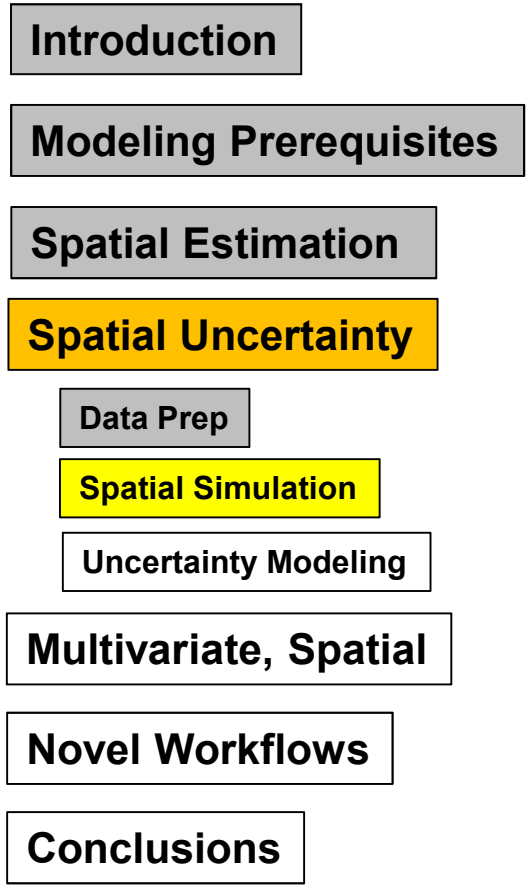
Data Analytics and Geostatistics:
Data Preparation

Lecture outline . . .

- **Gaussian Simulation**
- **Indicator Simulation**

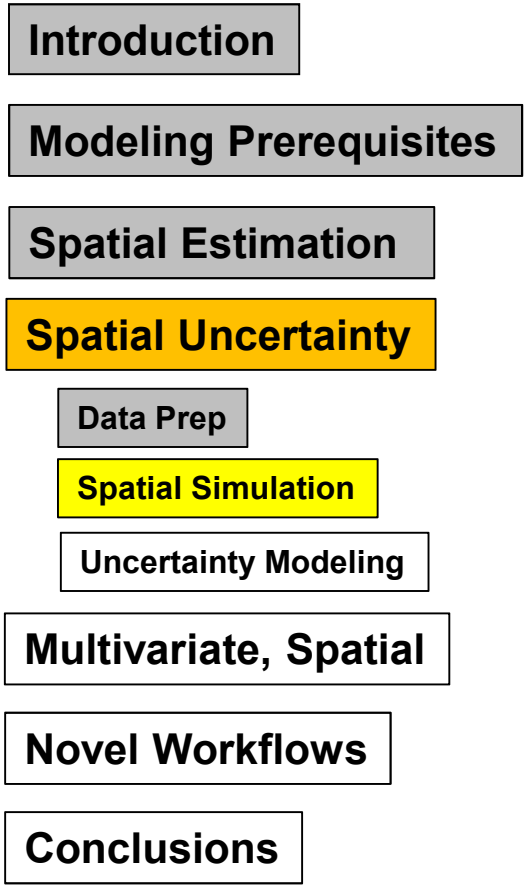


Instructor: Michael Pyrcz, the University of Texas at Austin

Data Analytics and Geostatistics:
Data Preparation

Lecture outline . . .

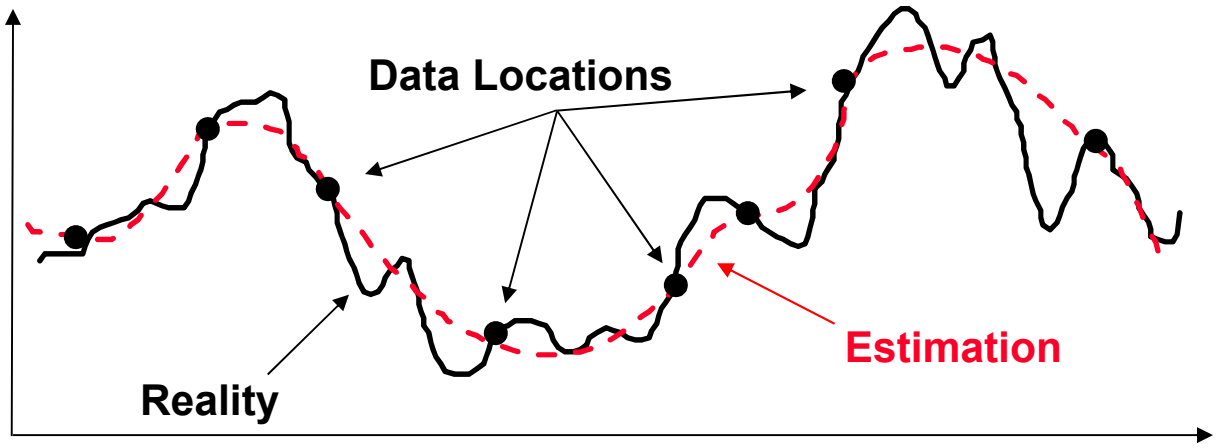
- **Gaussian Simulation**



Instructor: Michael Pyrcz, the University of Texas at Austin

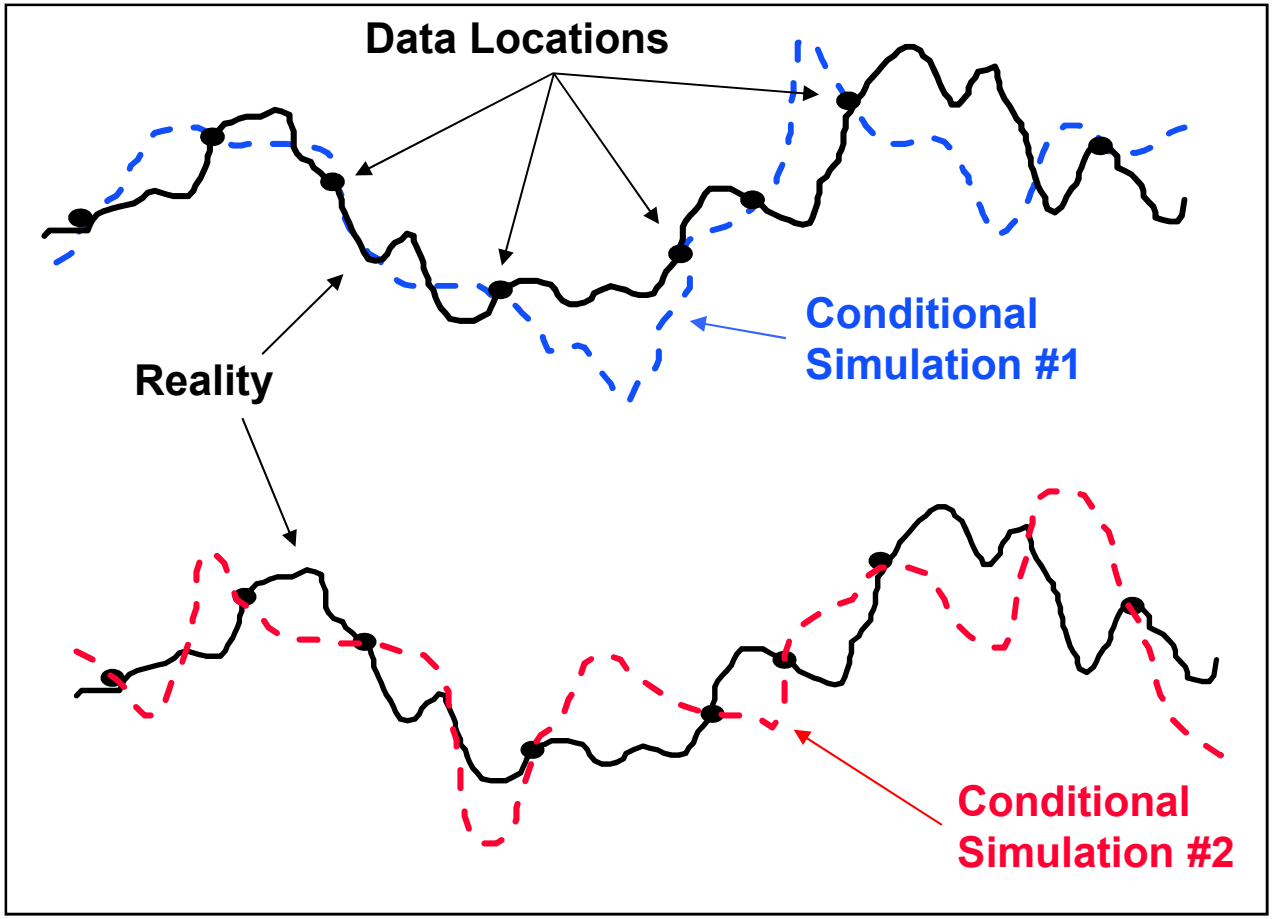
Motivation for Simulation

- Recall estimation: assign the most accurate value at each location.



Motivation for Simulation

- What do we accomplish with simulation?



Motivation for Simulation

- What do we accomplish with simulation?

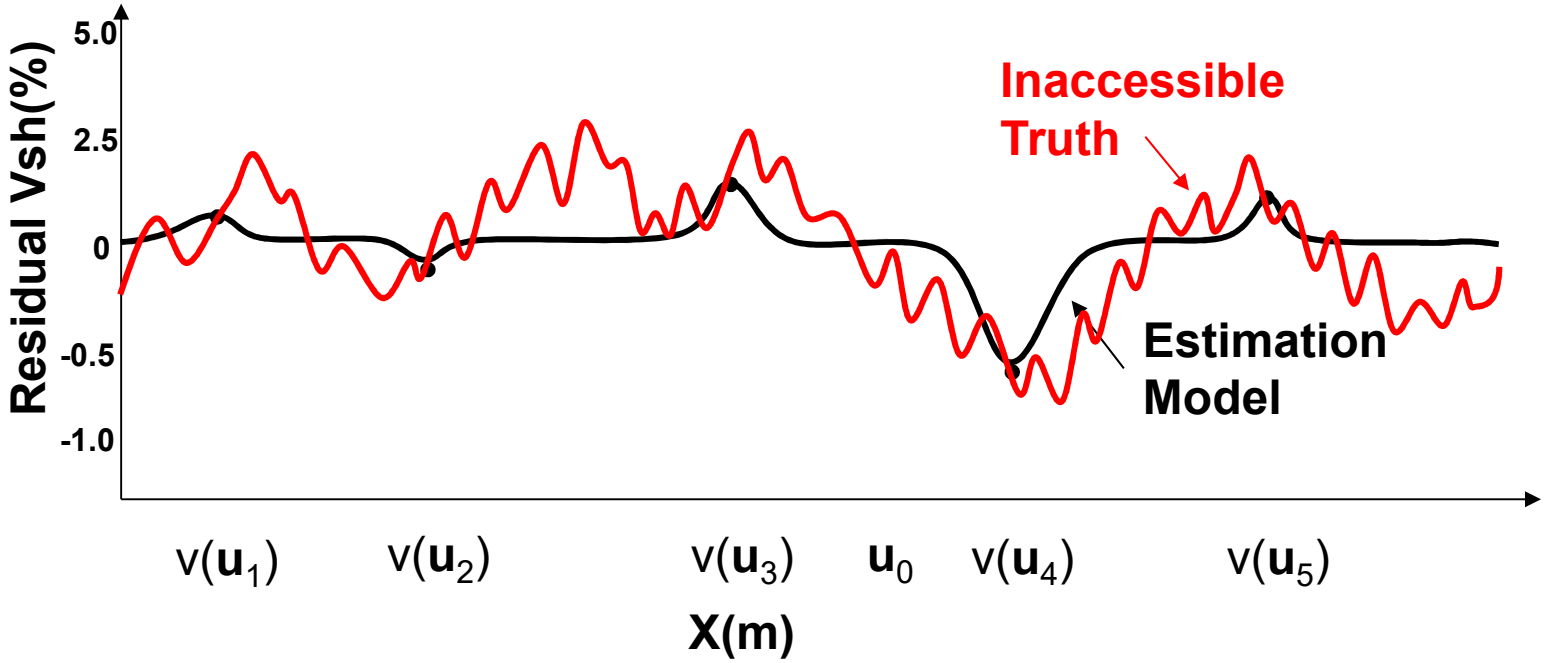
What does simulated dill pickle potato chips taste like?

What does a simulated reservoir model look like?

What does a simulated reservoir flow like?

What's Wrong with Kriging?

- Kriging is an estimation method. The goal of estimation is the most likely value at each location. Too smooth!
  - Would kriging honor the global distribution? Variance is too low!
  - Would kriging honor the variogram model? Super continuous!



  - We need a method that corrects for the smoothing ⇨ simulation

Estimation vs. Simulation

Estimation:

- honors local data (with discontinuity)
- locally accurate
- smooth appropriate for visualizing trends
- inappropriate for flow simulation
- no assessment of global uncertainty

Simulation:

- honors local data
- reproduces histogram
- honors spatial variability → appropriate for flow simulation
- alternative realizations possible → change random number seed
- assessment of global uncertainty is possible

Estimation and Simulation Definition

Estimation:

- a method to calculate the best estimate at each location
- focus on local accuracy, globally too smooth

Simulation:

- a method to calculate a good / reasonable estimate at each location
- focus on global accuracy, sacrifice local accuracy

Smoothing Effect of Kriging

- Kriging is locally accurate and smooth, appropriate for visualizing trends, inappropriate for any applications where heterogeneity is important

- The "variance" of the kriged estimates is too small:

$$Var\{Y^*(\mathbf{u})\} = \sigma^2 - \sigma^2_{SK}(\mathbf{u})$$

Simple Kriging Estimation Variance

  - σ^2 is variance of the property, $Var\{Y^*(\mathbf{u})\}$ is the variance of the estimates

Consider the following:

  - $\sigma^2_{SK}(\mathbf{u})$ is zero at the data locations → no smoothing
  - $\sigma^2_{SK}(\mathbf{u})$ is variance σ^2 beyond range of data locations → complete smoothing
  - spatial variations of $\sigma^2_{SK}(\mathbf{u})$ depend on the variogram and data spacing

Proposal to 'Correct Kriging'

- Missing variance in the estimates, $Var\{Y^*(\mathbf{u})\}$, *is* the kriging variance, $\sigma^2_{SK}(\mathbf{u})$
- Monte Carlo *Simulation* corrects the variance (get right histogram)

Add random residual $Y_s(\mathbf{u}) = Y^*(\mathbf{u}) + R(\mathbf{u})$ where *R*(u) is random residual that adds back in the missing variance.

and correct the covariance (get the right variogram)

Sequential simulation – add simulated values to data

- Simulation reproduces histogram, honors spatial variability (variogram)
  → appropriate for process evaluation where heterogeneity is important
- Allows an assessment of uncertainty with alternative realizations
- We now prove that this will work:

Covariance Reproduction and Kriging

- Recall the simple kriging estimator: $Y^*(\mathbf{u}) = \sum_{\beta=1}^{n} \lambda_\beta \cdot Y(\mathbf{u}_\beta)$

  and the corresponding simple kriging system: $\sum_{\beta=1}^{n} \lambda_\beta C(\mathbf{u}_\alpha, \mathbf{u}_\beta) = C(\mathbf{u}, \mathbf{u}_\alpha), \ \forall \mathbf{u}_\alpha$

- Let's calculate the covariance between the kriged estimate and one of the data values:

$$
\begin{aligned}
Cov\left\{Y^*(u), Y(u_\alpha)\right\} &= E\left\{Y^*(u), Y(u_\alpha)\right\} \\
&= E\left\{\left[\sum_{\beta=1}^{n} \lambda_\beta \cdot Y(u_\beta)\right] \cdot Y(u_\alpha)\right\} \\
&= \sum_{\beta=1}^{n} \lambda_\beta \cdot E\left\{Y(u_\beta) \cdot Y(u_\alpha)\right\} \\
&= \sum_{\beta=1}^{n} \lambda_\beta C(u_\alpha, u_\beta) \\
&= C(u, u_\alpha)
\end{aligned}
$$

Assuming mean of zero

The covariance between the data and the estimates is correct!

Proof is from Pyrcz and Deutsch, 2014, p. 122

Covariance Reproduction and Kriging

- The kriging equations forces the covariance between the **data values** and the **kriging estimate** to be correct

- Let's check the three parts of covariance in kriged models:

  1. between data values → **correct**
  2. between data values and predicted values → **correct**
  3. between predicted values → **incorrect**

- Recall: that variance of kriged estimates is too small

Addition of Missing Variance

- The variance of our random function: $\sigma^2 = C(0)$

- Stationary variance should be constant everywhere:

$$\sigma^2(\mathbf{u}) = \sigma^2, \ \forall \ \mathbf{u} \in A$$

- Although the covariance between the kriged estimates and the data is correct, the variance is too small:

$$Var\{Y^*(\mathbf{u})\} = C(0) - \sigma^2_{SK}(\mathbf{u})$$

the *missing variance* is the kriging variance $\sigma^2_{SK}(\mathbf{u})$!!

- We need to add back in the missing variance without changing the covariance reproduction

Addition of Missing Variance

- Add an independent component with zero mean and the correct variance:

$$Y_s(\mathbf{u}) = Y^*(\mathbf{u}) + R(\mathbf{u})$$

- Covariance is unchanged:

$$Cov\{Y_s(u), Y(u_\alpha)\} = E\{Y_s(u) \bullet Y(u_\alpha)\}$$

$$= E\left\{\left[\sum_{\beta=1}^{n} \lambda_\beta \cdot Y(u_\beta) + R(u)\right] \cdot Y(u_\alpha)\right\}$$

$$= \sum_{\beta=1}^{n} \lambda_\beta \cdot E\{Y(u_\beta) \cdot Y(u_\alpha)\} + E\{R(u) \cdot Y(u_\alpha)\}$$

- note that $E\{R(\mathbf{u}) \cdot Y(\mathbf{u}_\alpha)\} = E\{R(\mathbf{u})\} \cdot E\{Y(\mathbf{u}_\alpha)\}$, since *R*(u) is random

$$E\{R(\mathbf{u})\} = 0.0 \Rightarrow E\{R(\mathbf{u}) \cdot Y(\mathbf{u}_\alpha)\} = E\{R(\mathbf{u})\} \cdot E\{Y(\mathbf{u}_\alpha)\} = 0$$

- Therefore, $Cov\{Y_s(\mathbf{u}), Y(\mathbf{u}_\alpha)\} = Cov\{Y^*(\mathbf{u}), Y(\mathbf{u}_\alpha)\} = C(\mathbf{u}), Y(\mathbf{u}_\alpha)\}$

given the simple kriging system, $\sum_{\beta=1}^{n} \lambda_\beta C(\mathbf{u}_\alpha, \mathbf{u}_\beta) = C(\mathbf{u}, \mathbf{u}_\alpha), \ \forall \mathbf{u}_\alpha$

Proof is from Pyrcz and Deutsch, 2014, p. 123

Sequential Gaussian Simulation Definition

- **Sequential** – sequential inclusion of simulated values to impose the correct spatial correlation between the simulated values.

- **Gaussian** – work in Gaussian space since the local conditional distribution shape is known and can be parameterized by mean (kriging estimate) and variance (estimation variance).

- **Simulation** – simulation through Monte Carlo simulation from the local distribution of uncertainty to add in the missing variance and construction of multiple, equiprobable realizations.

