidimensionales, haciendo estimaciones para 10 valores (niveles) de z fijos, para distintos valores de N, intentando entregar un valor de N adecuado.

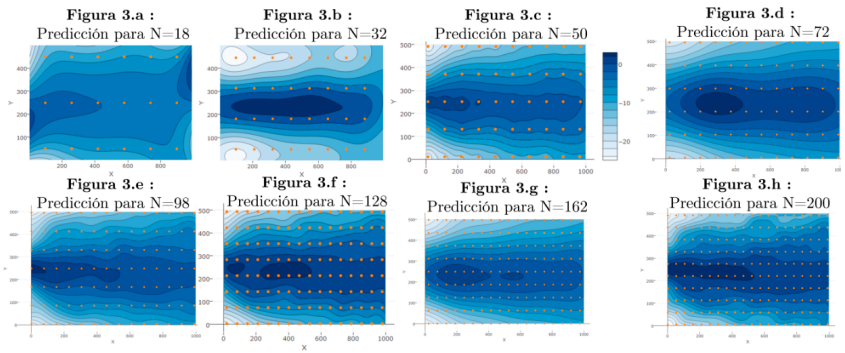
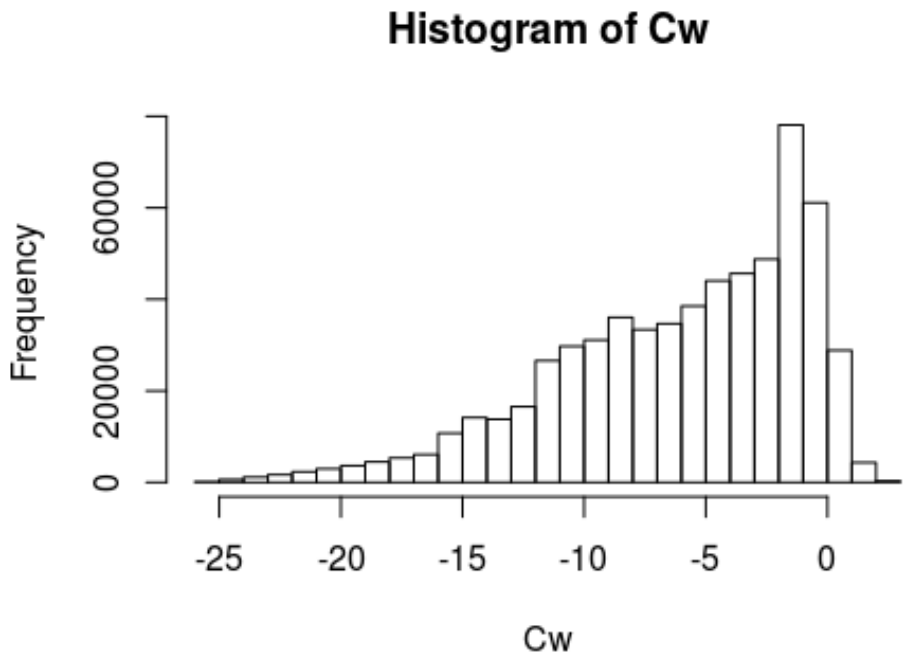


Figure: Sketch domain.

- El principal inconveniente del método anterior es que no se toma en cuenta relación entre capas, pudiendo omitir información relevante, por lo que propone desarrollar un modelo 3D para esta segunda parte del proyecto.

