Spatial estimation: Concentration of a contaminant in groundwater

Project [ MAT-288: Modeling Laboratory I ]

Professor: Pablo Aguirre
Tutors: Andrés Fuentes, Jason Gerhard

Gabriel Molina H.

Universidad Técnica Federico Santa María

Tuesday July 12, 2016
1. Introduction
   - Motivation
   - The Problem
   - How looks the data?

2. Modeling
   - The Model: Ordinary Kriging
   - Results
   - Error

3. Conclusions
Introduction
- Motivation
- The Problem
- How looks the data?

Modeling
- The Model: Ordinary Kriging
- Results
- Error

Conclusions
**Groundwater Contamination**

- Groundwater is the water present beneath Earth’s surface in soil pore spaces and in the fractures of rock formations.

- Actually, mainly in USA, exists filtration of contaminants in the soil. These contaminants can arriving at populated areas because of groundwater.

**Figure:** Example of groundwater contamination.
Groundwater is the water present beneath Earth’s surface in soil pore spaces and in the fractures of rock formations.

Actually, mainly in the USA, exists filtration of contaminants in the soil. These contaminants can arriving at populated areas because of groundwater.

Figure: Example of groundwater contamination.
The Problem

- $C_w$ is the logarithm of a concentration $\hat{C}_w [ML^{-3}]$ of a contaminant.

This project’s main objective is that from $N$ spatial samples of $C_w$ we can estimate the value of $C_w$ for an domain (3D or 2D).

- For this we have 625000 values of $C_w$ in a 3D domain, these data was obtain of a validated simulation.
The Problem

- $C_w$ is the logarithm of a concentration $\hat{C}_w [ML^{-3}]$ of a contaminant.
  This project’s main objective is that from $N$ spatial samples of $C_w$ we can estimate the value of $C_w$ for an domain (3D or 2D).
- For this we have 625000 values of $C_w$ in a 3D domain, these data was obtain of a validated simulation.
How looks the data: Domain

- We have 625000 values of $C_w$ associated with a 3D domain $[0, 1000] \times [0, 500] \times [0, 10]$ as the figure.

![Sketch domain](image)

**Figure:** Sketch domain.

- The $C_w$ data are located in the nodes of the grid with $\Delta x = 4$, $\Delta y = 2$ and $\Delta z = 1$.

So, we can separate the 3D-domain in 10 2D-layers with dimensions $[0, 1000] \times [0, 500]$. 

Gabriel E. Molina  
Spatial estimation: Concentration of a contaminant in groundwater
We have 625000 values of $C_w$ associated with a 3D domain $[0, 1000] \times [0, 500] \times [0, 10]$ as the figure.

The $C_w$ data are located in the nodes of the grid with $\Delta x = 4$, $\Delta y = 2$ and $\Delta z = 1$.

So, we can separate the 3D-domain in 10 2D-layers with dimensions $[0, 1000] \times [0, 500]$. 

Figure: Sketch domain.
How looks the data: Statistics Analysis of $C_w$

- **In the nature we don’t know this!**, but in our case (because $C_w$ comes of a simulation) this is useful for the comparison with the results of our prediction.

- For each layer (here’s the first one) $C_w$ we analyse the data.

- In ‘Figura 2.a’ we can see the quartiles increasing with the intensity of blue.

- In ‘Figura 2.b’ and ‘Figura 2.c’ we can see (in order) ‘$Y$ vs $C_w$’ and ‘$C_w$ vs $X$’.

- In ‘Figura 2.d’ we can see the histogram for $C_w$. 

---

**Gabriel E. Molina**  
**Spatial estimation: Concentration of a contaminant in groundwater**
How looks the data: Statistics Analysis of $C_w$

- **In the nature we don’t know this!,** but in our case (because $C_w$ comes of a simulation) this is useful for the comparison with the results of our prediction.

- For each layer (here’s the first one) $C_w$ we analyse the data.