Sequential Simulation Workflow

- Transform data to standard normal distribution (all work will be done in "normal" space)

- Go to a location and perform kriging to get mean and corresponding kriging variance:

$$Y^*(\mathbf{u}) = \sum_{\beta=1}^{n} \lambda_\beta \cdot Y(\mathbf{u}_\beta)$$

$$\sigma^2_{SK}(\mathbf{u}) = C(0) - \sum_{\alpha=1}^{n} \lambda_\alpha C(\mathbf{u}, \mathbf{u}_\alpha)$$

- Draw a random residual *R*(u) that follows a normal distribution with mean of 0.0 and variance of $\sigma^2_{SK}(\mathbf{u})$

- Add the kriged estimate and residual to get simulated value:

$$Y_s(\mathbf{u}) = Y^*(\mathbf{u}) + R(\mathbf{u})$$

- Note that $Y_s(\mathbf{u})$ could be equivalently obtained by drawing from a normal distribution with mean $Y^*(\mathbf{u})$ and variance $\sigma^2_{SK}(\mathbf{u})$

Sequential Simulation Workflow

- Add $Y_s(\mathbf{u})$ to the set of data to ensure that the covariance with this value and all future predictions is correct
- A key idea of sequential simulation is to use previously kriged/simulated values as data so that we reproduce the covariance between *all* of the simulated values!
- Visit all locations in random order (to avoid artifacts of limited search)
- Back-transform all data values and simulated values when model is populated
- Create another equiprobable realization by repeating with different random number seed

Why Gaussian Simulation?

- Local estimate is given by kriging
- Mean of residual is zero and variance is given by kriging; however, what "shape" of distribution should we consider?
- Advantage of normal / Gaussian distribution is that the global N(0,1) distribution will be preserved if we always use Gaussian distributions
- Transform data to normal scores in the beginning (before variography)
- Simulate 3-D realization in "normal space"

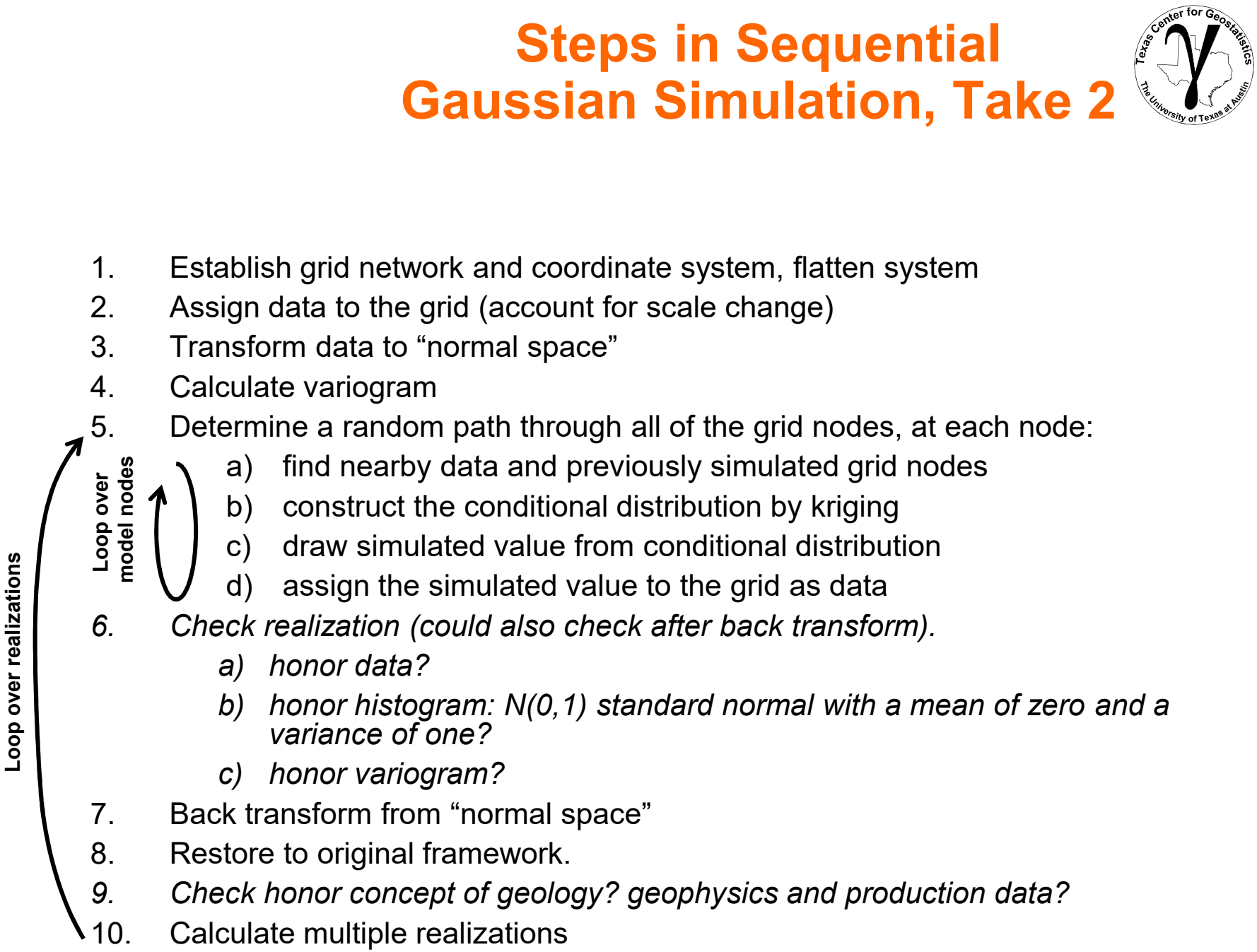
Why Gaussian Simulation?

- Back-transform all of the values when finished
- Price of mathematical simplicity is the characteristic of maximum spatial entropy, i.e., low and high values are disconnected. Not appropriate for permeability.
- Of all unbounded distributions of finite variance, the Gaussian distribution maximizes entropy:

$$H = -\int_{-\infty}^{+\infty} \left[ln f(z) \right] f(z) dz$$

- Consequences:
  - maximum spatial disorder beyond the variogram
  - maximum disconnectedness of extreme values
  - median values have greatest connectedness
  - symmetric disconnectedness of extreme low / high values

Steps in Sequential Gaussian Simulation, Take 2



1. Establish grid network and coordinate system, flatten system
2. Assign data to the grid (account for scale change)
3. Transform data to “normal space”
4. Calculate variogram
5. Determine a random path through all of the grid nodes, at each node:
   a) find nearby data and previously simulated grid nodes
   b) construct the conditional distribution by kriging
   c) draw simulated value from conditional distribution
   d) assign the simulated value to the grid as data
6. *Check realization (could also check after back transform).*
   a) honor data?
   b) honor histogram: N(0,1) standard normal with a mean of zero and a variance of one?
   c) honor variogram?
7. Back transform from “normal space”
8. Restore to original framework.
9. *Check honor concept of geology? geophysics and production data?*
10. Calculate multiple realizations

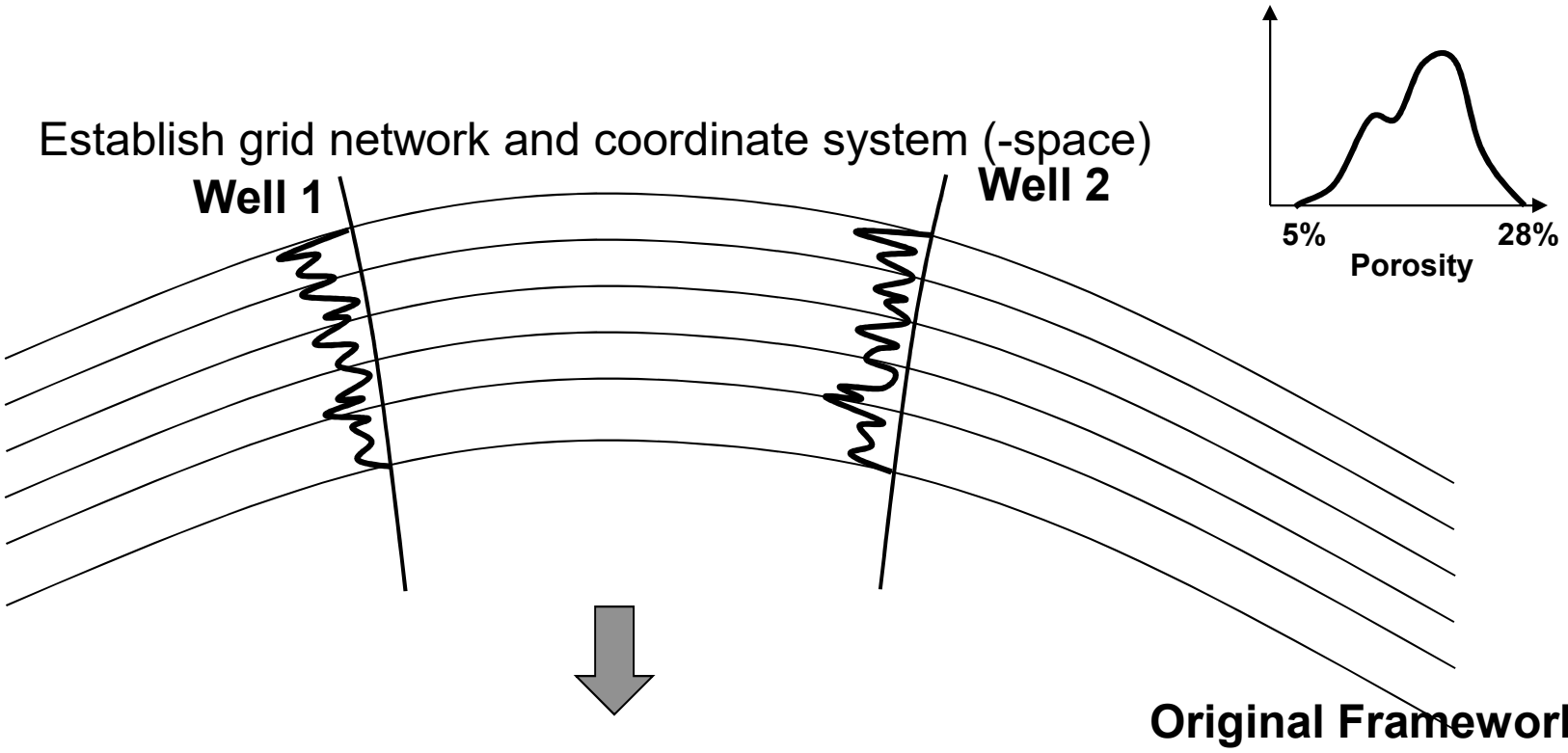
Steps in Sequential Gaussian Simulation, Take 2

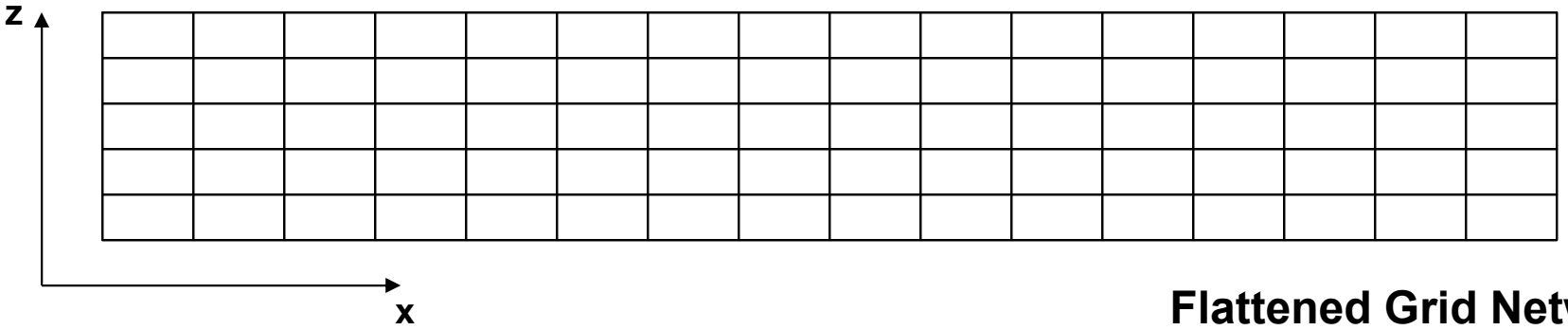
Why you should know the sequential Gaussian (SGS) simulation workflow?

1. >90% of reservoir models use SGS in the workflow!
2. The workflow links directly to theory.
   a) Correcting for missing variance with addition of a random residual / kriging variance
   b) Ensuring correct covariance between simulated values with sequential approach
   c) Use of a random path and Monte Carlo simulation to calculate realizations.
3. You'll understand when things go wrong!
   a) Limited search artifacts
   b) Mismatch with input statistics / ergodic fluctuations
   c) String effect
4. The sequential framework is used in many other methods
   a) cosimulations, indicator simulation, truncated Gaussian, MPS etc.

Steps in Sequential Gaussian Simulation, Take 2

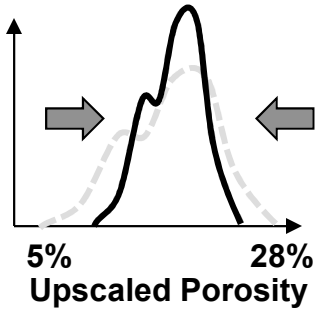
1. Establish grid network and coordinate system (-space)

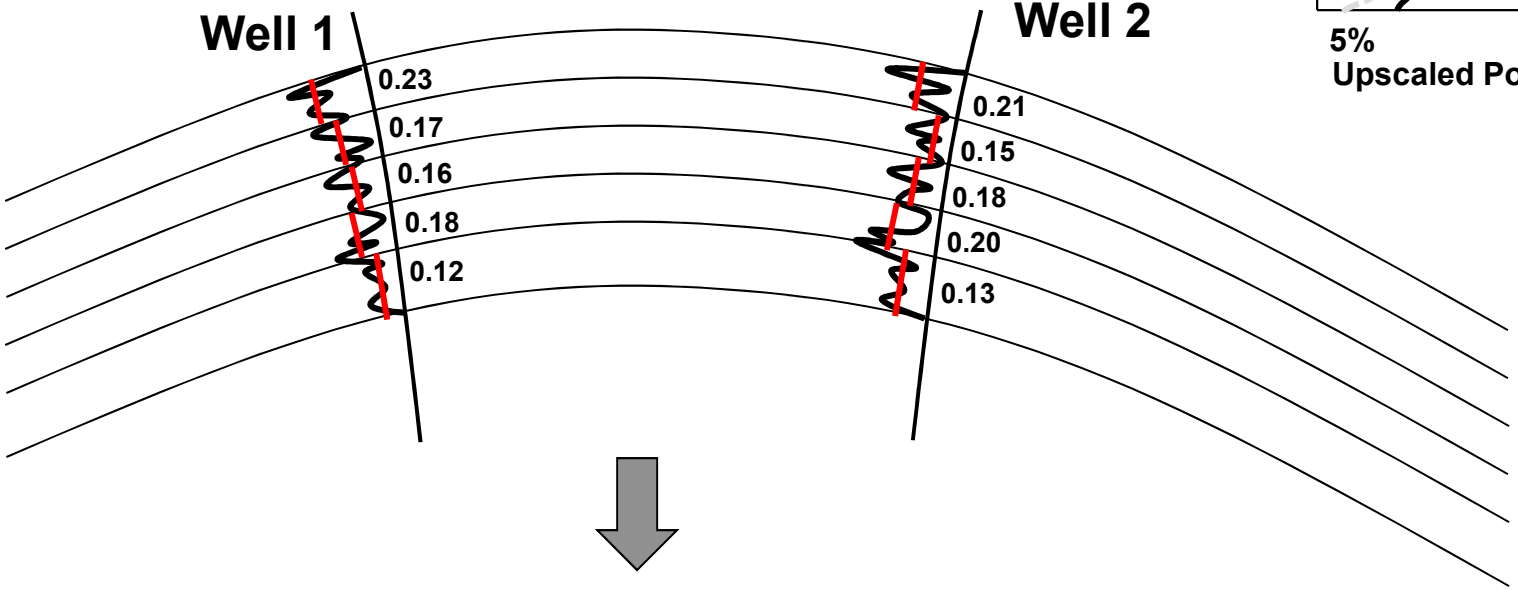


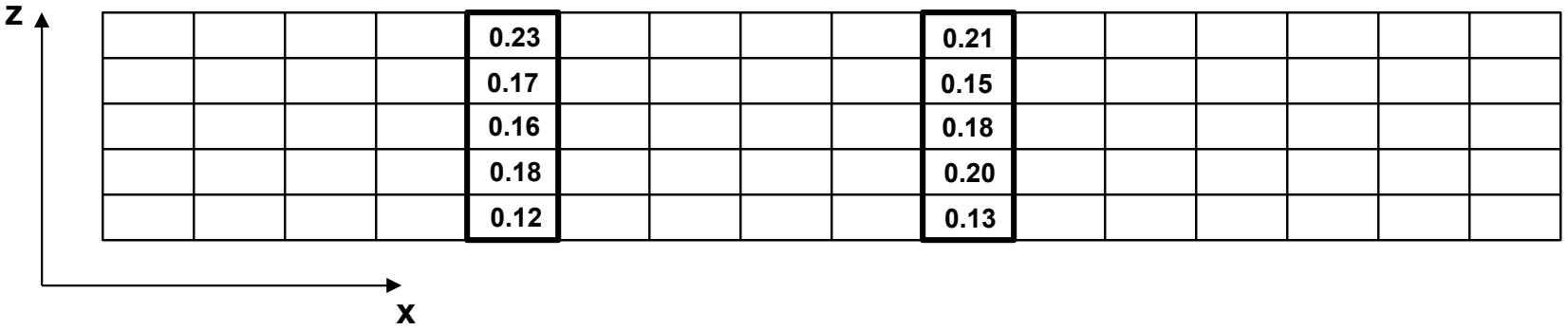


Steps in Sequential Gaussian Simulation, Take 2

2. Assign data to the grid (account for scale change)

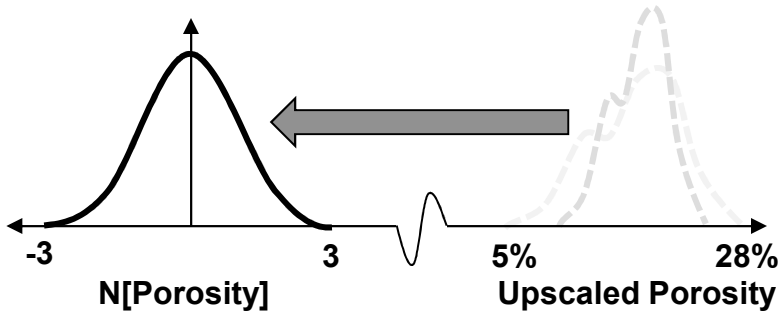


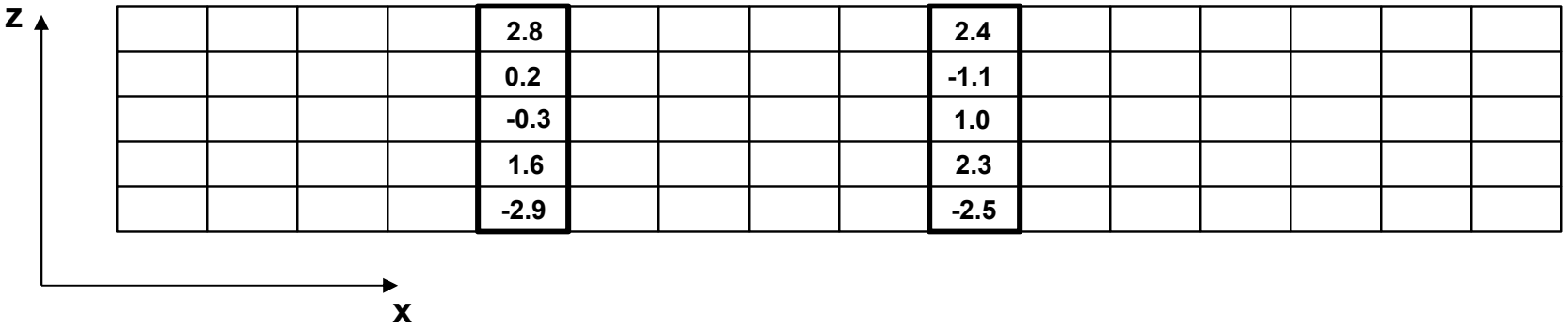




Steps in Sequential Gaussian Simulation, Take 2

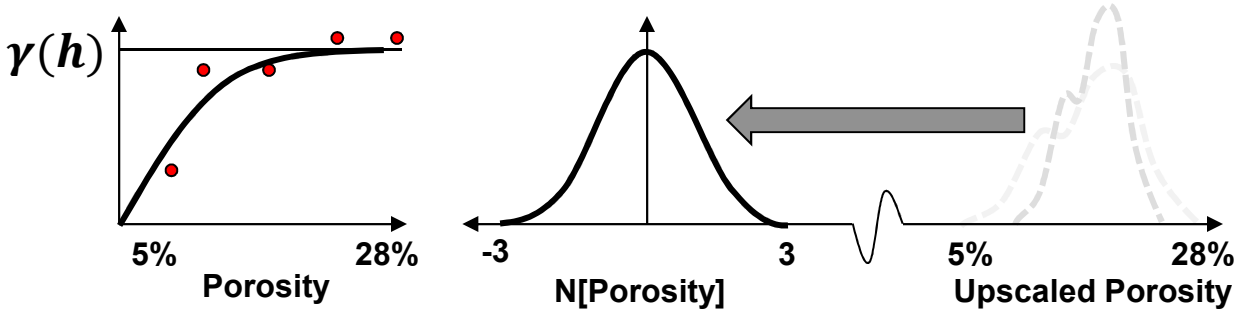
3. Transform data to “normal space”