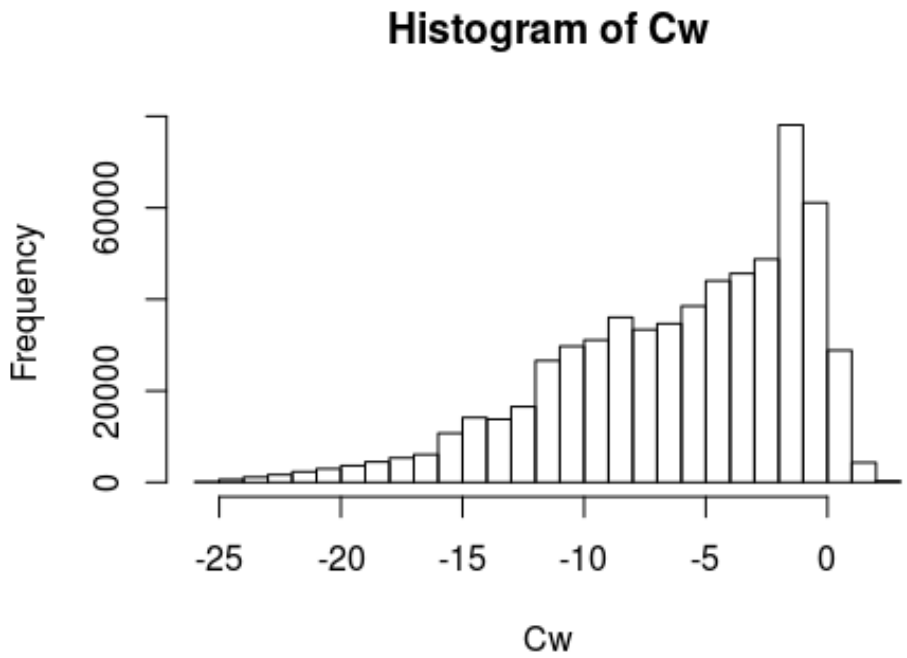
Cómo lucen los datos: Análisis exploratorio de C_w

- **En la realidad no contamos con esta información!**, pero en nuestro caso (dado que C_w proviene de una simulación) será de utilidad realizar un análisis exploratorio en para fines comparativos con nuestra propuesta de predicción.
- Se visualizan y analizan los valores de C_w.



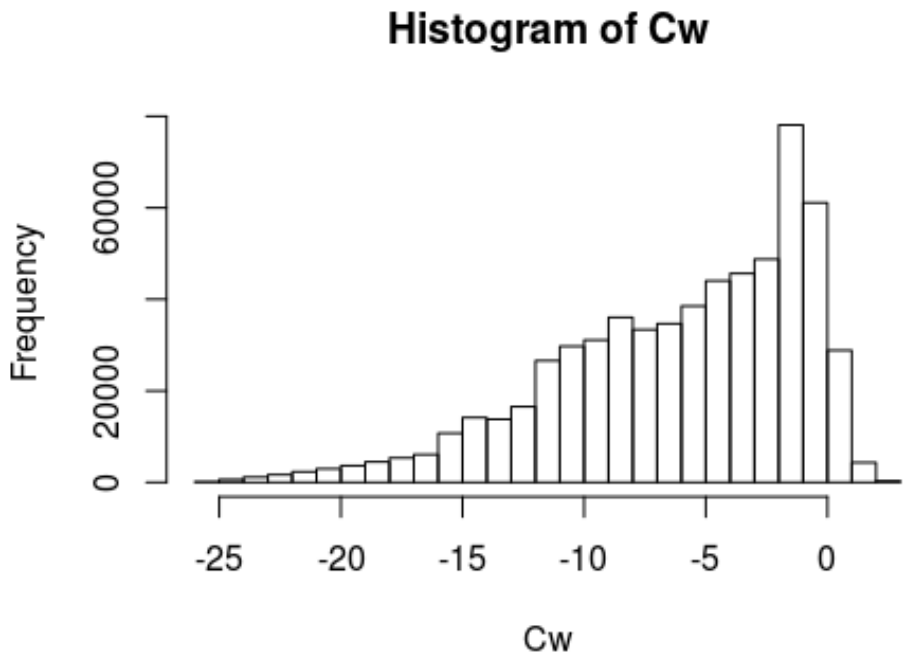
Cómo lucen los datos: Análisis exploratorio de C_w

- **En la realidad no contamos con esta información!**, pero en nuestro caso (dado que C_w proviene de una simulación) será de utilidad realizar un análisis exploratorio en para fines comparativos con nuestra propuesta de predicción.
- Se visualizan y analizan los valores de C_w.