- In ‘Figura 2.a’ we can see the quartiles increasing with the intensity of blue.
- In ‘Figura 2.b’ and ‘Figura 2.c’ we can see (in order) ‘$Y$ vs $C_w$’ and ‘$C_w$ vs $X$’.
- In ‘Figura 2.d’ we can see the histogram for $C_w$. 

---

Gabriel E. Molina  
Spatial estimation: Concentration of a contaminant in groundwater
How looks the data: Statistics Analysis of $C_w$

- **In the nature we don’t know this!**, but in our case (because $C_w$ comes of a simulation) this is useful for the a comparison with the results of our prediction.

- For each layer (here’s the first one) $C_w$ we analyse the data.

  - In 'Figura 2.a' we can see the quartiles increasing with the intensity of blue.
  - In 'Figura 2.b' and 'Figura 2.c' we can see (in order) 'Y vs $C_w$' and '$C_w$ vs $X$'.
  - In 'Figura 2.d' we can see the histogram for $C_w$
How looks the data: Statistics Analysis of $C_w$

- **In the nature we don’t know this!**, but in our case (because $C_w$ comes of a simulation) this is useful for the a comparison with the results of our prediction.
- For each layer (here’s the first one) $C_w$ we analyse the data.

In ’Figura 2.a’ we can see the quartiles increasing with the intensity of blue.
In ’Figura 2.b’ and ’Figura 2.c’ we can see (in order) ’$Y$ vs $C_w$’ and ’$C_w$ vs $X$’.
In ’Figura 2.d’ we can see the histogram for $C_w$. 

**Gabriel E. Molina**
**Spatial estimation: Concentration of a contaminant in groundwater**
How looks the data: Statistics Analysis of $C_w$

- **In the nature we don’t know this!**, but in our case (because $C_w$ comes of a simulation) this is useful for the a comparison with the results of our prediction.
- For each layer (here’s the first one) $C_w$ we analyse the data.

In ’Figura 2.a’ we can see the quartiles increasing with the intensity of blue.
In ’Figura 2.b’ and ’Figura 2.c’ we can see (in order) ’$Y$ vs $C_w$’ and ’$C_w$ vs $X$’.
In ’Figura 2.d’ we can see the histogram for $C_w$
How looks the data: Statistics Analysis of $C_w$

- In the next figure we can see a contour plot of the values of $C_w$ in the first layer.

**Figure:** Contour Plot of Cw (upper layer)
Contents

1 Introduction
   - Motivation
   - The Problem
   - How looks the data?

2 Modeling
   - The Model: Ordinary Kriging
   - Results
   - Error

3 Conclusions
The Model: Ordinary Kriging

Idea

Let be $D = \{ s \in [0, 1000] \times [0, 500] \subset \mathbb{R}^2 \}$ the domain for a layer, defines $Z(s), s \in D$ as a random variable, $Z(s)$ represents the value of $C_w$ at the point $s$.

Suppose $Z(s) = \mu(s) + e(s)$ with $\mu$ constant but unknown and $e(s) \sim (0, \Sigma)$. We want to obtain $\hat{Z}(s_0)$ an estimator of $C_w$ in $s_0$ since we know $N$ values of $C_w$ in the locations $\{s_1, s_2, \ldots, s_N\}$ with values $\{Z(s_1), Z(s_2), \ldots, Z(s_N)\}$. 
The Model: Ordinary Kriging

Idea

- Let be $D = \{ s \in [0, 1000] \times [0, 500] \subset \mathbb{R}^2 \}$ the domain for a layer, defines $Z(s), \ s \in D$ as a random variable, $Z(s)$ represents the value of $C_w$ at the point $s$

- Suppose $Z(s) = \mu(s) + e(s)$ with $\mu$ constant but unknown and $e(s) \sim (0, \Sigma)$. We want to obtains $\hat{Z}(s_0)$ an estimator of $C_w$ in $s_0$ since we know $N$ values of $C_w$ in the locations $\{s_1, s_2, ..., s_N\}$ with values $\{Z(s_1), Z(s_2), ..., Z(s_N)\}$. 
The Model: Ordinary Kriging

Definition 1

We define the variogram \( \gamma(h) \) as:

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

It’s important to say that exists a few \textit{preset families} of variograms that we can associate to a random variable.