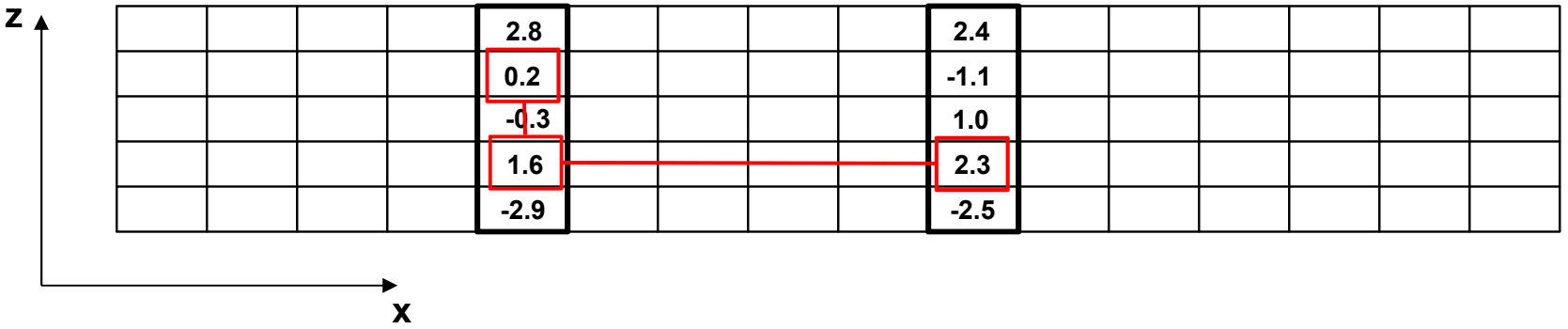




Steps in Sequential Gaussian Simulation, Take 2

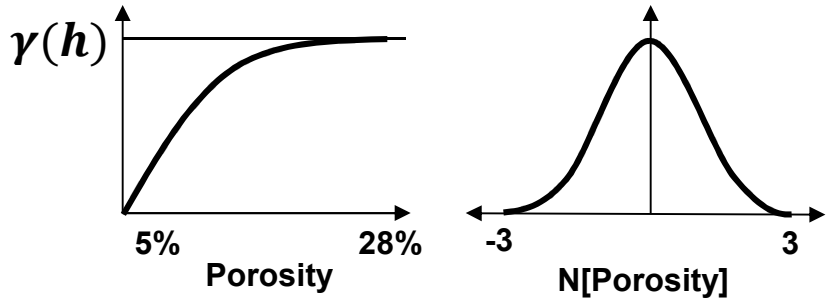
4. Calculate and model variogram

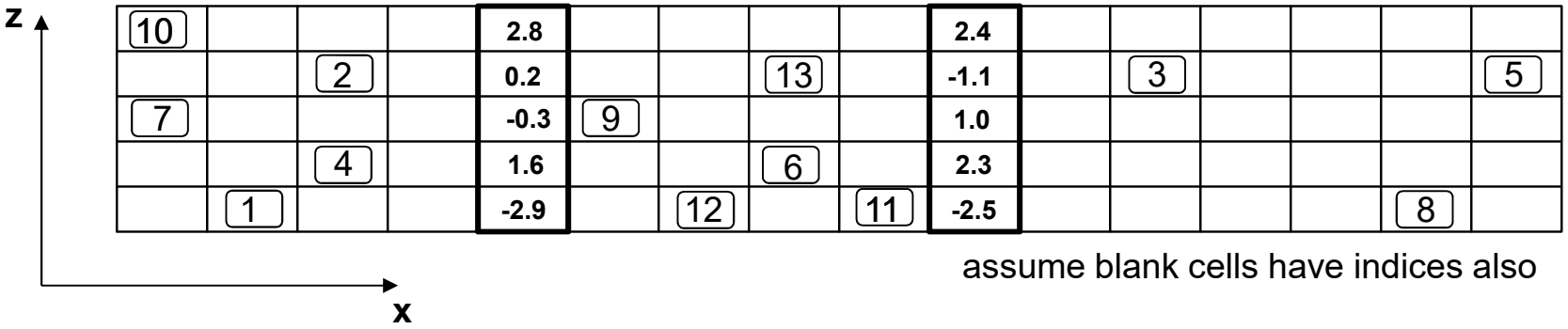




Steps in Sequential Gaussian Simulation, Take 2

5. Determine a random path through all of the grid nodes
   - only included 1st 13 nodes

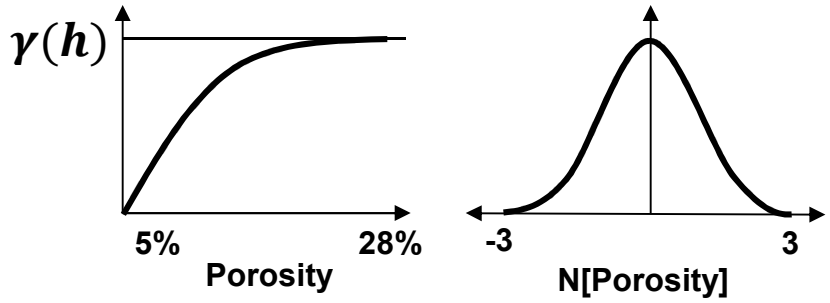


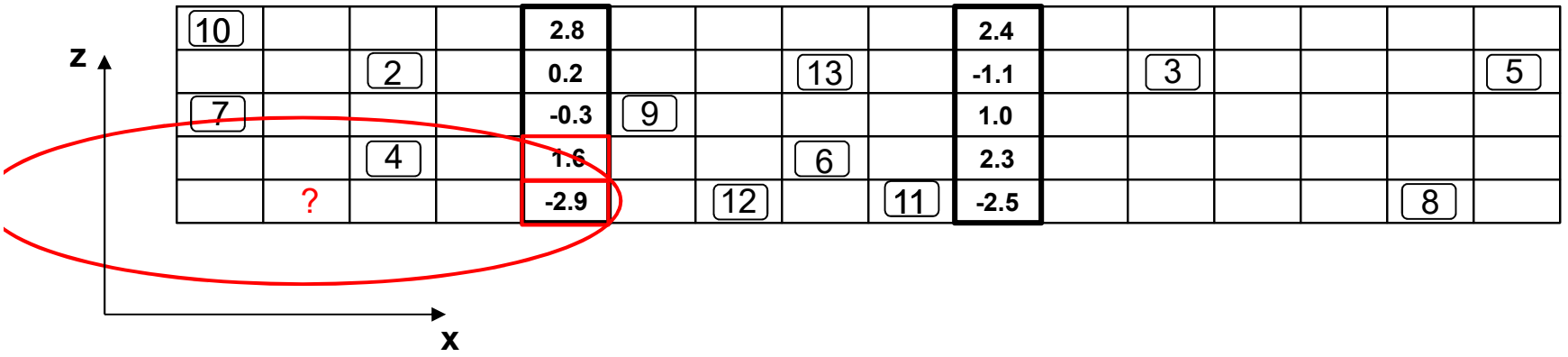


assume blank cells have indices also

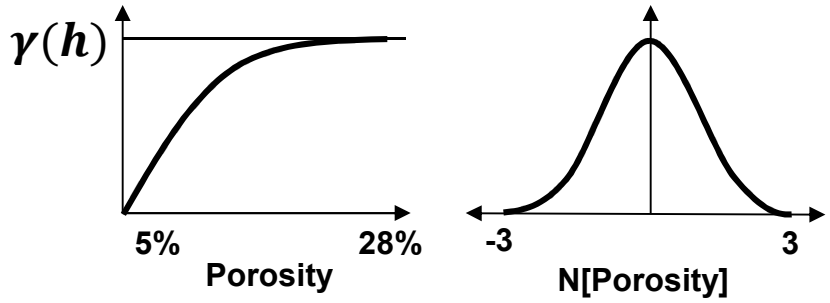
Steps in Sequential Gaussian Simulation, Take 2

5. For each node on random path:
   a) find nearby data and previously simulated grid nodes

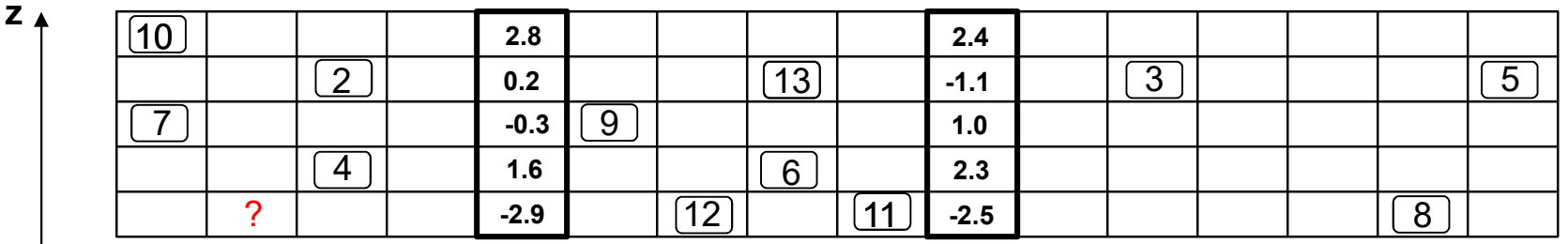




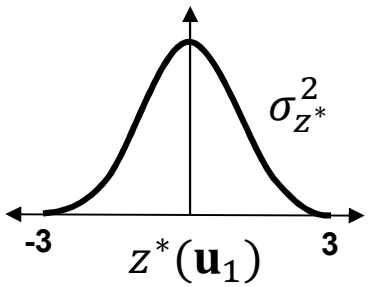
Steps in Sequential Gaussian Simulation, Take 2



5. For each node on random path:
   a) find nearby data and previously simulated grid nodes
   b) construct the conditional distribution by kriging

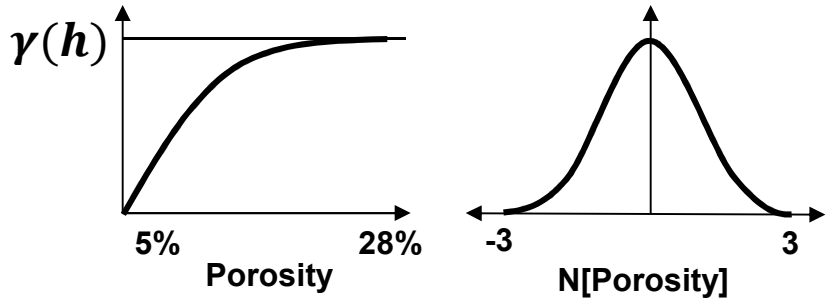


$$\begin{bmatrix} C_{1,1} & & \\ & & \\ & & C_{n,n} \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \\ \lambda_n \end{bmatrix} = \begin{bmatrix} C_{0,1} \\ \\ C_{0,n} \end{bmatrix}$$

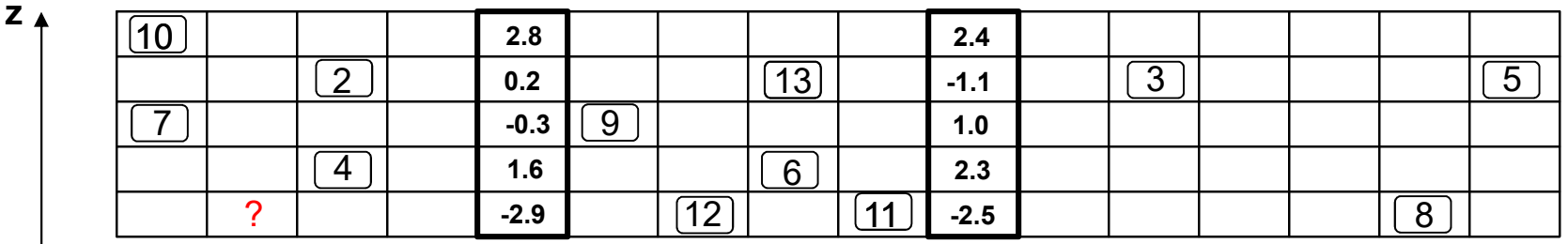


Simple Kriging
Indicator Kriging
Multiple Point Template and Training Image

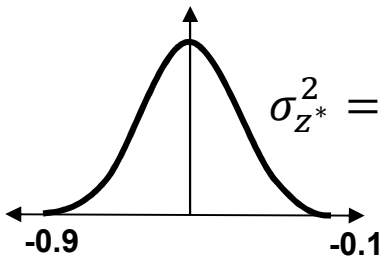
Steps in Sequential Gaussian Simulation, Take 2



5. For each node on random path:
   a) find nearby data and previously simulated grid nodes
   b) construct the conditional distribution by kriging
   c) draw simulated value from conditional distribution

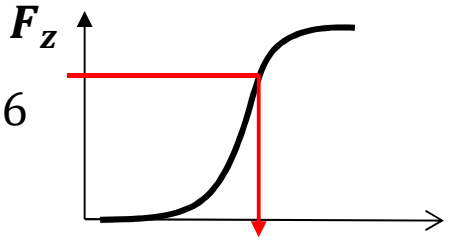


$$\begin{bmatrix} C_{1,1} & & \\ & & C_{n,n} \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_n \end{bmatrix} = \begin{bmatrix} C_{0,1} \\ C_{0,n} \end{bmatrix}$$



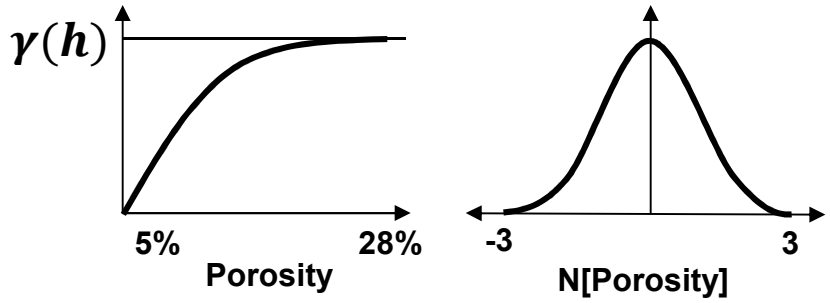
$\sigma^2_{z^*} = 0.16$

$z^*(\mathbf{u}_1) = -0.5$

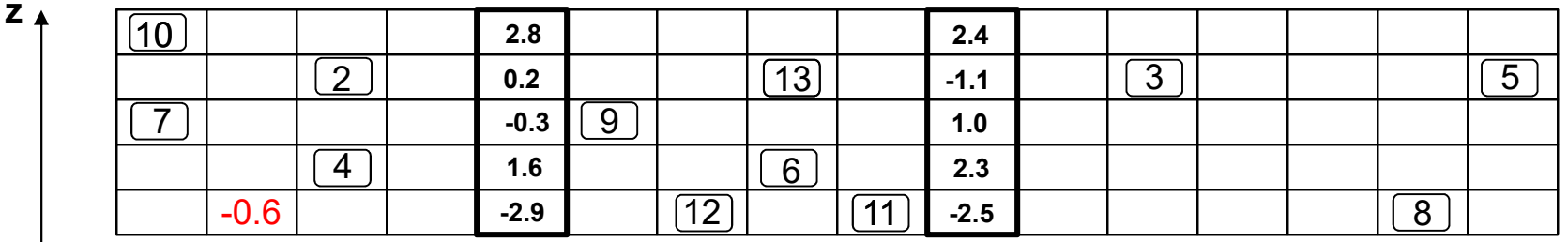


$z_s^{\ell=1}(\mathbf{u}_1) = -0.6$

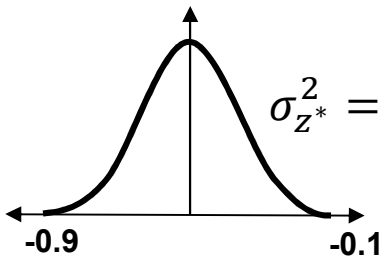
Steps in Sequential Gaussian Simulation, Take 2



5. For each node on random path:
   a) find nearby data and previously simulated grid nodes
   b) construct the conditional distribution by kriging
   c) draw simulated value from conditional distribution
   d) assign the simulated value to the grid as data

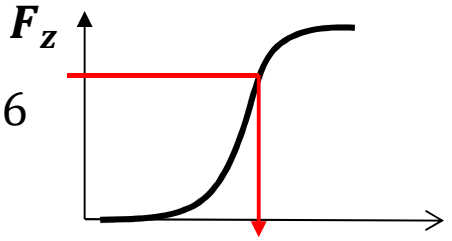


$$\begin{bmatrix} C_{1,1} & & \\ & & C_{n,n} \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_n \end{bmatrix} = \begin{bmatrix} C_{0,1} \\ C_{0,n} \end{bmatrix}$$