Cómo lucen los datos: Análisis exploratorio de C_w

- **En la realidad no contamos con esta información!**, pero en nuestro caso (dado que C_w proviene de una simulación) será de utilidad realizar un análisis exploratorio en para fines comparativos con nuestra propuesta de predicción.
- Se visualizan y analizan los valores de C_w.



Cómo lucen los datos: Análisis exploratorio de C_w

- En la siguiente figura podemos visualizar los cuartiles de C_w (creciendo en intensidad de naranjos) en el dominio.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-25.870	-9.474	-5.030	-6.115	-1.761	2.740

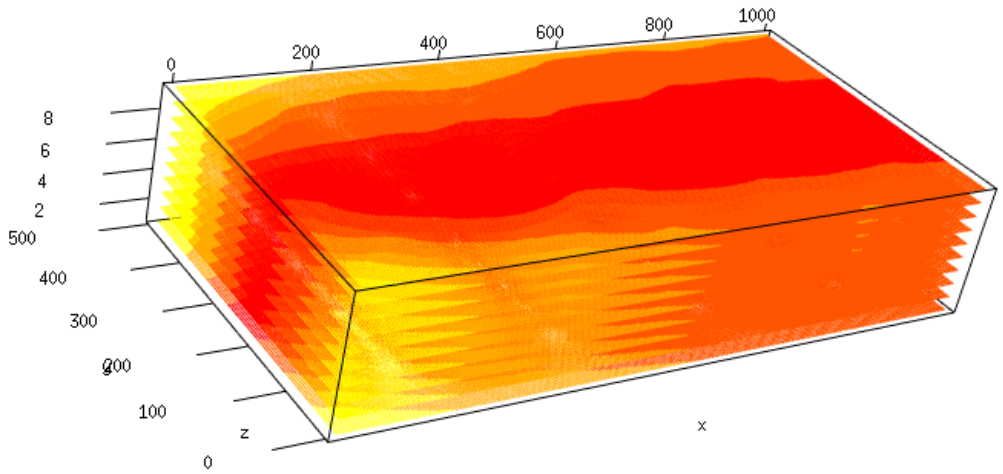


Figure: Visualización de los Cuartiles de Cw en el dominio 3D

Contents

Estimación basada en análisis del variograma
Idea

- Sea $D = \{s \in [0, 1000] \times [0, 500] \times [0, 10] \subset \mathbb{R}^2\}$ el dominio 3D, se define $Z(s)$, $s \in D$ como una variable aleatoria, donde $Z(s)$ representa el valor de C_w en el punto s
- Suponga que $Z(s) = \mu(s) + e(s)$ con μ constante pero desconocido y $e(s) \sim (0, \Sigma)$. Queremos obtener un estimador $\hat{Z}(s_0)$ para C_w in s_0 dado que conocemos N valores de C_w en los puntos $\{s_1, s_2, ..., s_N\}$ con sus correspondientes valores $\{Z(s_1), Z(s_2), ..., Z(s_N)\}$.

Estimación basada en análisis del variograma
Idea

- Sea $D = \{s \in [0, 1000] \times [0, 500] \times [0, 10] \subset \mathbb{R}^2\}$ el dominio 3D, se define $Z(s)$, $s \in D$ como una variable aleatoria, donde $Z(s)$ representa el valor de C_w en el punto s
- Suponga que $Z(s) = \mu(s) + e(s)$ con μ constante pero desconocido y $e(s) \sim (0, \Sigma)$. Queremos obtener un estimador $\hat{Z}(s_0)$ para C_w in s_0 dado que conocemos N valores de C_w en los puntos $\{s_1, s_2, ..., s_N\}$ con sus correspondientes valores $\{Z(s_1), Z(s_2), ..., Z(s_N)\}$.

Estimación basada en análisis del variograma

Definiciónn 1

Se define la función **Variograma** $\gamma(h)$ como:

$$\gamma(h) = \frac{1}{2}\mathbb{V}[Z(s+h) - Z(s)]$$

Es importante mnencionar que existen contadas **familias predefinidas** de variogramas que se pueden asociar a un proceso aleatorio.