Definition 2

We define the \textit{empirical variogram} \( \hat{\gamma}(h) \) as:

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

\[
N(h) := \{(s_i, s_j) : \|s_i - s_j\| = h, \quad s_i, s_j \in D\}
\]
Now when we have $\hat{\gamma}(h)$ for all the distances $h$ of the domain $D$ we can adjust (minimizing MeanSquaredError) some preset function $\gamma(h)$, and now we have the theoretical variogram.
The Model: Ordinary Kriging

So, in Ordinary Kriging we obtain the best estimator with the form

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

where we demand \( \hat{Z}(s_0) \) be an Minimum-variance unbiased estimator.(ie: \( \mathbb{E}[\hat{Z}(s_0) - Z(s_0)] = 0 \) and minimize

\[ \sigma^2_E = \mathbb{V}[\hat{Z}(s_0) - Z(s_0)] \]

So it reduces to solving the system

\[
\begin{bmatrix}
\gamma(s_1 - s_1) & \cdots & \gamma(s_1 - s_N) \\
\vdots & \ddots & \vdots \\
1 & \cdots & 1
\end{bmatrix}
\begin{bmatrix}
\lambda_{ok} \\
\vdots \\
\mu_{ok}
\end{bmatrix}
= 
\begin{bmatrix}
\gamma(s_1 - s_0) \\
\vdots \\
\gamma(s_N - s_0)
\end{bmatrix}
\]
Results

The next figure consists in contours plots of the results of Ordinary Kriging to some sizes of samples (the samples are coloured in orange).

Figure: Prediction of Cw for some sizes of sample N

Figure: Real Values of Cw
Particularly for $N = 50$ we analyse the prediction.

Figure: Statistics analysis of the prediction for N=50

In ’Figura 7.a’ we can see the quartiles increasing with the intensity of blue. In ’Figura 7.b’ and ’Figura 7.c’ we can see (in order) ’$Y$ vs $C_w$’ and ’$C_w$ vs $X$’.

In ’Figura 7.d’ we can see the histogram for $C_w$

Figure: Statistics analysis Cw (Real Data)
Particularly for $N = 50$ we observe the next figure that represents the variance of prediction in a contour plot. Again, the orange dots represent the sample.

**Figure:** Variance of prediction for $N=50$

We note the variance is higher away from samples.
Particularly for $N = 50$ we observe the next figure that represents a 3D plot of the prediction $P$ (red), the real values $C_w$ (blue) and the residue $R = P - C_w$ (green).

![3D comparison plot](image)

**Figure:** 3D comparison plot

OBS: If the prediction is perfect, $R = P - C_w$ correspond to the plane $z = 0$. $R$ is positive where prediction is overestimated, and is negative where the prediction is underestimated.
A measure of the goodness of prediction is the MSE. So in the next figure we can see 'MSE vs N'

Figure: Mean Squared Error vs N
Tabla de Contenidos

1 Introduction
   • Motivation
   • The Problem
   • How looks the data?

2 Modeling
   • The Model: Ordinary Kriging
   • Results
   • Error

3 Conclusions
A prediction for $C_w$ was achieved with good goodness of fit. In particular for $N=50$ obtains $\text{MSE}=0.74$.

‘$\text{MSE vs N}$’ provides a comparison between the error of prediction and the number of samples $N$.

This method is ‘cheaper’ and faster than trying to solve an diffusion equation, and do not handle assumptions regarding soil or nature parameters.

It is possible in the future to develop a time serie to analyze the evolution over time of $C_w$. 
Thanks for your attention!