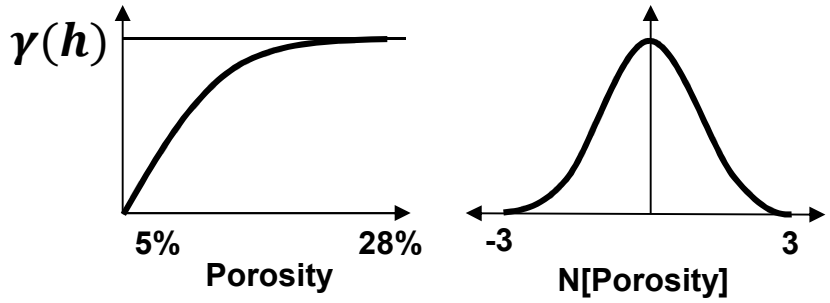
$\sigma^2_{z^*} = 0.16$

$z^*(\mathbf{u}_1) = -0.5$

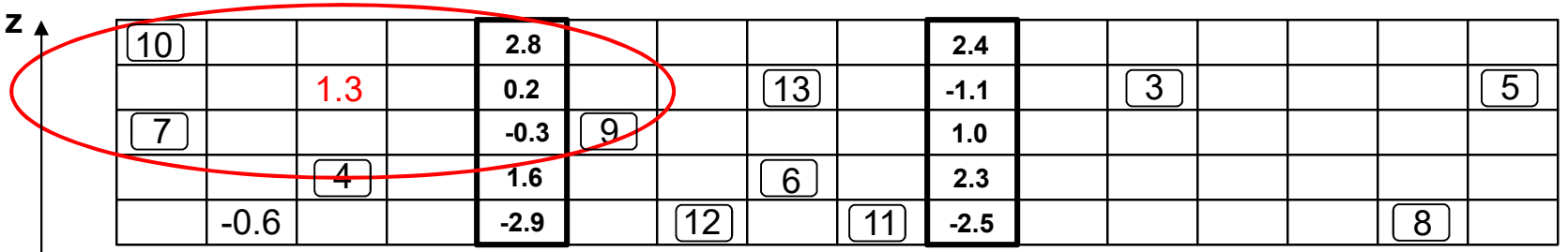


$z_s^{\ell=1}(\mathbf{u}_1) = -0.6$

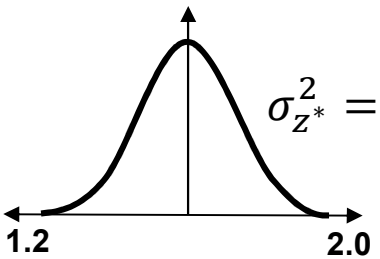
Steps in Sequential Gaussian Simulation, Take 2



5. For each node on random path:
   a) find nearby data and previously simulated grid nodes
   b) construct the conditional distribution by kriging
   c) draw simulated value from conditional distribution
   d) assign the simulated value to the grid as data

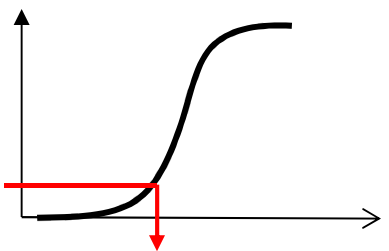


$$\begin{bmatrix} C_{1,1} & & \\ & & C_{n,n} \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \\ \lambda_n \end{bmatrix} = \begin{bmatrix} C_{0,1} \\ \\ C_{0,n} \end{bmatrix}$$



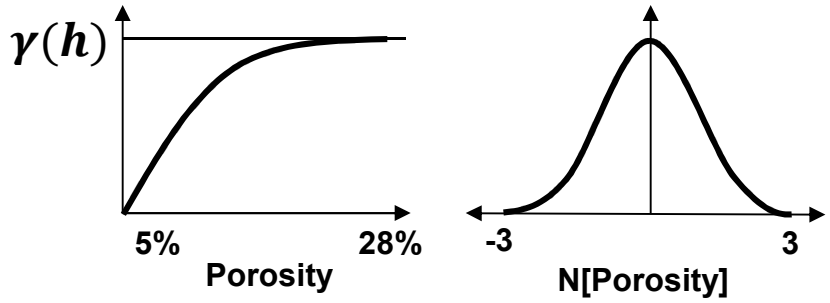
$\sigma^2_{z^*} = 0.11$

$z^*(\mathbf{u}_1) = 1.6$

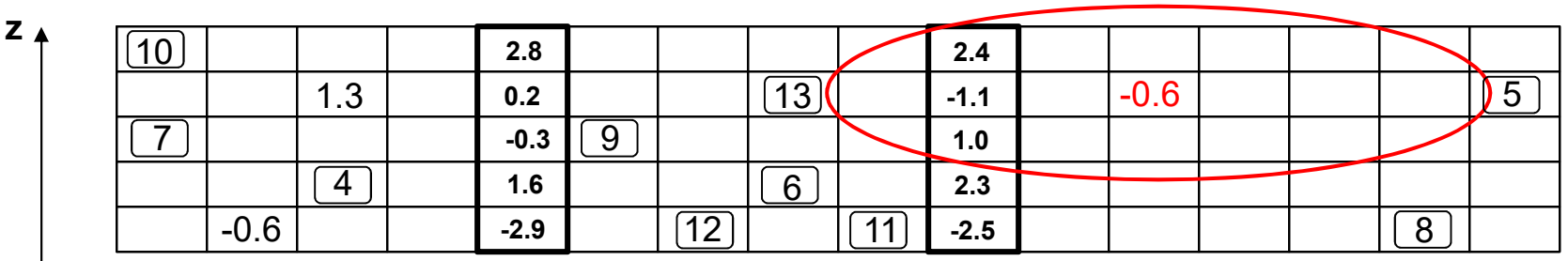


$z_s^{\ell=2}(\mathbf{u}_2) = 1.3$

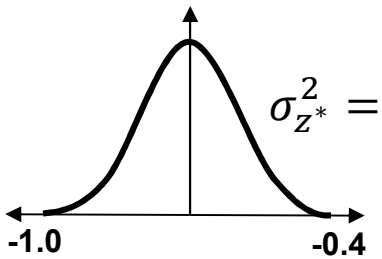
Steps in Sequential Gaussian Simulation, Take 2



5. For each node on random path:
   a) find nearby data and previously simulated grid nodes
   b) construct the conditional distribution by kriging
   c) draw simulated value from conditional distribution
   d) assign the simulated value to the grid as data

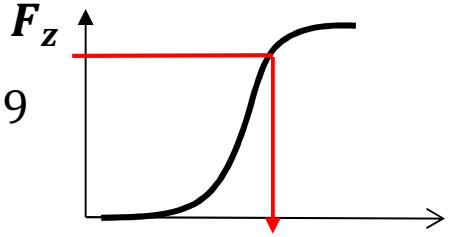


$$\begin{bmatrix} C_{1,1} & & \\ & & C_{n,n} \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \\ \lambda_n \end{bmatrix} = \begin{bmatrix} C_{0,1} \\ \\ C_{0,n} \end{bmatrix}$$



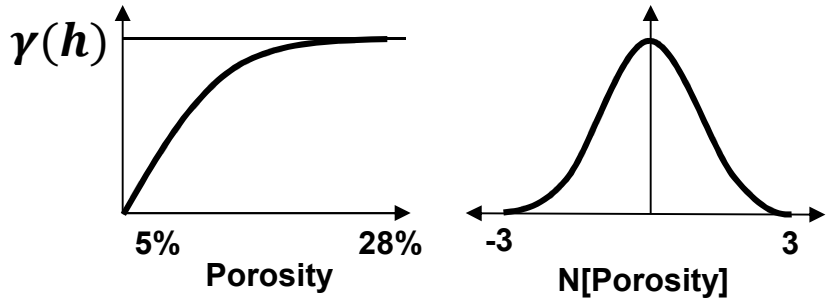
$\sigma^2_{z^*} = 0.09$

$z^*(\mathbf{u}_1) = -0.7$

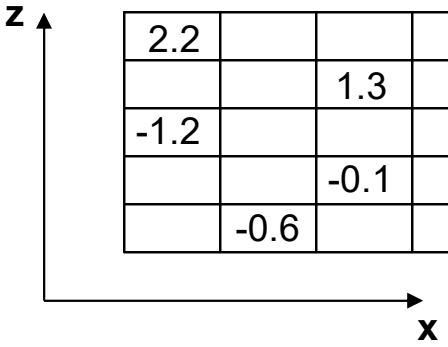


$z_s^{\ell=3}(\mathbf{u}_3) = -0.6$

Steps in Sequential Gaussian Simulation, Take 2



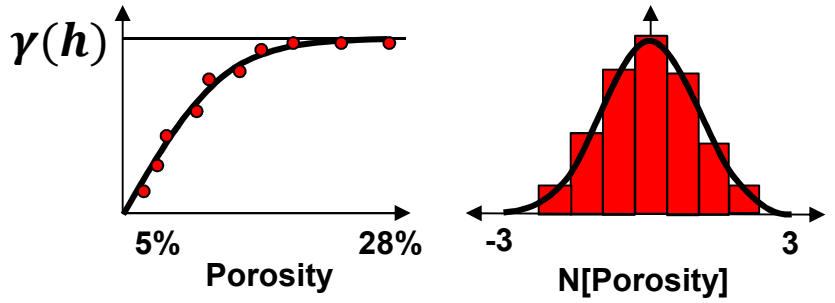
5. For each node on random path:
   a) find nearby data and previously simulated grid nodes
   b) construct the conditional distribution by kriging
   c) draw simulated value from conditional distribution
   d) assign the simulated value to the grid as data



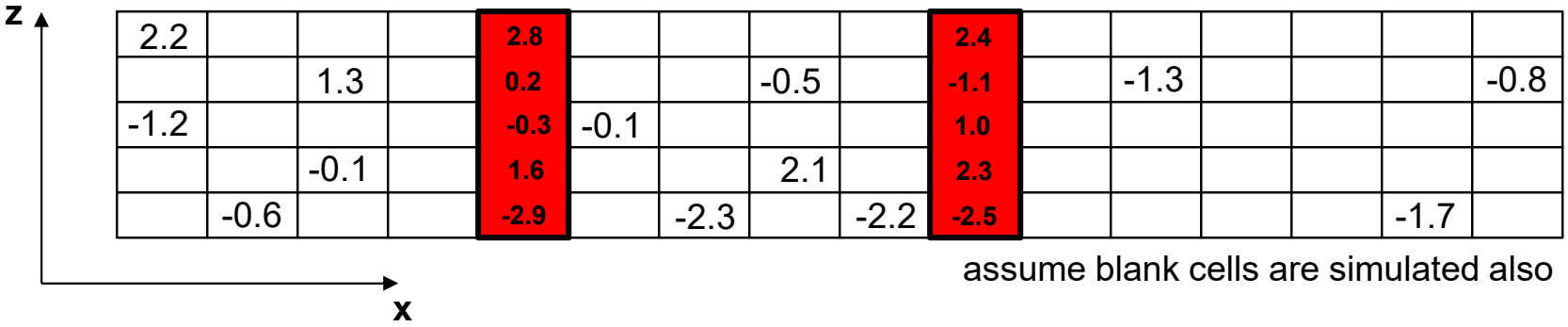
2.2				**2.8**					**2.4**						
		1.3		**0.2**			-0.5		**-1.1**		-1.3				-0.8
-1.2				**-0.3**	-0.1				**1.0**						
		-0.1		**1.6**			2.1		**2.3**						
	-0.6			**-2.9**		-2.3		-2.2	**-2.5**					-1.7	

assume blank cells are simulated also

Steps in Sequential Gaussian Simulation, Take 2



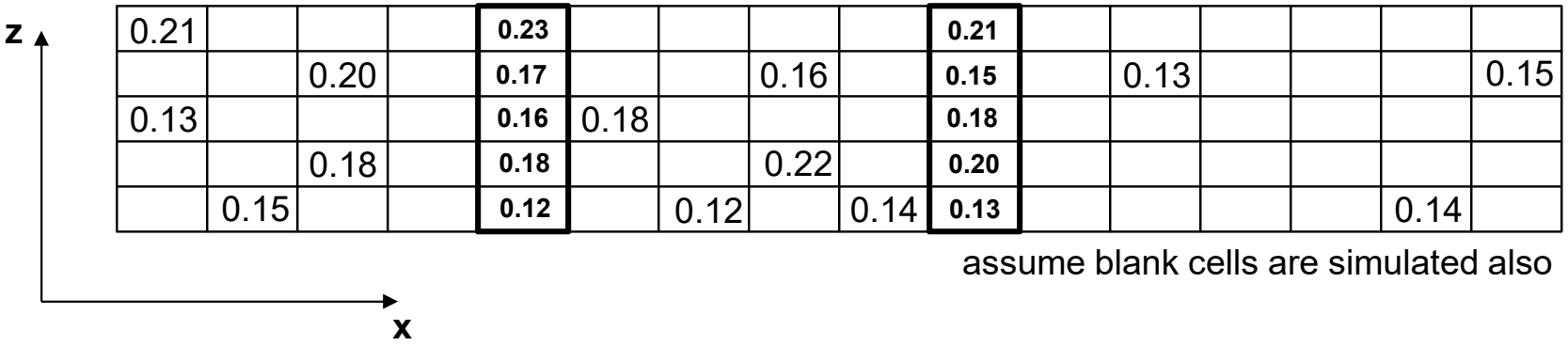
6. Check results
   a) honor data?
   b) honor histogram: N(0,1) standard normal with a mean of zero and a variance of one?
   c) honor variogram?



assume blank cells are simulated also

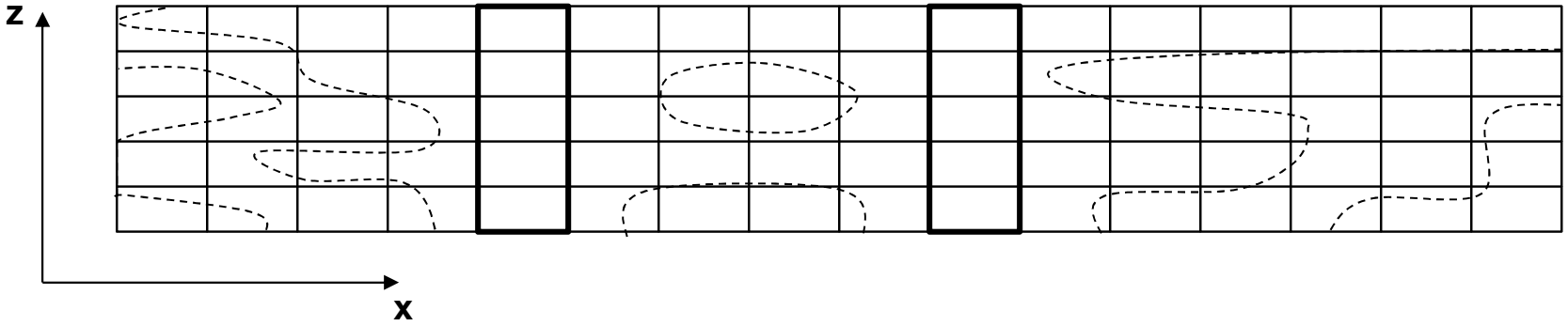
Steps in Sequential Gaussian Simulation, Take 2

7. Back transform from “normal space”



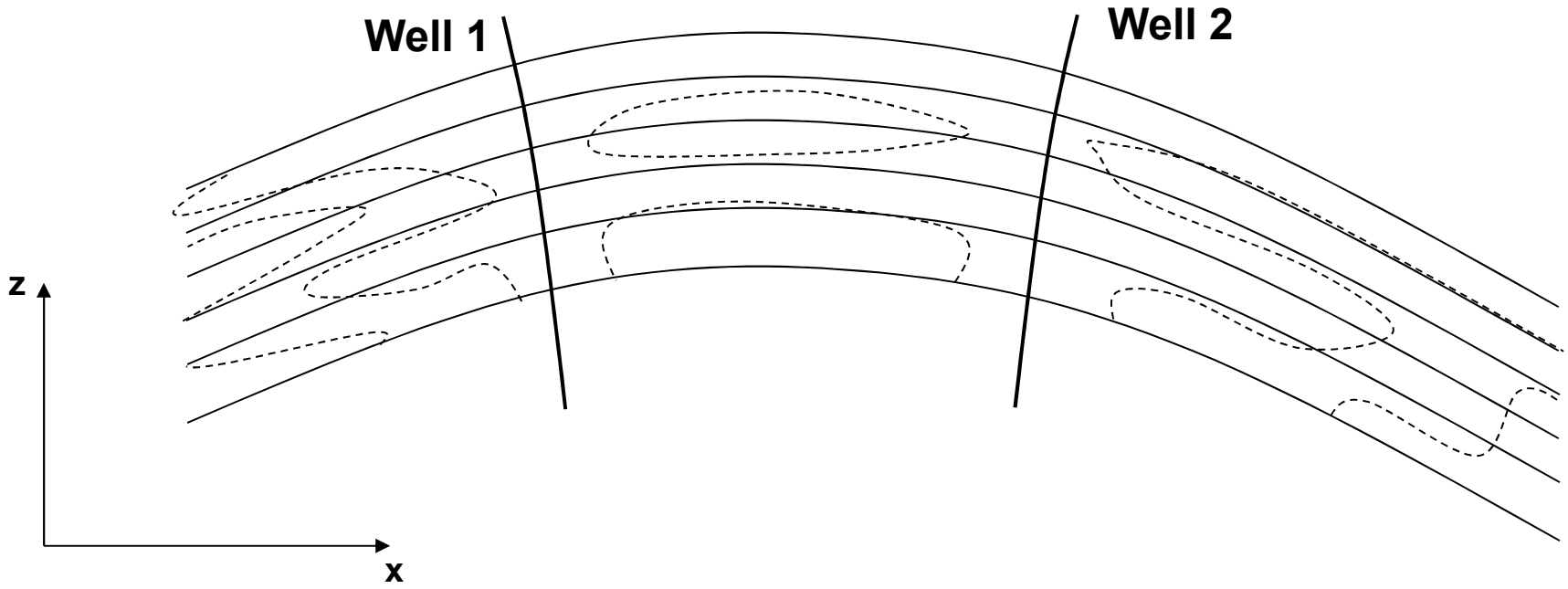
assume blank cells are simulated also

Reservoir Property Model



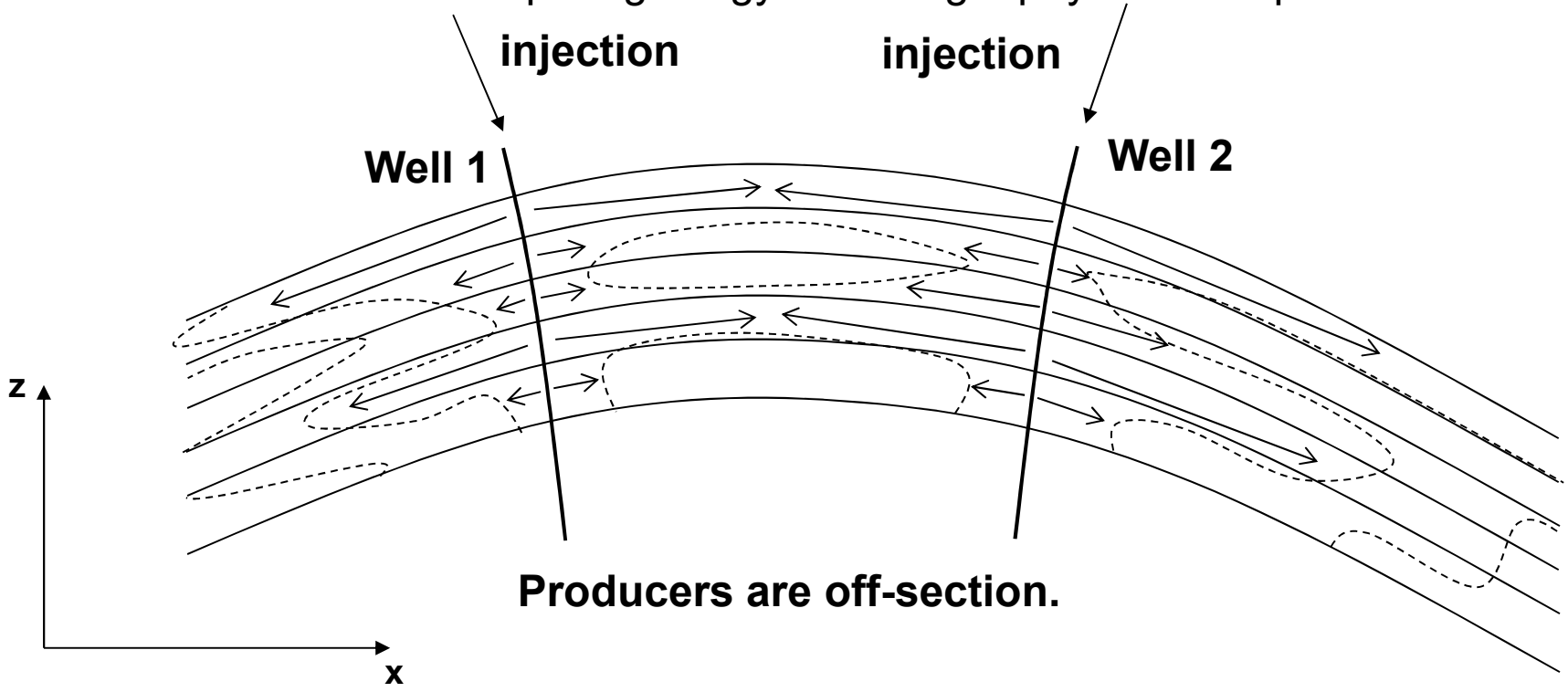
Steps in Sequential Gaussian Simulation, Take 2