Definition 2

Definimos el **Variograma Empírico** $\hat{\gamma}(h)$ como:

$$\hat{\gamma}(h) = \frac{1}{2N(h)} \sum_{s_i, s_j \in N(h)} (Z(s_i) - Z(s_j))^2$$

$$N(h) := \{(s_i, s_j) : \|s_i - s_j\| = h, \quad s_i, s_j \in D\}$$

Estimación basada en análisis del variograma

- Ahora cuando tenemos $\hat{\gamma}(h)$ para todas las distancias h muestreadas del dominio D es posible ajustar (minimizando Error cuadrático medio) alguna función teórica $\gamma(h)$ de las familias predefinidas, obteniendo así el variograma teórico asociado al proceso aleatorio $Z(s)$.
- El inconveniente en este caso es que en la actualidad existen pocas familias de variogramas asociados a procesos tridimensionales (la mayoría son bidimensionales), y el ajuste puede ser inconcluyente.

Estimación basada en análisis del variograma

- Ahora cuando tenemos $\hat{\gamma}(h)$ para todas las distancias h muestreadas del dominio D es posible ajustar (minimizando Error cuadrático medio) alguna función teórica $\gamma(h)$ de las familias predefinidas, obteniendo así el variograma teórico asociado al proceso aleatorio $Z(s)$.
- El inconveniente en este caso es que existen pocas familias de variogramas asociados a procesos tridimensionales, y el ajuste puede ser inconcluyente.

Estimación basada en análisis del variograma

- Opción 1: Intentar ajustar un variograma a capas bidimensionales horizontales y verticales, y lograr asociar una Matriz que ajuste ambas direcciones para un único variograma 3D... Inconcluyente.
- Opción 2: Investigar y asociar un variograma 3D apropiado para $Z(s)$.
  Variograma de Matern 3D:

$$\gamma(h) = c_0 + c_1 \left(1 - \frac{1}{2^{\nu-1}}\Gamma(\nu)\left(\frac{h}{a}\right)^{\nu} K_\nu\left(\frac{h}{a}\right)\right)$$

  donde K_ν es una función de Bessel Modificada de segundo tipo de orden ν. Γ es la función gamma.

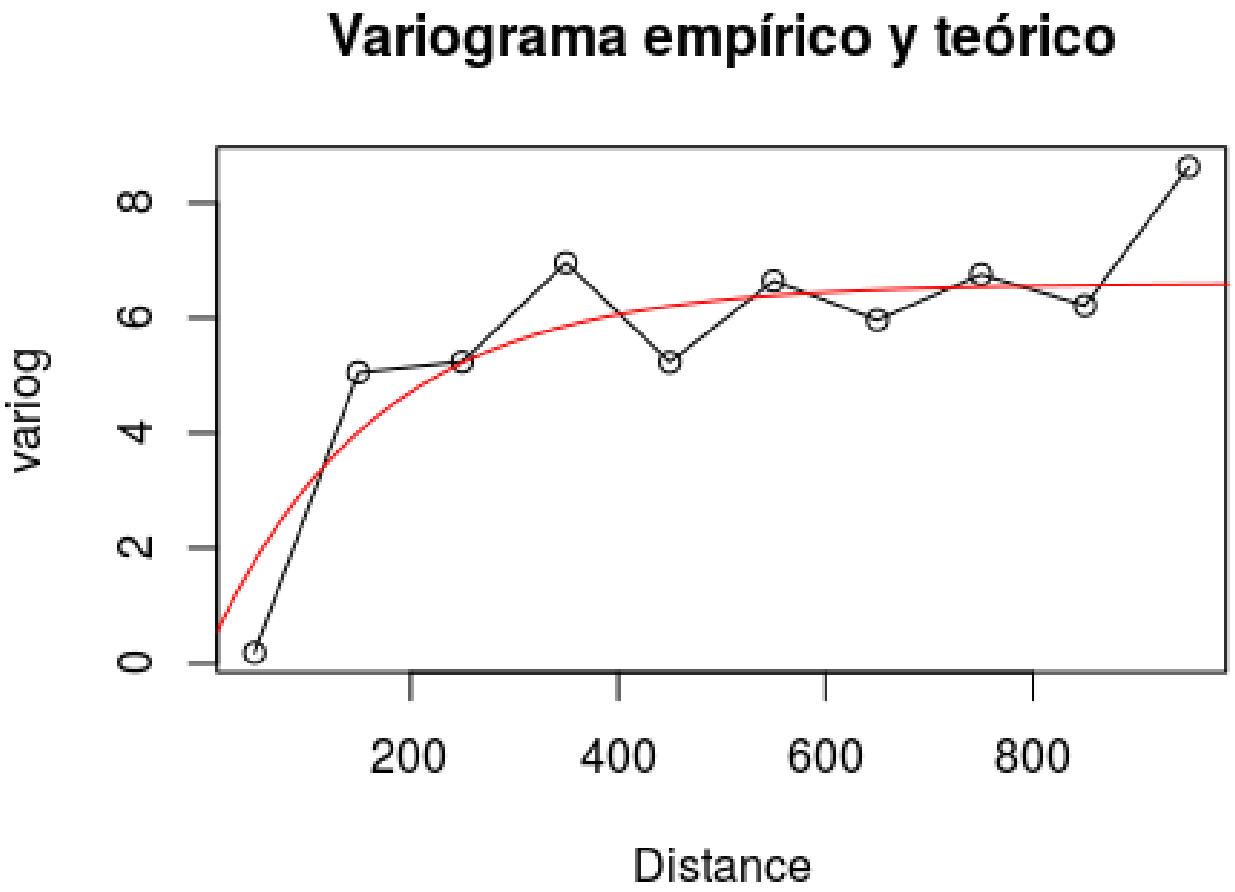


Figure: Puntos del variograma empírico y en rojo la curva variograma de MATERN 3D ajustado al proceso.

Estimación basada en análisis del variograma

Entonces se busca obtener el mejor estimador de la forma

$$\hat{Z}(s_0) = \lambda_0 + \sum_{i=1}^{N} \lambda_i Z(s_i)$$

Donde demandamos que $\hat{Z}(s_0)$ sea un estimador insesgado de mínima varianza.(ie: $\mathbb{E}[\hat{Z}(s_0) - Z(s_0)] = 0$ y se minimiza $\sigma_E^2 = \mathbb{V}[\hat{Z}(s_0) - Z(s_0)]$)
Lo anterior se reduce a resolver el sistema:

$$\begin{bmatrix} \gamma(s_1 - s_1) & \cdots & \gamma(s_1 - s_N) & 1 \\ \vdots & \ddots & \vdots & 1 \\ 1 & \cdots & 1 & 0 \end{bmatrix} \begin{bmatrix} \lambda_1^{ok} \\ \vdots \\ \lambda_N^{ok} \\ \mu_{ok} \end{bmatrix} = \begin{bmatrix} \gamma(s_1 - s_0) \\ \vdots \\ \gamma(s_N - s_0) \\ 1 \end{bmatrix}$$

Resultado

Se utilizó un muestreo con $N = 250$, $\Delta x = 220$, $\Delta y = 110$ y $\Delta z = 1$.