8. Restore to original framework



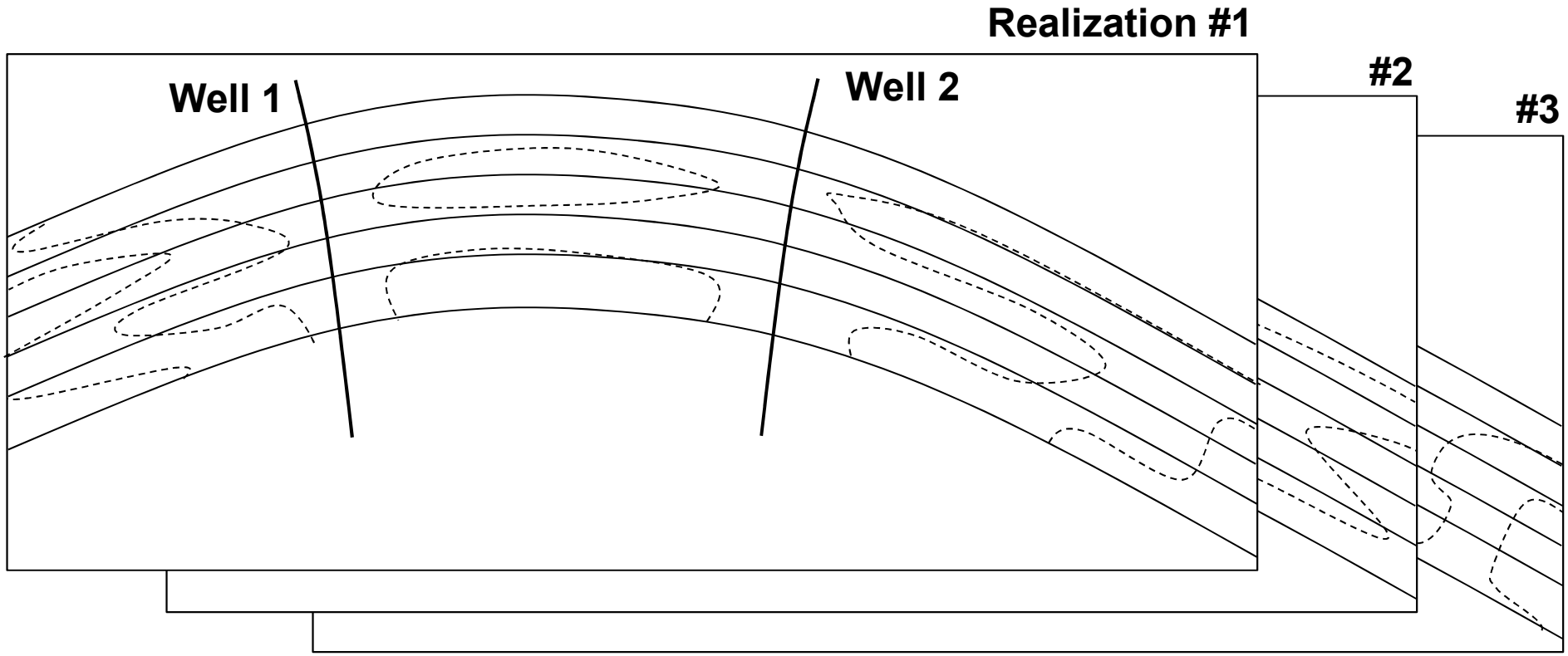
Steps in Sequential Gaussian Simulation, Take 2

9. Check honor concept of geology? Match geophysical and production data?

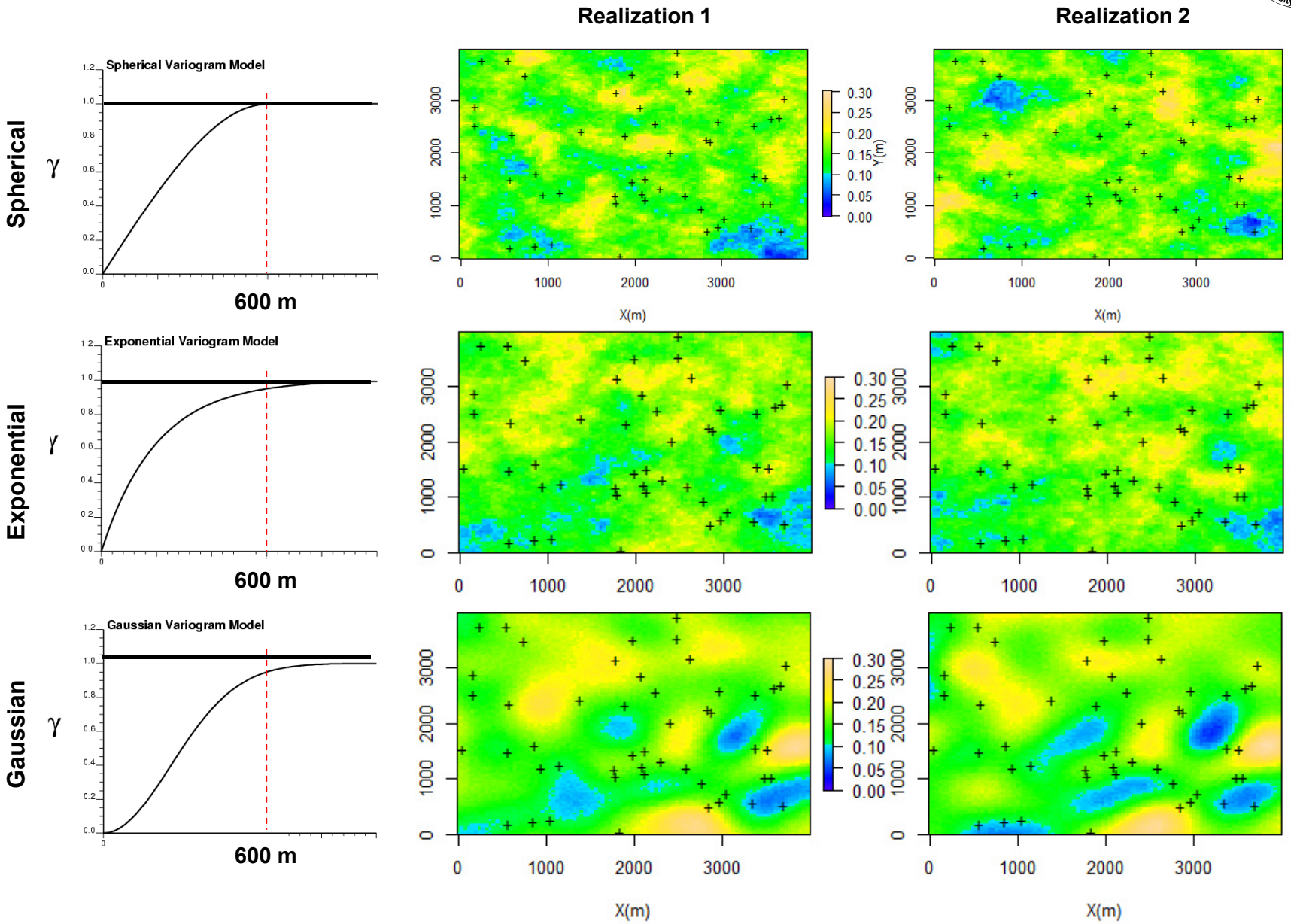


Steps in Sequential Gaussian Simulation, Take 2

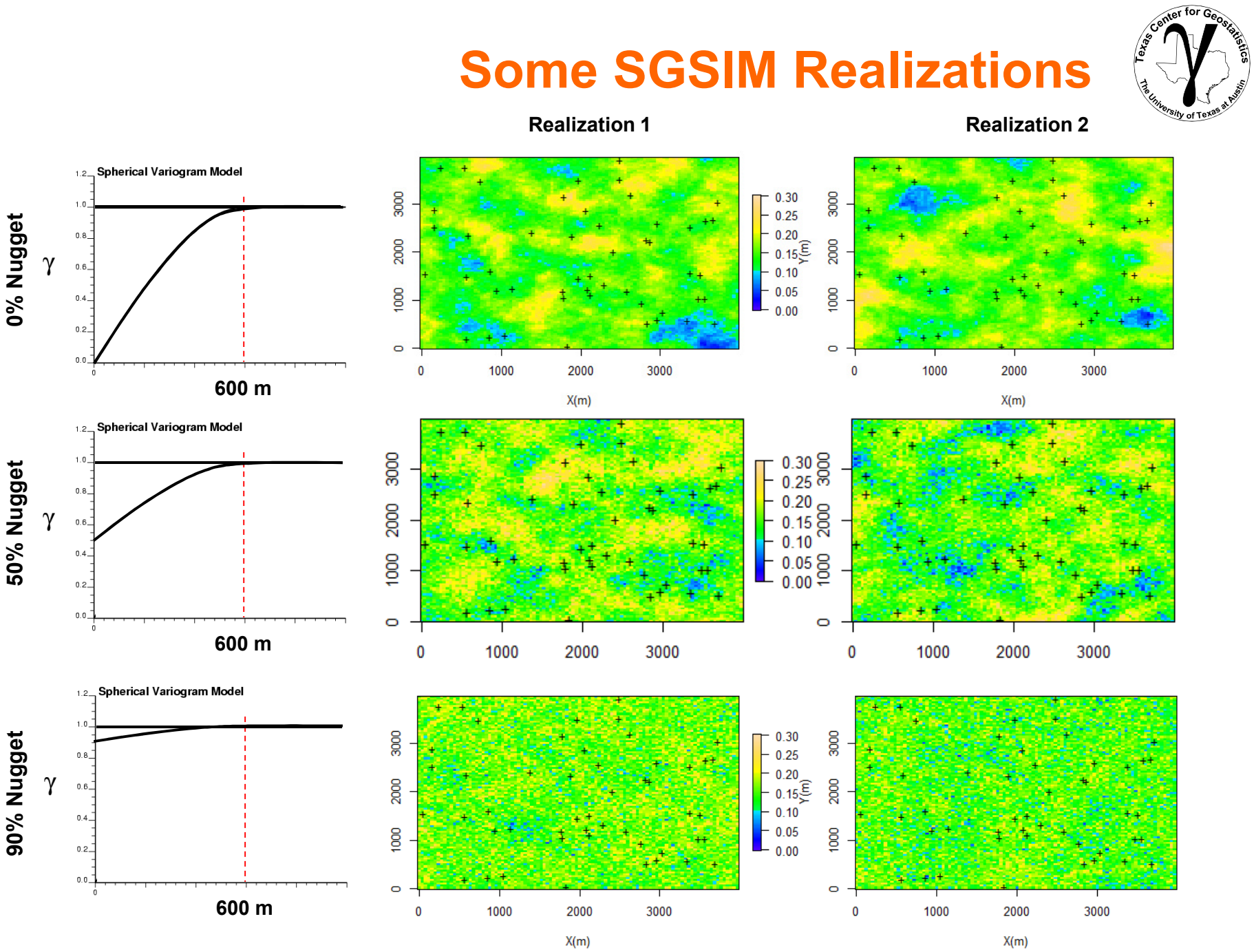
10. Calculate multiple realizations to represent uncertainty.



Some SGSIM Realizations

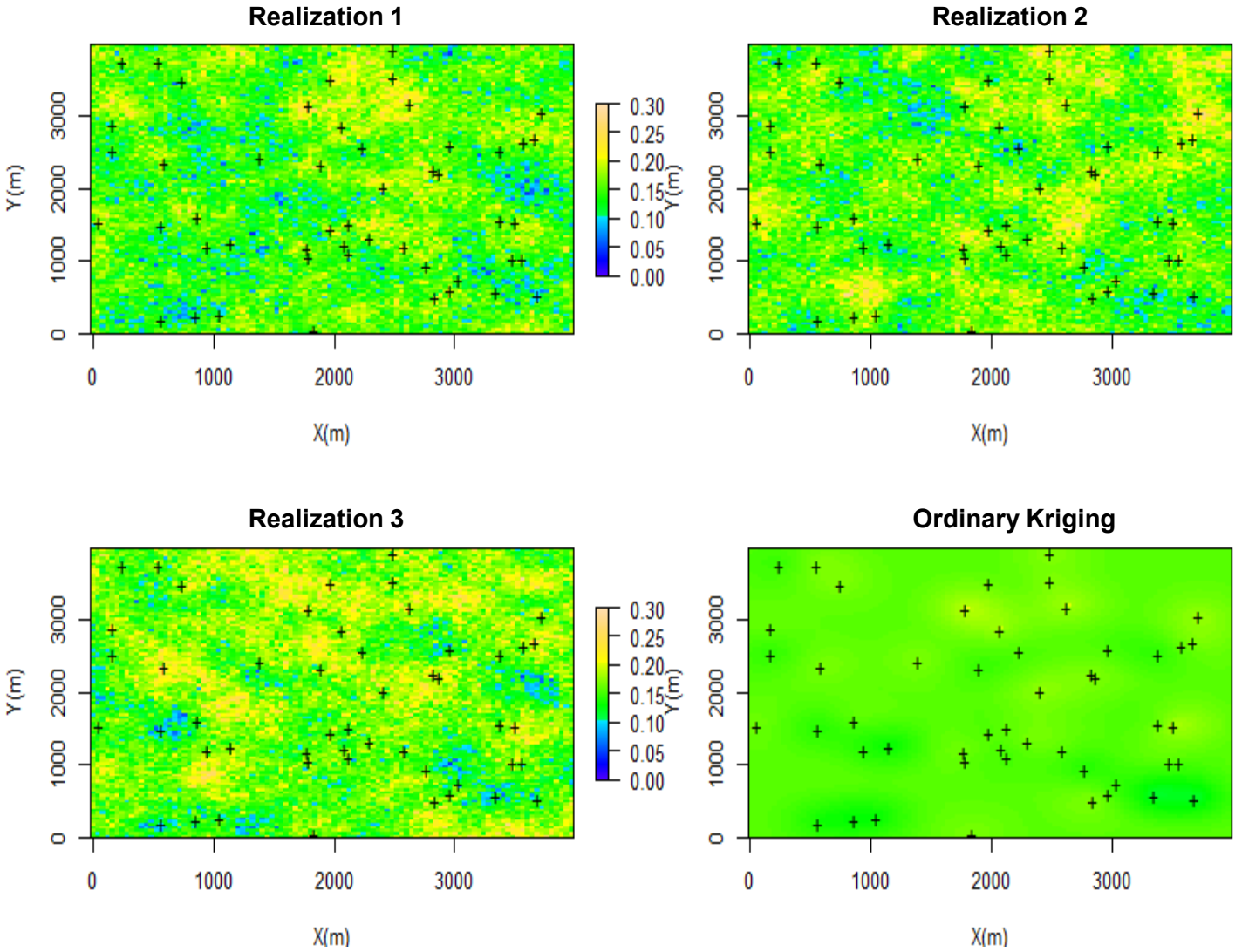


Some SGSIM Realizations



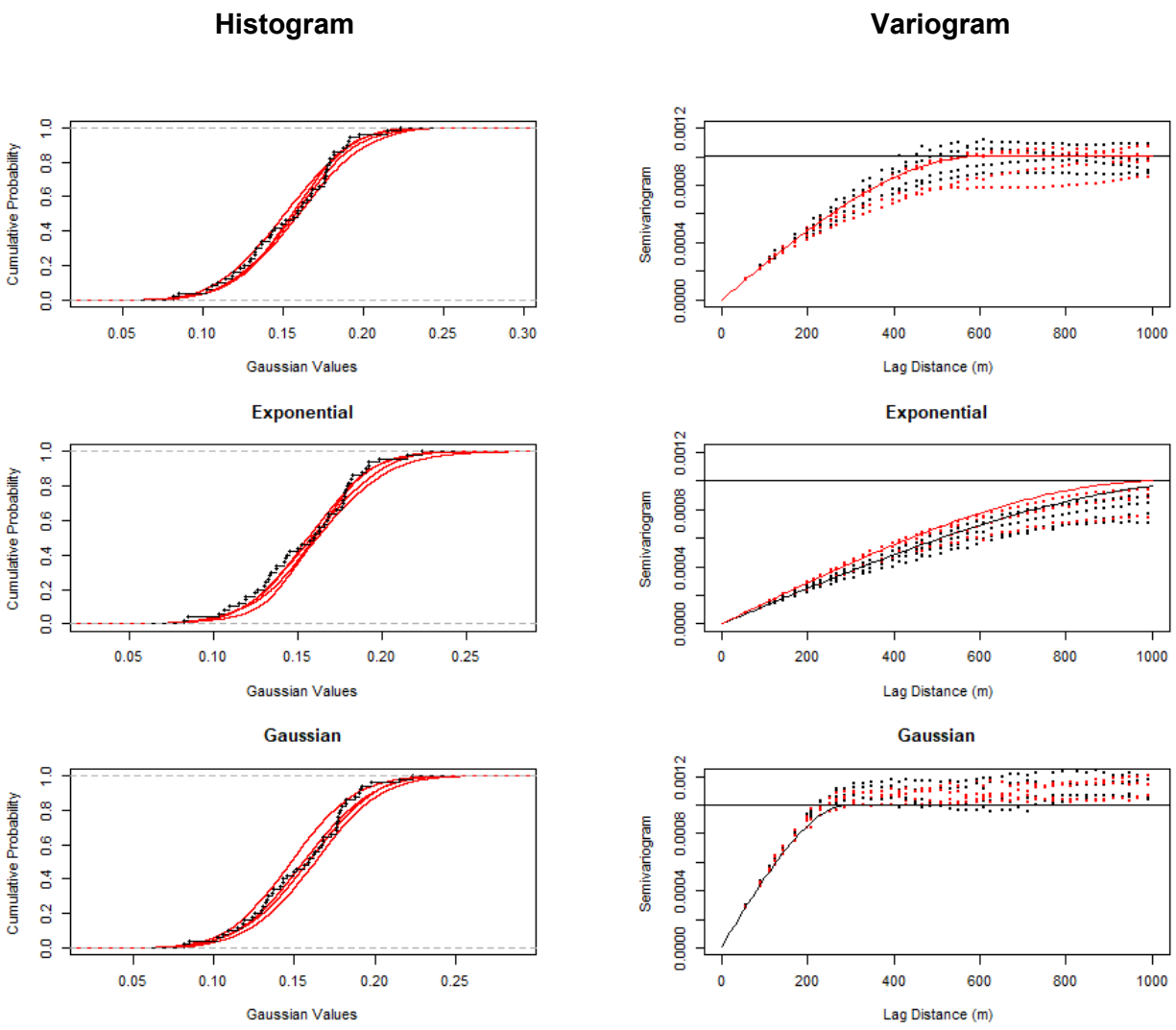
Example of Estimation and Simulation

- Compare kriging and simulation.

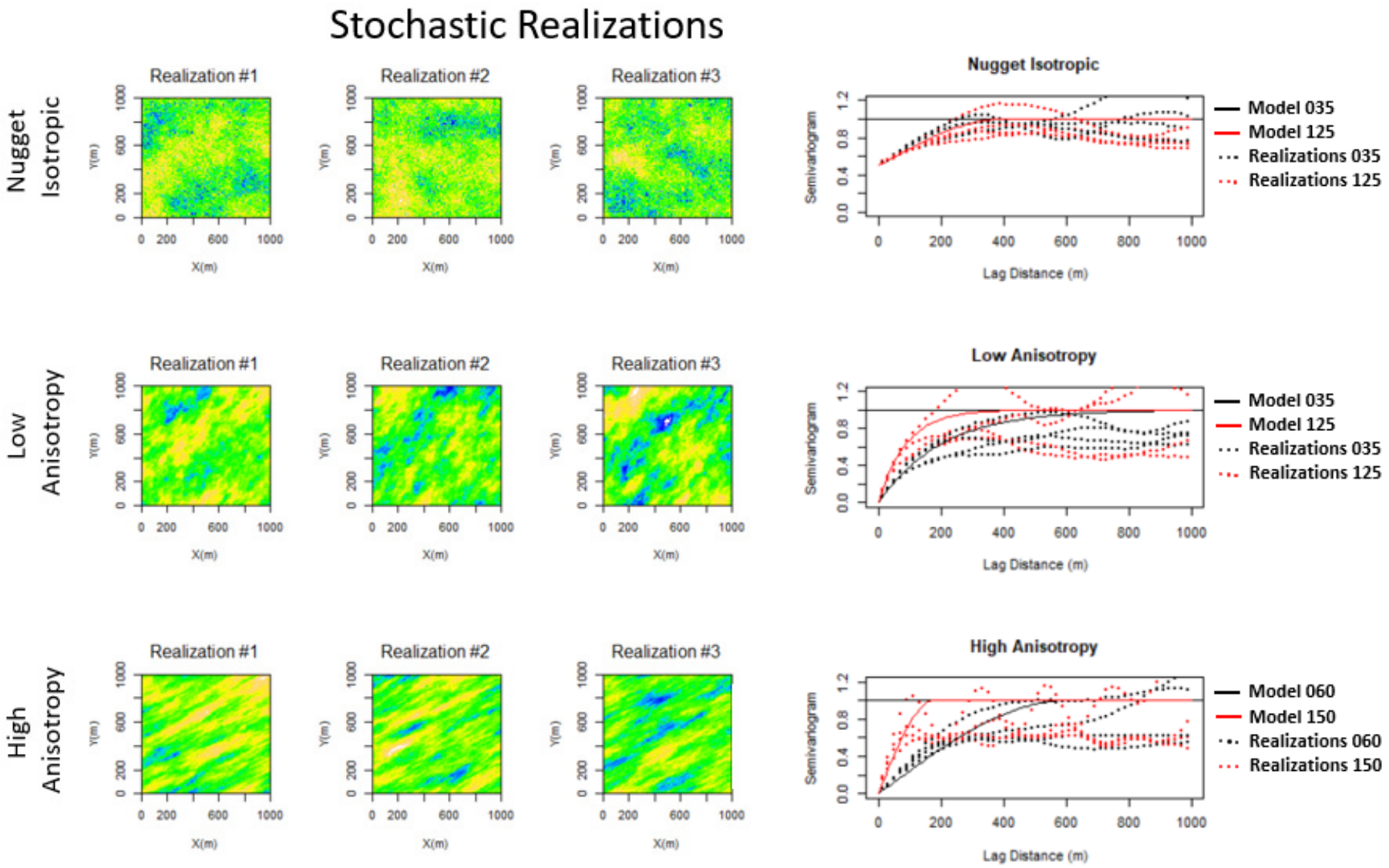


Ergodic Fluctuations

- Expect some statistical fluctuation in the input statistics
- These are a function of the ratio of spatial continuity to the size of the model.
  - If model is large relative to spatial continuity range then fluctuations should minimal
  - If model is small relative to spatial continuity range then fluctuations may be extreme



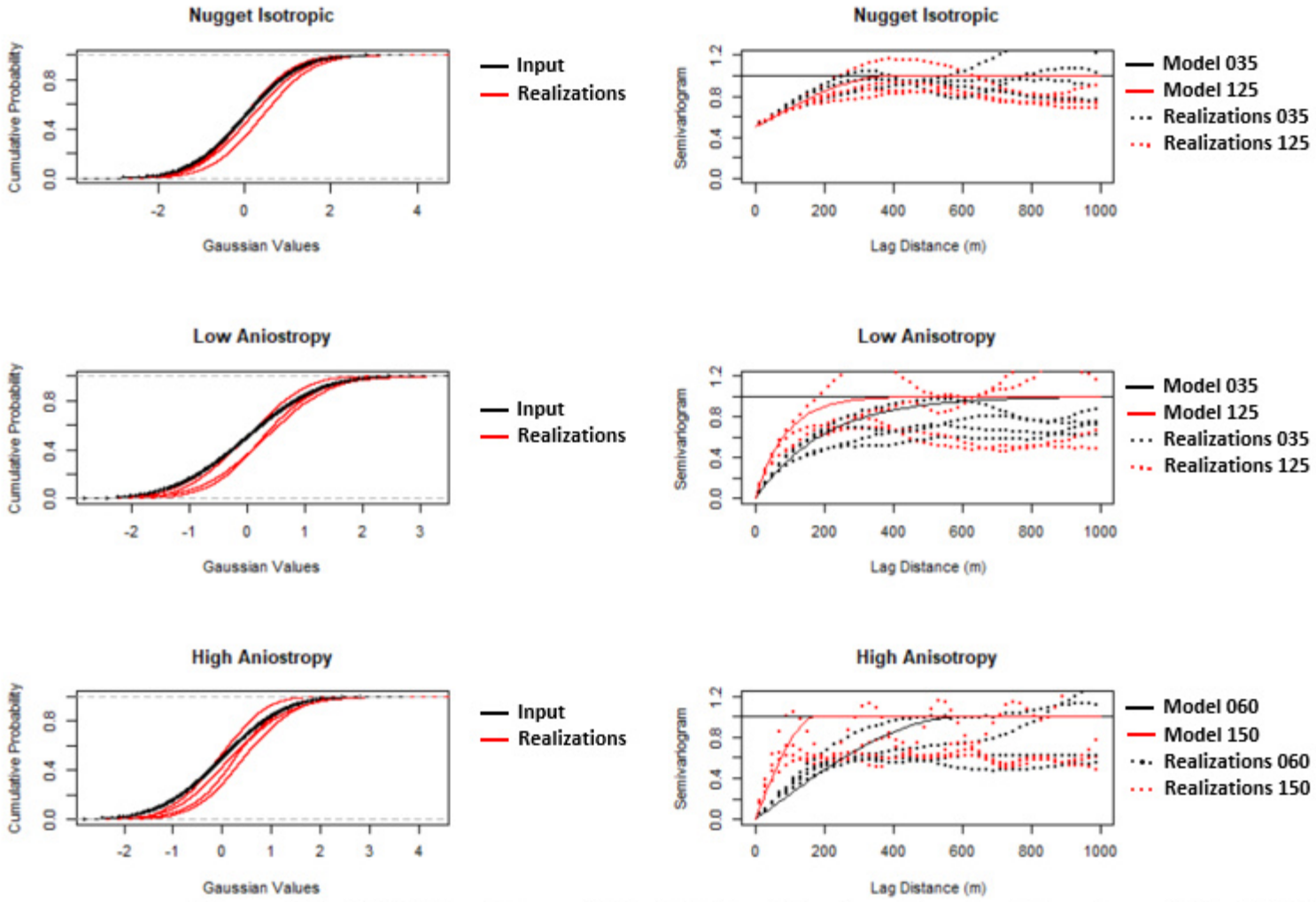
Sequential Gaussian Simulation in R with gstat Package



Results of the kriging_demo.r demo from Files/Code

Sequential Gaussian Simulation in R with gstat Package

Minimum Acceptance Checks (Distribution, Variogram)



Pebesma, E.J., 2004. Multivariable geostatistics in S: the gstat package. Computers & Geosciences 30 (7), 683-691

Results of the kriging_demo.r demo from Files/Code

Gaussian Simulation Hands-on

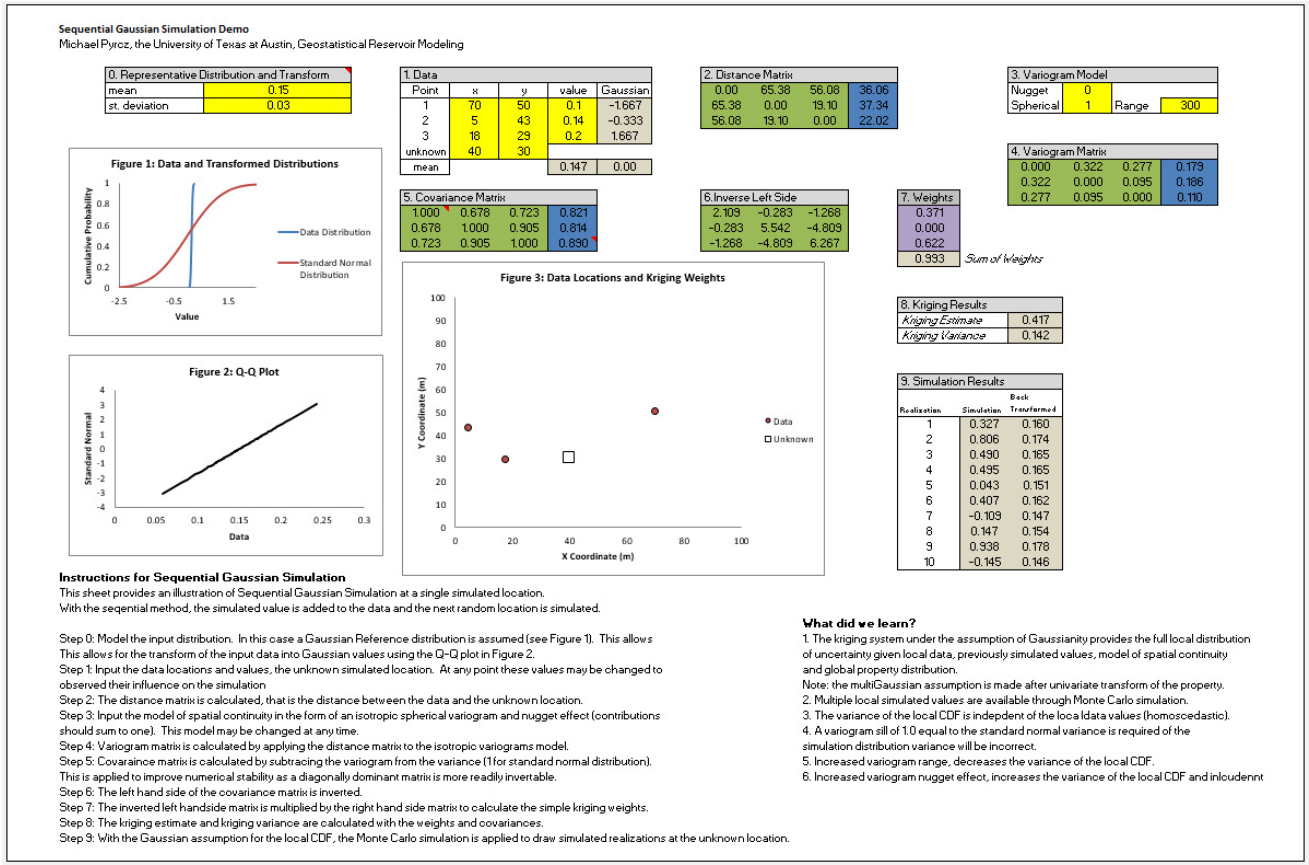
Sequential Gaussian Simulation:

Things to try:

1. Set the variogram range very small relative to data spacing.

2. Set the variogram range very large.

Evaluate the behavior of the 10 local realizations.



Sequential Gaussian Simulation Demo

Michael Pyrcz, the University of Texas at Austin, Geostatistical Reservoir Modeling

0. Representative Distribution and Transform	
mean	0.15
st. deviation	0.03

1. Data				
Point	x	y	value	Gaussian
1	70	50	0.1	-1.667
2	5	43	0.14	-0.333
3	18	29	0.2	1.667
unknown	40	30		
mean			0.147	0.00

2. Distance Matrix			
0.00	65.38	56.08	36.06
65.38	0.00	19.10	37.34
56.08	19.10	0.00	22.02

3. Variogram Model			
Nugget	0		
Spherical	1	Range	300

4. Variogram Matrix			
0.000	0.322	0.277	0.179
0.322	0.000	0.095	0.186
0.277	0.095	0.000	0.110

5. Covariance Matrix			
1.000	0.678	0.723	0.821
0.678	1.000	0.905	0.814
0.723	0.905	1.000	0.890

6. Inverse Left Side			
2.109	-0.283	-1.268	
-0.283	5.542	-4.809	
-1.268	-4.809	6.267	

7. Weights	
0.371	
0.000	
0.622	
0.993	*Sum of Weights*

8. Kriging Results	
Kriging Estimate	0.417
Kriging Variance	0.142

9. Simulation Results		
Realization	Simulation	Back Transformed
1	0.327	0.160
2	0.806	0.174
3	0.490	0.165
4	0.495	0.165
5	0.043	0.151
6	0.407	0.162
7	-0.109	0.147
8	0.147	0.154
9	0.938	0.178
10	-0.145	0.146

Figure 1: Data and Transformed Distributions

Figure 2: Q-Q Plot

Figure 3: Data Locations and Kriging Weights

Instructions for Sequential Gaussian Simulation

This sheet provides an illustration of Sequential Gaussian Simulation at a single simulated location.
With the seqential method, the simulated value is added to the data and the next random location is simulated.

Step 0: Model the input distribution. In this case a Gaussian Reference distribution is assumed (see Figure 1). This allows
This allows for the transform of the input data into Gaussian values using the Q-Q plot in Figure 2.
Step 1: Input the data locations and values, the unknown simulated location. At any point these values may be changed to observed their influence on the simulation
Step 2: The distance matrix is calculated, that is the distance between the data and the unknown location.
Step 3: Input the model of spatial continuity in the form of an isotropic spherical variogram and nugget effect (contributions should sum to one). This model may be changed at any time.
Step 4: Variogram matrix is calculated by applying the distance matrix to the isotropic variograms model.
Step 5: Covarainice matrix is calculated by subtracing the variogram from the variance (1 for standard normal distribution). This is applied to improve numerical stability as a diagonally dominant matrix is more readily invertable.
Step 6: The left hand side of the covariance matrix is inverted.
Step 7: The inverted left handside matrix is multiplied by the right hand side matrix to calculate the simple kriging weights.
Step 8: The kriging estimate and kriging variance are calculated with the weights and covariances.
Step 9: With the Gaussian assumption for the local CDF, the Monte Carlo simulation is applied to draw simulated realizations at the unknown location.

What did we learn?

1. The kriging system under the assumption of Gaussianity provides the full local distribution of uncertainty given local data, previously simulated values, model of spatial continuity and global property distribution.
Note: the multiGaussian assumption is made after univariate transform of the property.
2. Multiple local simulated values are available through Monte Carlo simulation.
3. The variance of the local CDF is indepdent of the loca ldata values (homoscedastic).
4. A variogram sill of 1.0 equal to the standard normal variance is required of the simulation distribution variance will be incorrect.
5. Increased variogram range, decreases the variance of the local CDF.
6. Increased variogram nugget effect, increases the variance of the local CDF and inloudennt

File Name: Sequential_Gaussian_Simulation_Demo.xlsx File is at: https://git.io/fNgB1

Bayesian Perspective on Sequential Simulation

- Consider N nodes in a simulation model, $A_i, i = 1, \dots, N$

- We need to sample a realization from the joint distribution, $P(A_1, \dots, A_1)$, this would not be possible, it is typically a vast solution space. We use Bayes to make it practical:

  From recursive application of Bayes law:

$$P(A_1, \dots, A_1) = P(A_N|A_1, \dots, A_{N-1}) \cdot P(A_1, \dots, A_{N-1})$$

$$P(A_N|A_1, \dots, A_{N-1}) \cdot P(A_{N-1}|A_1, \dots, A_{N-2}) \cdot P(A_1, \dots, A_{N-2})$$

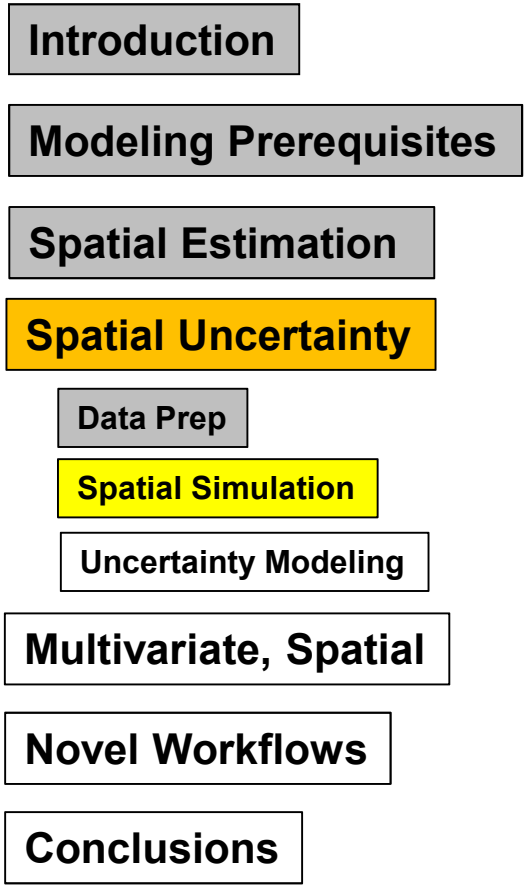
$$P(A_N|A_1, \dots, A_{N-1}) \cdot P(A_{N-1}|A_1, \dots, A_{N-2}) \cdot \cdots \cdot P(A_2|A_1) \cdot P(A_1)$$

- Now we can proceed sequentially to jointly simulate the *N* events A_j:
  - Draw A_1 from the marginal, $P(A_1)$
  - Draw A_2 from the conditional, $P(A_2|A_1 = a_1)$
  - Draw A_3 from the conditional, $P(A_3|A_1 = a_1, A_2 = a_2)$
  - …
  - Draw A_N from the conditional, $P(A_N|A_1 = a_1, A_2 = a_2, \dots, A_{N-1} = a_{N-1})$
- This is theoretically valid with no approximations/assumptions

Data Analytics and Geostatistics:
Data Preparation

Lecture outline . . .

- **Indicator Simulation**



Instructor: Michael Pyrcz, the University of Texas at Austin

The Indicator Formalism and Indictor Kriging (IK)

- Indicator Methods:
  - Estimation and Simulation with categorical variables with explicit control of spatial continuity of each category
  - Estimation and simulation with continuous variables with explicit control of the spatial continuity of different magnitudes
  - Requires indicator coding of data, a probability coding based on category or threshold
  - Requires indicator variograms to describe the spatial continuity.