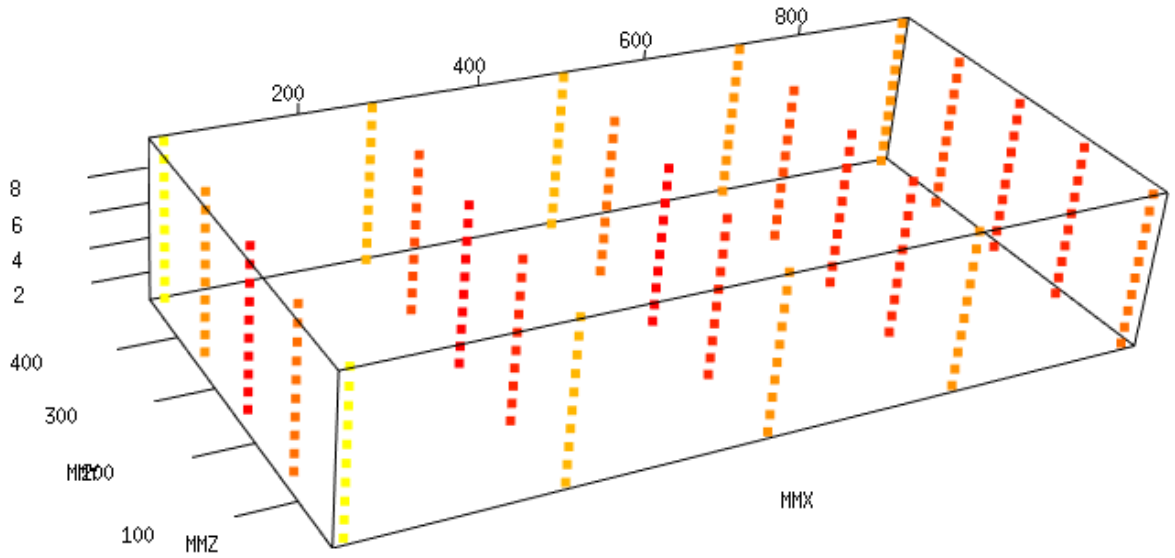


Figure: Visualización de la muestra tomada

Resultado

Para $N = 250$

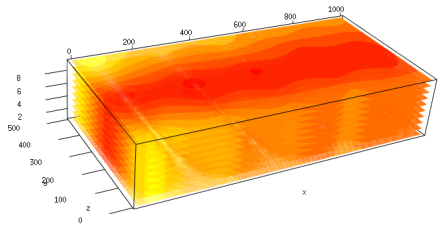


Figure: Visualización de la Predicción de Cw para N=250

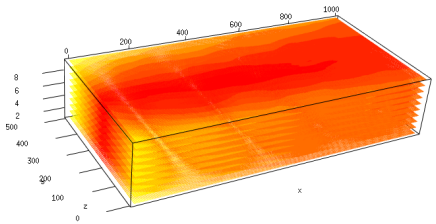


Figure: Visualización de los valores reales de Cw

Resultado

Histogramas

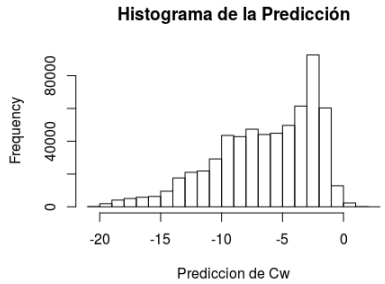


Figure: Histograma de la predicción de Cw

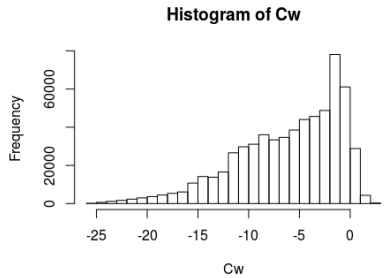


Figure: Histograma de los valores reales de Cw

Resultado

Analizamos ahora el residuo $e(s) = Z(s) - \hat{Z(s)}$

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-9.4170	-0.3507	0.5153	0.3310	1.2160	6.2950