The Indicator Formalism and Indictor Kriging (IK)

Indicator coding is transforming a random variable / function to a probability relative to a category or a threshold.

- If $I\{\mathbf{u}: z_k\}$ is an indicator for a categorical variable,
  - What is the probability of a realization equal to a category?

$$I(\mathbf{u};\boldsymbol{z_k}) = \begin{cases} 1, & \text{if } Z(\mathbf{u}) = \boldsymbol{z_k} \\ 0, & \text{otherwise} \end{cases}$$

  - e.g. given threshold, $z_2 = 2$, and data at $\mathbf{u}_1, z(\mathbf{u}_1) = 2$, then $I\{\mathbf{u}_1; z_2\} = 1$
  - e.g. given threshold, $z_1 = 1$, and a RV away from data, $\mathrm{Z}(\mathbf{u}_2)$ then $I\{\mathbf{u}_2; z_1\} = 0.25$

- If $I\{\mathbf{u}: z_k\}$ is an indicator for a continuous variable,
  - What is the probability of a realization less than or equal to a threshold?

$$I(\mathbf{u};\boldsymbol{z_k}) = \begin{cases} 1, & \text{if } Z(\mathbf{u}) \leq \boldsymbol{z_k} \\ 0, & \text{otherwise} \end{cases}$$

  - e.g. given threshold, $z_1 = 6\%$, and data at $\mathbf{u}_1, z(\mathbf{u}_1) = 8\%$, then $I\{\mathbf{u}_1; z_1\} = 0$
  - e.g. given threshold, $z_4 = 18\%$, and a RV,$\mathrm{Z}(\mathbf{u}_2) = N[16\%, 3\%]$ then $I\{\mathbf{u}_1; z_k\} = 0.75$

The Indicator Formalism and Indictor Kriging (IK)

Example of indicator transforms for a categorical variable.

Original Data	$I\{\mathbf{u}_\alpha; z_1 = 1\}$	$I\{\mathbf{u}_\alpha; z_2 = 2\}$	$I\{\mathbf{u}_\alpha; z_3 = 3\}$
$z(\mathbf{u}_1) = 3$	0	0	1
$z(\mathbf{u}_2) = 1$	1	0	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$z(\mathbf{u}_n) = 2$	0	1	0

Our $\mathrm{z}(\mathbf{u}_\alpha)$, $\alpha = 1, \ldots, n$, data become $k = 1, \ldots, K$ sets of n data, a new variable that indicates the probability of being each category.

The Indicator Formalism and Indictor Kriging (IK)

Example of indicator transforms for a continuous variable.

Original Data	$I\{\mathbf{u}_\alpha; z_1 = 5\%\}$	$I\{\mathbf{u}_\alpha; z_2 = 10\%\}$	$I\{\mathbf{u}_\alpha; z_3 = 15\%\}$
$z(\mathbf{u}_1) = 12\%$	0	0	1
$z(\mathbf{u}_2) = 4\%$	1	1	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$z(\mathbf{u}_n) = 17\%$	0	0	0

Our $z(\mathbf{u}_\alpha)$, $\alpha = 1, \ldots, n$, data become $k = 1, \ldots, K$ sets of n data, a new variable that indicates the probability of being less than or equal to each threshold.

The Indicator Formalism and Indictor Kriging (IK)

Then we can use these indicator-based probabilities for spatial estimates for each category or threshold.

- The IK process consists of discretizing the interval of variability of the continuous attribute z with a series of *K* threshold values $z_k, k = 1, \ldots, K$.
- A conditional CDF is built by assembling the *K* indicator kriging estimates

$$z_I^*(\mathbf{u}_\alpha; z_k) = P^*(Z(\mathbf{u}_\alpha) \leq z_k | n)$$

  - where $z_I^*(\mathbf{u}_\alpha; z_k)$ is the indicator kriging estimate at location, $\mathbf{u}_\alpha$ for threshold z_k

or for the categorical case we work with each category.

$$z_I^*(\mathbf{u}_\alpha; z_k) = P^*(Z(\mathbf{u}_\alpha) = z_k | n)$$

  - where $z_I^*(\mathbf{u}_\alpha; z_k)$ is the indicator kriging estimate at location, $\mathbf{u}_\alpha$ for threshold z_k

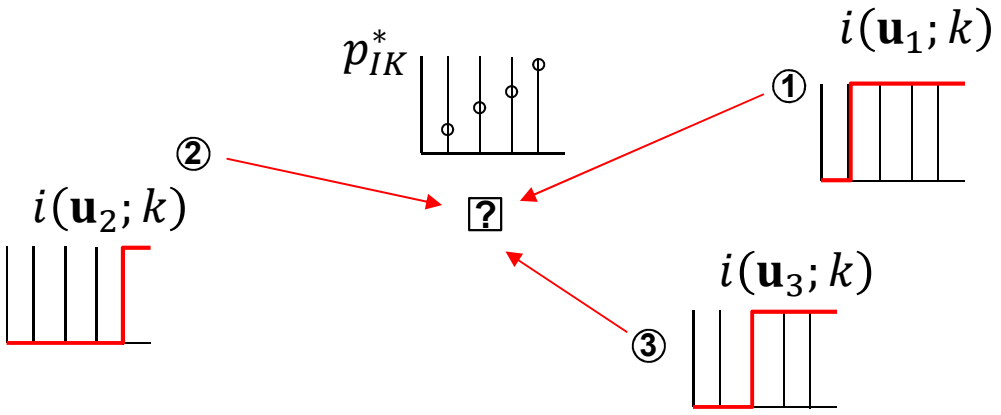
Conditional CDF / CCDF is the CDF at an unsampled locations estimated by local data.

The Indicator Formalism and Indictor Kriging (IK)

The indicator kriging estimator:

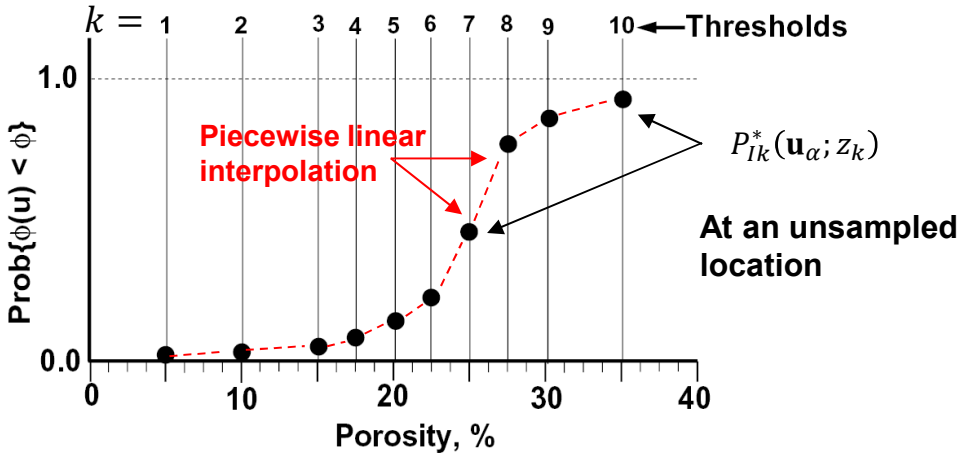
$$p^*_{IK}(\mathbf{u};k) = \sum_{\alpha=1}^{n} \lambda_\alpha(k) \cdot i(\mathbf{u}_\alpha;k) + \left(1 - \sum_{\alpha=1}^{n} \lambda_\alpha(k)\right) \cdot p(k)$$

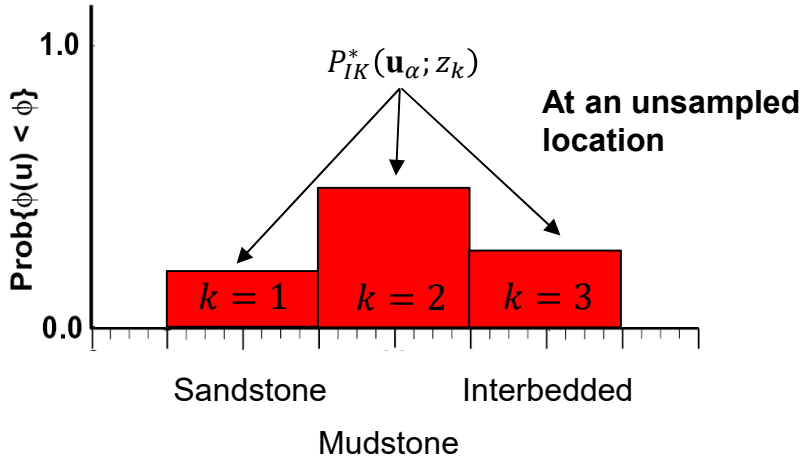
where $\lambda_\alpha(k)$ is the indicator kriging weight for data α and category / threshold k, $i(\mathbf{u}_\alpha;k)$ is the k category / threshold indicator transform of the data at location α and $p(k)$ is the global or local mean categorical probability / continuous cumulative probability.



Result of Indicator Kriging (IK)

- **Indicator kriging estimates the conditional CDF at thresholds or categories at an unsampled location.**





- Establish a series of thresholds / categories:
  - May be related to critical thresholds and should enough thresholds to represent the local distributions of uncertainty
  - By estimating **probability $\leq$ for each threshold and interpolation** or **probabilities for each category** we are directly estimating the distribution of uncertainty at an unsampled location without distribution assumption.

The Indicator Formalism and Indictor Kriging (IK)

Indicator kriging includes the following steps, for each threshold or category:

1. Application of indicator transform
2. Calculation of an indicator variogram
3. Indicator kriging to estimate the probability at an unsampled location

Indicator Variogram Models

How do we calculate spatial continuity for the indicator approach?

The indicator variogram:

$$\gamma_I(\mathbf{h}; z_k) = \frac{1}{2N(\mathbf{h})} \sum_{i=1}^{N(\mathbf{h})} [i(\mathbf{u}; z_k) - i(\mathbf{u}+\mathbf{h}; z_k)]^2$$

- Note: for hard data the indicator transform $i(\mathbf{u}; z_k)$ is either 0 or 1, in which case the $[i(\mathbf{u}; z_k) - i(\mathbf{u}+\mathbf{h}; z_k)]^2$ is equal to 0 when the values at head and tail are $\leq \boldsymbol{z_k}$ (continuous) or $= \boldsymbol{z_k}$ (categorical) or 1 when they are different.

- Therefore, the indicator variogram is ½ the proportion of pairs that change! The indicator variogram can be related to probability of change over a lag distance.

Indicator Variogram Models

	Cumulative Class %	Nugget	Exponential			Spherical		
			3a	sill	anis	range	sill	anis
First Cutoff	6.1	0.17	18.0	0.50	2.3	100.0	0.33	10.0
Second Cutoff	15.5	0.11	47.7	0.54	3.3	150.0	0.35	10.0
Third Cutoff	23.3	0.13	90.0	0.58	5.0	170.0	0.29	10.0
Fourth Cutoff	32.7	0.13	90.0	0.61	6.2	160.0	0.26	10.0
Fifth Cutoff	43.4	0.12	108.0	0.68	6.2	91.0	0.20	7.1
Sixth Cutoff	57.7	0.12	108.0	0.68	7.1	85.0	0.20	7.1
Seventh Cutoff	75.4	0.22	144.0	0.69	6.7	66.0	0.09	5.9

- Standardize all points and models to a unit variance from the original variance.
- Note $p(1 - p)$ is the variance of an indicator with proportion of 1's as p.

- Model the variograms with smoothly changing parameters for a *consistent* description
  - Common dataset → imparts consistency
  - Consistency between indicator variograms reduces order relations in IK/SIS (more later)
  - Allows straightforward interpolation of models for new cutoffs

Table from Pyrcz and Deutsch (2014)

The Indicator Formalism and Indictor Kriging (IK)

Some Comments on the Indicator Approach for Estimation and Simulation

- A variogram is needed for each threshold → more difficult inference problem, however, there is greater flexibility
- Resulting model of uncertainty is not Gaussian (avoid maximum entropy issue)
- More readily integrates data of different types (more later on soft data)
- May use Sequential Indicator Simulation (will demonstrate next as extension to indicator kriging).
  - » Commonly used for categorical variable like facies
  - » Sometimes used for continuous variables like porosity

Indicator Kriging (IK)

Indicator Kriging Workflow

For all locations:

- For all thresholds:
  - find all relevant data: *n*
  - code all data as indicator data at the current threshold:

$$i(\mathbf{u}_\alpha; z_k) = Prob^* \left\{ Z(\mathbf{u})_\alpha \leq z_k \right\}$$

  - estimate the indicator function at the current threshold at this location:

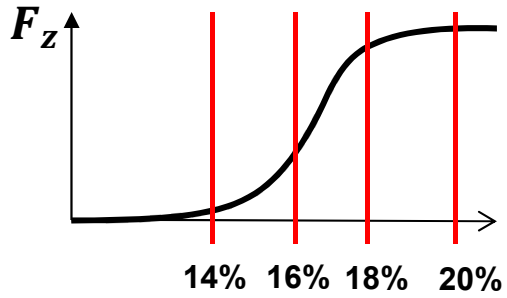
$$p^*_{IK}(\mathbf{u}; k) = \sum_{\alpha=1}^{n} \lambda_\alpha(k) \cdot i(\mathbf{u}_\alpha; k) + \left(1 - \sum_{\alpha=1}^{n} \lambda_\alpha(k) \right) \cdot p(k)$$

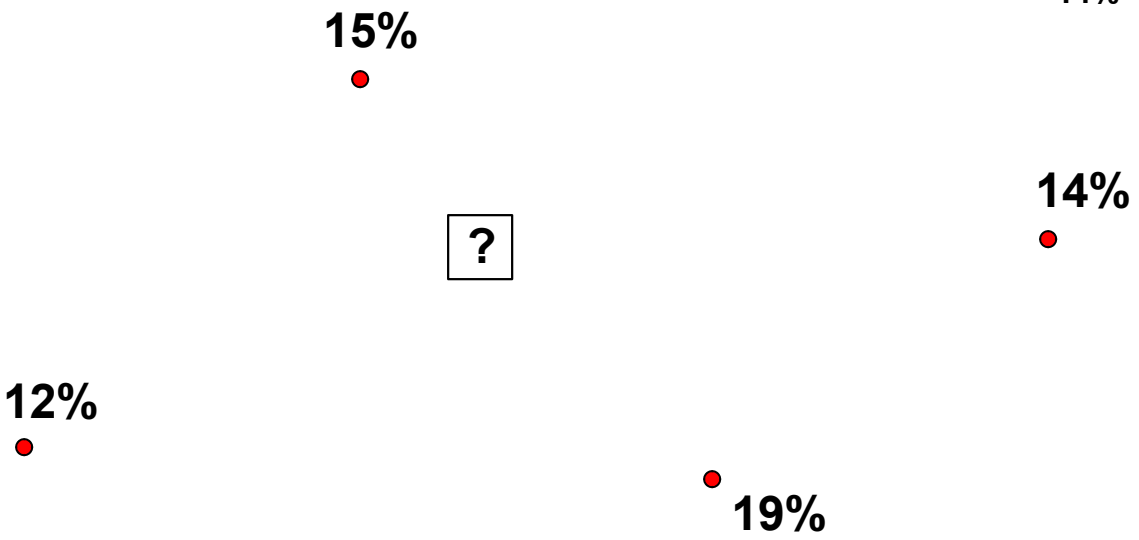
- Correct local distribution for order relations
- Use distribution for:
  - measure of uncertainty, probability intervals
  - probability to exceed given thresholds
  - E-type mean estimate, truncated statistics
  - stochastic simulation

Indicator Kriging (IK)

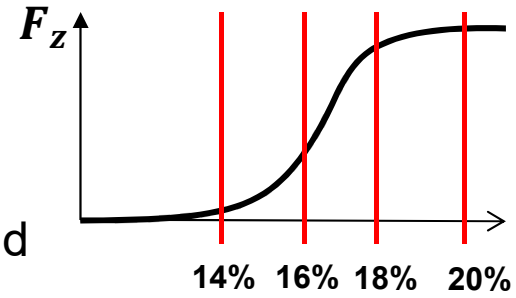
- Demonstration of Indicator Kriging

1. Assign thresholds





Indicator Kriging (IK)



- Demonstration of Indicator Kriging

1. Assign thresholds
2. Calculate indicator variograms for each threshold
   a) indicator transform
   b) calculate variogram

$i(\mathbf{u}_1; z = z_1)$ **0**

$i(\mathbf{u}_4; z = z_1)$ **1**

?

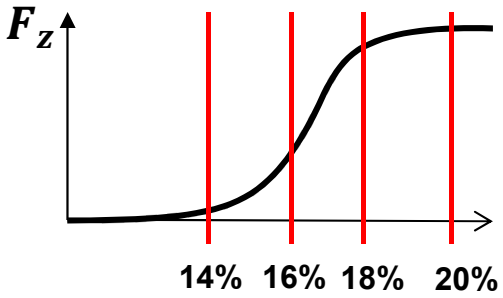
$i(\mathbf{u}_2; z = z_1)$ **1**

$i(\mathbf{u}_3; z = z_1)$ **0**

- Indicator Variogram:

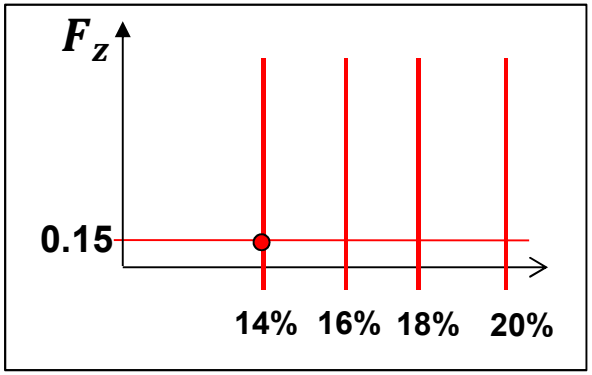
$$\gamma_I(\mathbf{h}; z_k) = \frac{1}{2N(\mathbf{h})} \sum_{i=1}^{N(\mathbf{h})} [i(\mathbf{u}; z_k) - i(\mathbf{u}+\mathbf{h}; z_k)]^2$$

Indicator Kriging (IK)



- Demonstration of Indicator Kriging

3. For each location:
4. For each threshold: $z = z_1 = 14\%$



$i(\mathbf{u}_1; z = z_1)$

0

$C_{1_{1,4}}$

?

$i(\mathbf{u}_4; z = z_1)$

0

$i(\mathbf{u}_2; z = z_1)$

1

$i(\mathbf{u}_3; z = z_1)$

0

- Solve for indicator kriging estimate of probability porosity at location ? is < z_1.

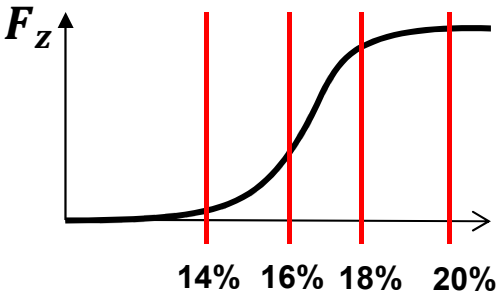
$$\begin{bmatrix} C_{1_{1,1}} & & C_{1_{1,n}} \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_n \end{bmatrix} = \begin{bmatrix} C_{1_{0,1}} \\ C_{1_{0,n}} \end{bmatrix}$$

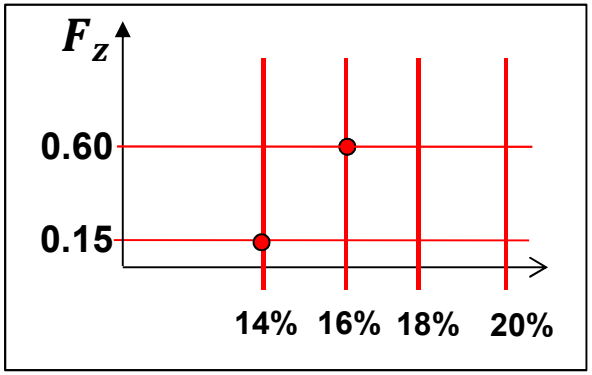
$$z^*(\mathbf{u}_1; z_1) = 0.15$$

Indicator Kriging (IK)

- Demonstration of Indicator Kriging

3. For each location:
4. For each threshold: $z = z_2 = 16\%$





$i(\mathbf{u}_1; z = z_2)$

1

?