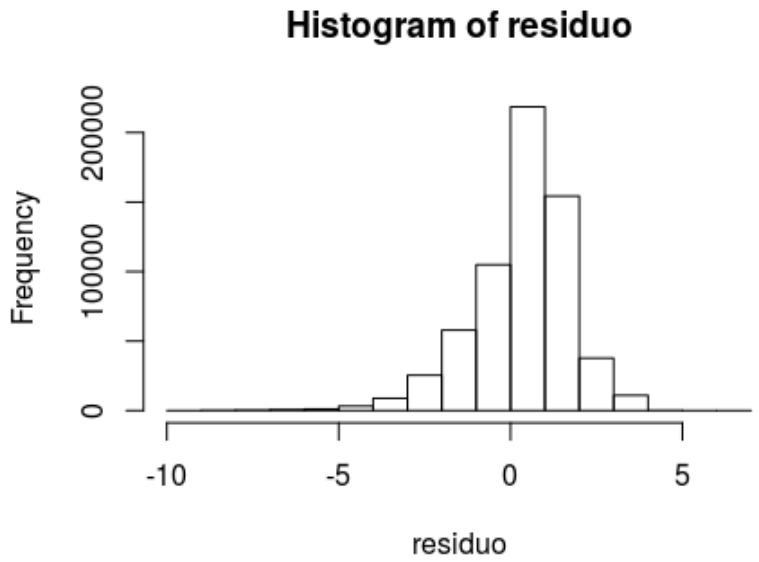


Figure: Histograma de los residuos

Resultado

Si los residuos son independientes al proceso original, se otorga una buena predicción, en nuestro caso el coeficiente de Tjostheim $\alpha \in [-1, 1]$ (medida de asociación de procesos espaciales) obtenido fué $\alpha = 0.008308827$, lo que nos hace sospechar independencia. Al aplicar test Chi-cuadrado se pudo confirmar nuestra hipótesis.

Results

El error cuadrático medio obtenido fué

$$ECM = 2.108657$$

Tabla de Contenidos

Conclusions

- Fué entregado un método tridimensional para estimar C_w con un error cuadrático medio 2.108657 para N=250.
- En este caso se obtiene un mejor desempeño para capas bidimensionales
- El método propuesto es rápido computacionalmente (en mi caso particular sólo 228.37 segundos sólo dedicados a la predicción) en contraste a métodos convencionales basados en resolución de EDP.
- El modelo no hace supuestos sobre las condiciones del suelo, por lo que es más flexible en comparación las alternativas presentes hoy en día
- Es posible generar una serie de tiempo de C_w a futuro.

Conclusions

- Fué entregado un método tridimensional para estimar C_w con un error cuadrático medio 2.108657 para N=250.
- En este caso se obtiene un mejor desempeño para capas bidimensionales
- El método propuesto es rápido computacionalmente (en mi caso particular sólo 228.37 segundos sólo dedicados a la predicción) en contraste a métodos convencionales basados en resolución de EDP.
- El modelo no hace supuestos sobre las condiciones del suelo, por lo que es más flexible en comparación las alternativas presentes hoy en día
- Es posible generar una serie de tiempo de C_w a futuro.

	Tiempo Propuesto[h]	Tiempo Real [h]
Análisis y Depuración	15	13
Investigación	30	47
Implementación teórica	30	39
Implementación computacional	50	43
Análisis de Resultados	25	17
Preparación entregables	15	15
TOTAL	165	174

Reuniones con asesores :

- Preparación y tiempo de reuniones con Profesor Pablo Aguirre : 1[h]
- Preparación y tiempo de reuniones con Profesor Ronny Vallejos : 8[h]

Total 9[h] versus 15[h] presupuestadas
Tiempo de preparación y comunicación con el mandante:

- Preparación y tiempo de reuniones con Profesor Jason Gerhard : 3[h]

Total 3[h] versus 15[h] presupuestadas

TIEMPO TOTAL : 186[h] utilizadas
(195[h] presupuestadas)

Bibliografía

- N. Cressie (1993) :Statistics for Spatial Data, Wiley, New York.
- K. Mumford, N. Mustafa, J. Gerhard: Probabilistic risk assesment of contaminant transfort in grownd water and vapour intrusion following remediation of a contaminant source, Elsevier, 2013.
- N. Mustafa, K. Mumford, J. Gerhard, D. O'Carrol: A three-dimensional numerical model for linking community-wide vapour risk, Elsevier, 2015.
- Dar Tjostheim: A Measure of assotiation for Spatial Variables, Biometrika, 1978.
- B. Marchant, R. Lark: The Mátern variogram model: Implications for uncertainty propagation and sampling in geostatistical surveys, Elsevier, 2007.
- B. Glick: A Spartial Rank-Order Correlation Measure, Wiley, 1982.
- R. Vallejos, R. Osorio, F. Cuevas: SpatialPack - An R package for computing spatial association between two stochastic processes defined on the plane, 2013.
- F. Cuevas, E. Porcu, R. Vallejos: Study of spatial relationships between two sets of variables: A nonparametric approach. Journal of Nonparametric Statistics, 2013.

Gracias por su atención!