$i(\mathbf{u}_4; z = z_2)$

1

$i(\mathbf{u}_2; z = z_2)$

1

$i(\mathbf{u}_3; z = z_2)$

0

- Solve for indicator kriging estimate of probability porosity at location ? is < z_1.

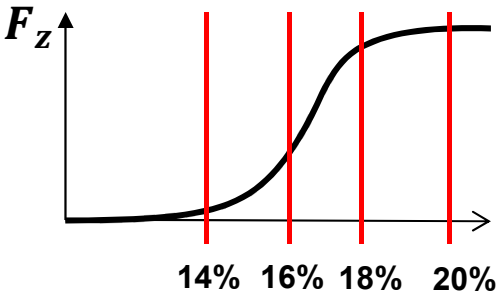
$$\begin{bmatrix} C_{2_{1,1}} & & C_{2_{n,n}} \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_n \end{bmatrix} = \begin{bmatrix} C_{2_{0,1}} \\ C_{2_{0,n}} \end{bmatrix}$$

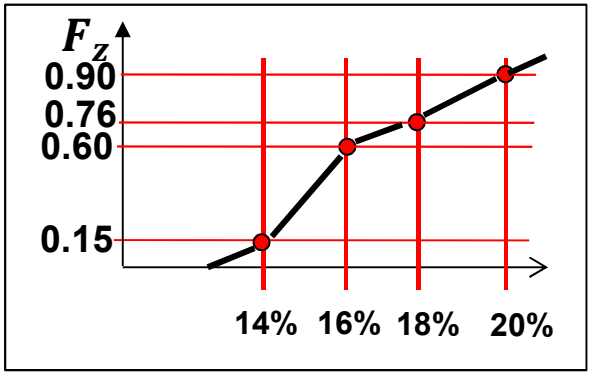
$$z^*(\mathbf{u}_0; z_2) = 0.60$$

Indicator Kriging (IK)

- Demonstration of Indicator Kriging

3. For each location:
4. For each threshold: $z = z_4 = 20\%$





$i(\mathbf{u}_1; z = z_4)$

1

?

$i(\mathbf{u}_4; z = z_4)$

1

$i(\mathbf{u}_2; z = z_4)$

1

$i(\mathbf{u}_3; z = z_4)$

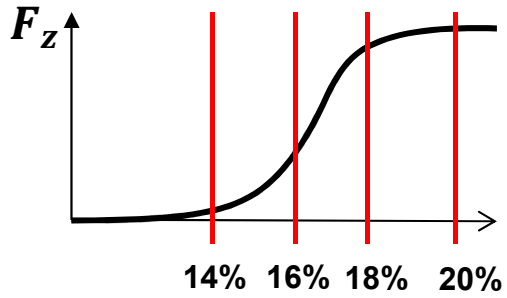
1

- Solve for indicator kriging estimate of probability porosity at location ? is < z_1.
- **Indicator kriging directly solves for the local distribution of CDF!**

$$\begin{bmatrix} C_{2_{1,1}} & & \\ & & \\ & & C_{2_{n,n}} \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \\ \lambda_n \end{bmatrix} = \begin{bmatrix} C_{2_{0,1}} \\ \\ C_{2_{0,n}} \end{bmatrix}$$

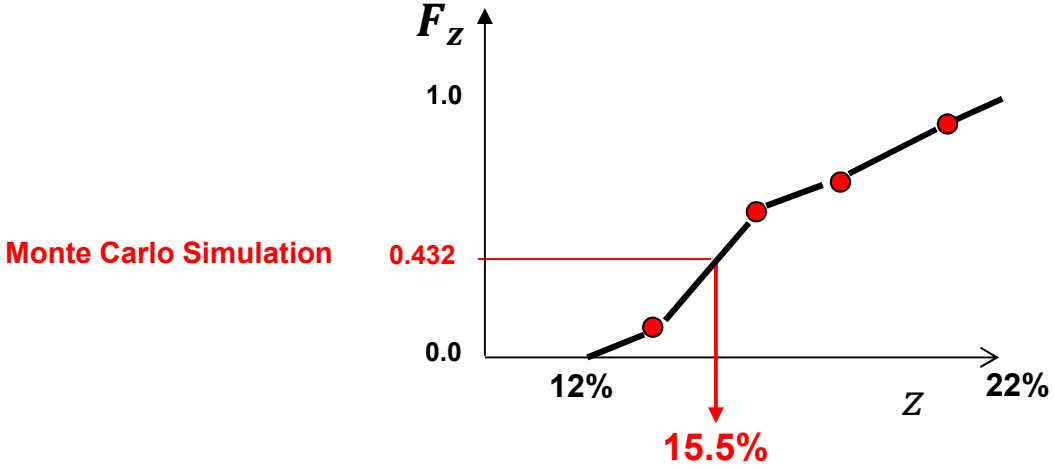
$$z^*(\mathbf{u}_0; z_2) = 0.60$$

Indicator Simulation (IS)



- Demonstration of Indicator Kriging

3. For each location:
4. For each threshold: $z = z_4 = 20\%$

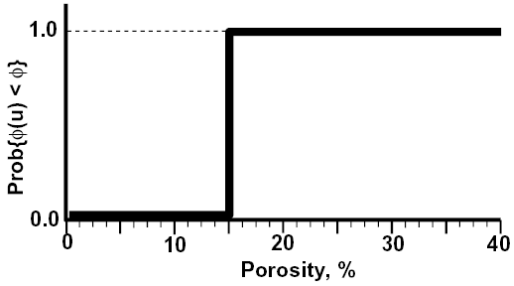


- Solve for indicator kriging estimate of probability porosity at location ? is < z_1.
- **Monte Carlo Simulation and treat the simulated value as data, indicator transform it and move to the next location on the random path!**

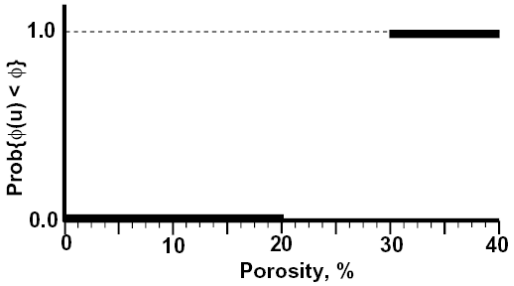
Data For Indicator Kriging

Hard and Soft Data with the Indicator Approach

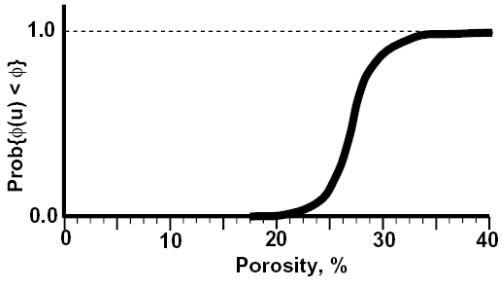
- Various constraints that may be applied to indicator coding



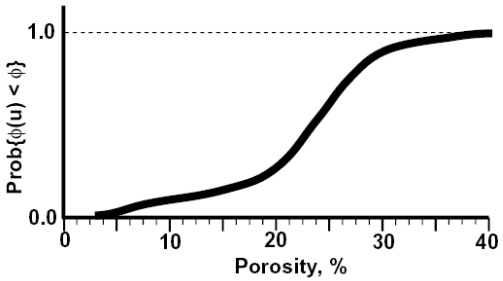
Local Hard Indicator Data



Inequality Constraint From Soft Secondary Data



Prior Distribution From Soft Secondary Data



Global Distribution From Stationary Histogram

Recall:

$$I(\mathbf{u};z_k)=\begin{cases}1, & \text{if } Z(\mathbf{u}) \le z_k \\ 0, & \text{otherwise}\end{cases}$$

Order Relations Correction

What is Order Relations? Why is it an Issue with Indicator Methods?

- Nonmonotonic behavior in continuous CDF (negative slope) or sum of categorical probabilities not equal to one.
- Cumulative probability at each threshold was solved with potentially a difference indicator variogram and with a separate kriging.
- There is no direct constraint to impose slope ≥ 0.0.
- Correction:
  - Continuous - take average of the forward and reverse constrained slope ≥ 0
  - Categorical - normalize sum of probabiiltes $= 0$

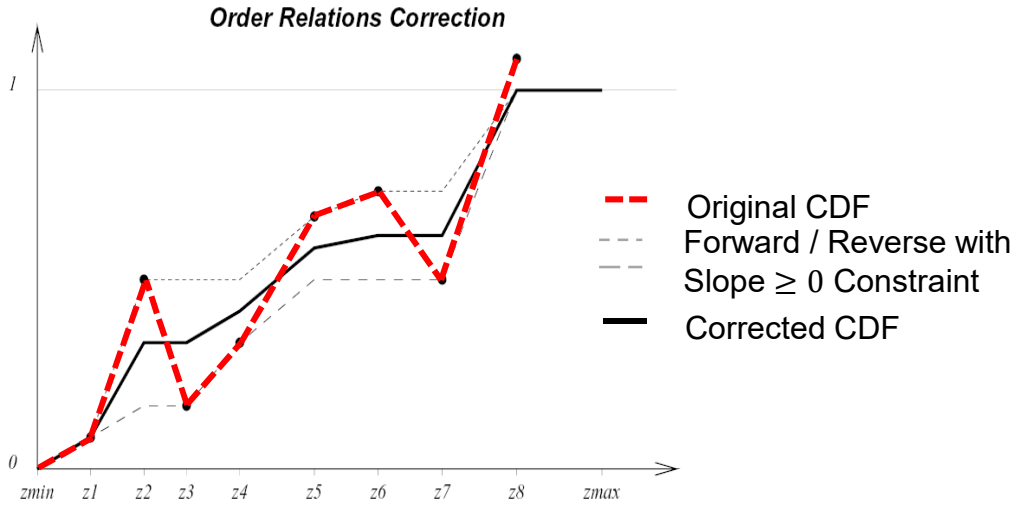
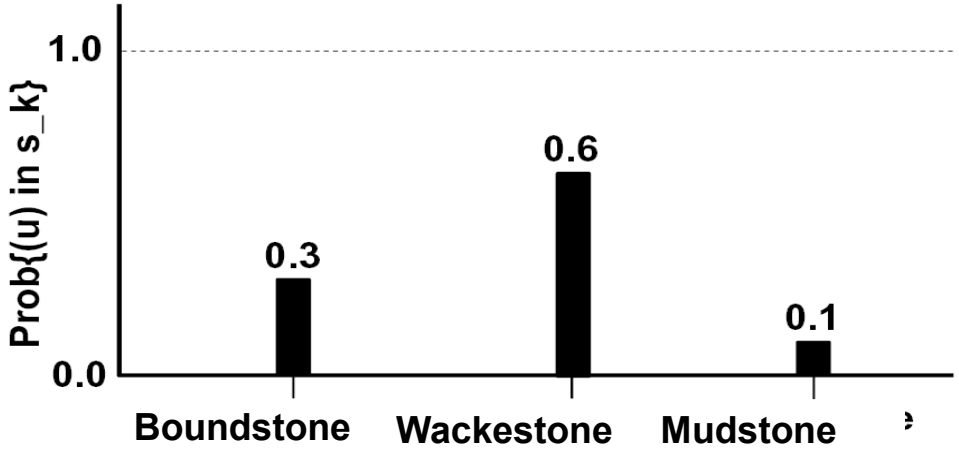


image from Pyrcz and Deutsch (2014)

Indicator Methods with Categorical Variables



- Consider a set of *K* categories: $\boldsymbol{z_k}, k = 1, \ldots, K$
- Indicator variable for each location and each category:

$$i(\mathbf{u}_\alpha; \boldsymbol{z_k}) = \begin{cases} 1 \text{ if category } \boldsymbol{z_k} \text{ prevails at location } \mathbf{u}_\alpha \\ 0 \text{ if not} \end{cases}$$

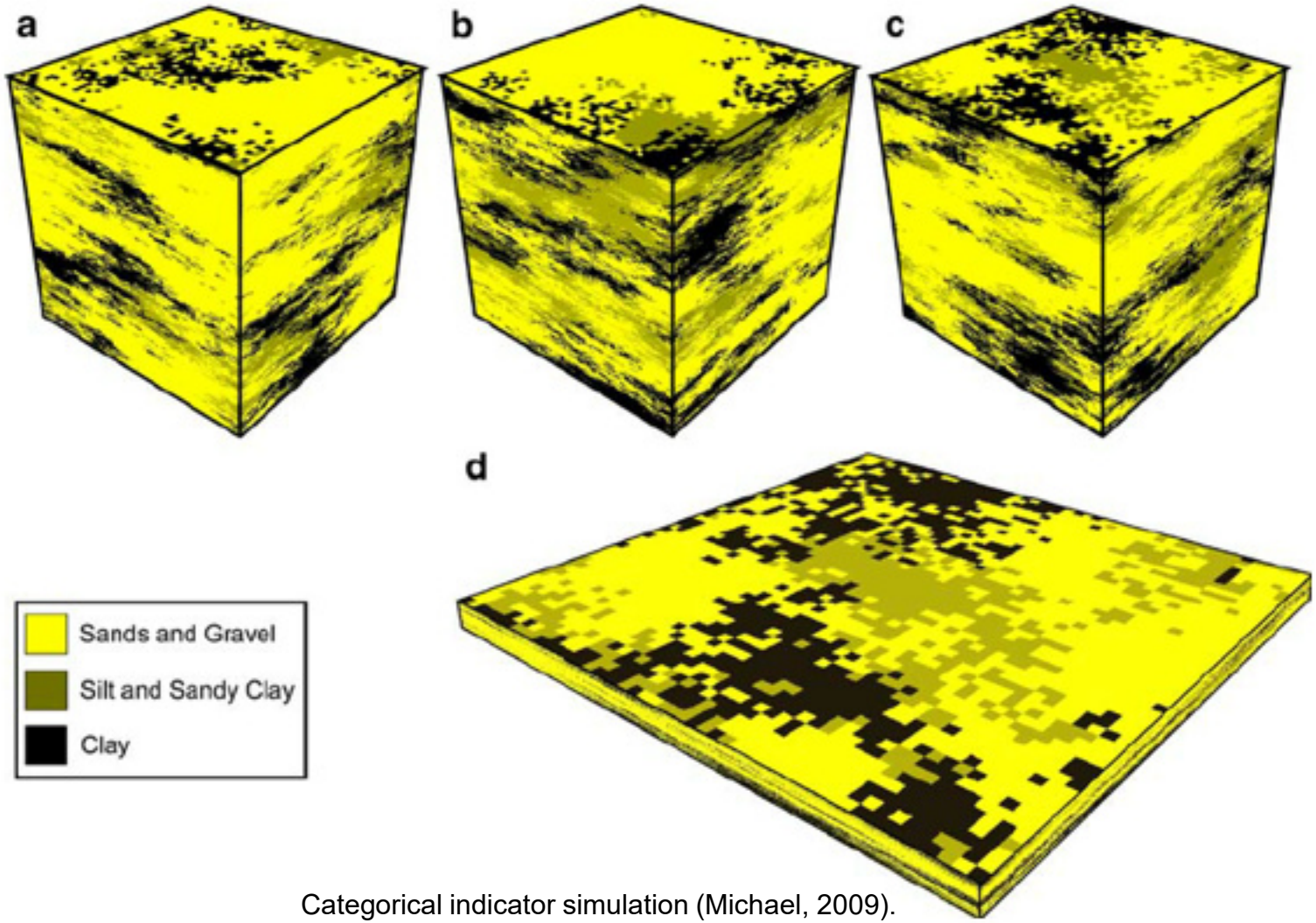
- Same procedure for indicator kriging with continuos variable
- Order relations:

$$\text{all } i(\mathbf{u}; \boldsymbol{z_k})^*, k = 1, \ldots, K \text{ must be } \geq 0$$

$$\sum_{k=1}^{K} i(\mathbf{u}; \boldsymbol{z_k})^* = 1.0$$

- Results are the probabilities that categories $\boldsymbol{z_k}, k = 1, \ldots, K$ at location $\mathbf{u}$

Example of Categorical Indicator Simulation

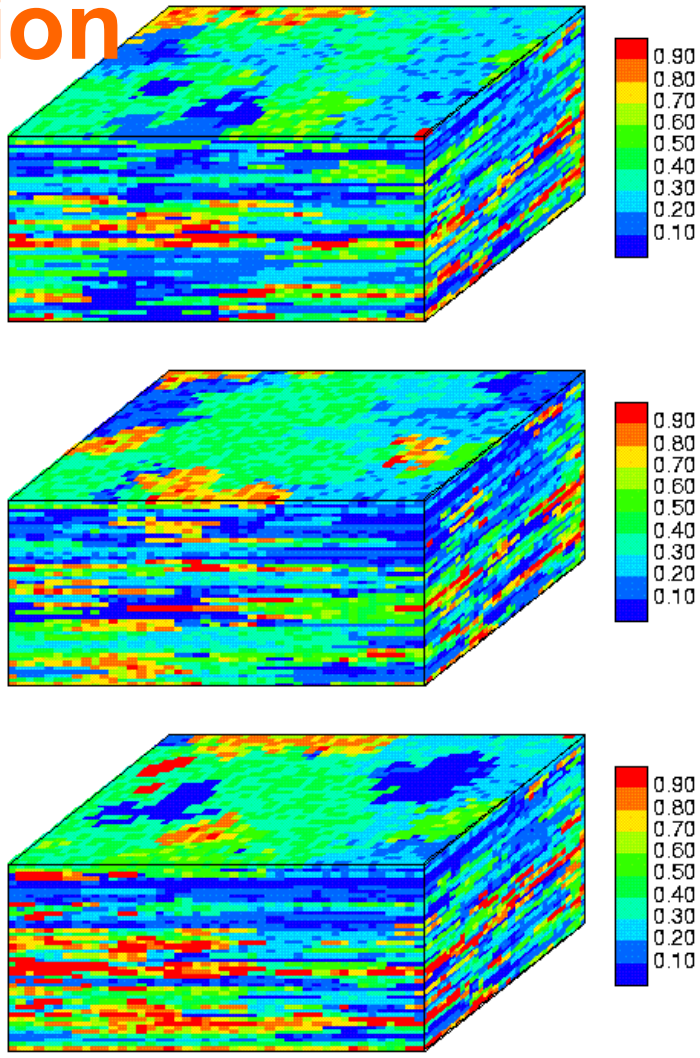


Categorical indicator simulation (Michael, 2009).

Example of Continuous Indicator Simulation

Continuous indicator simulation

- See the discontinuity across the continuous thresholds?
- My estimate is 0.7, 0.5, 0.3 and 0.1 were used.
- This is a known artifact with the continuous indicator simulation.



Three Vsh realizations for a shale group (Meehan and Verma, 1994)

Recall Indicator Kriging Hands-on

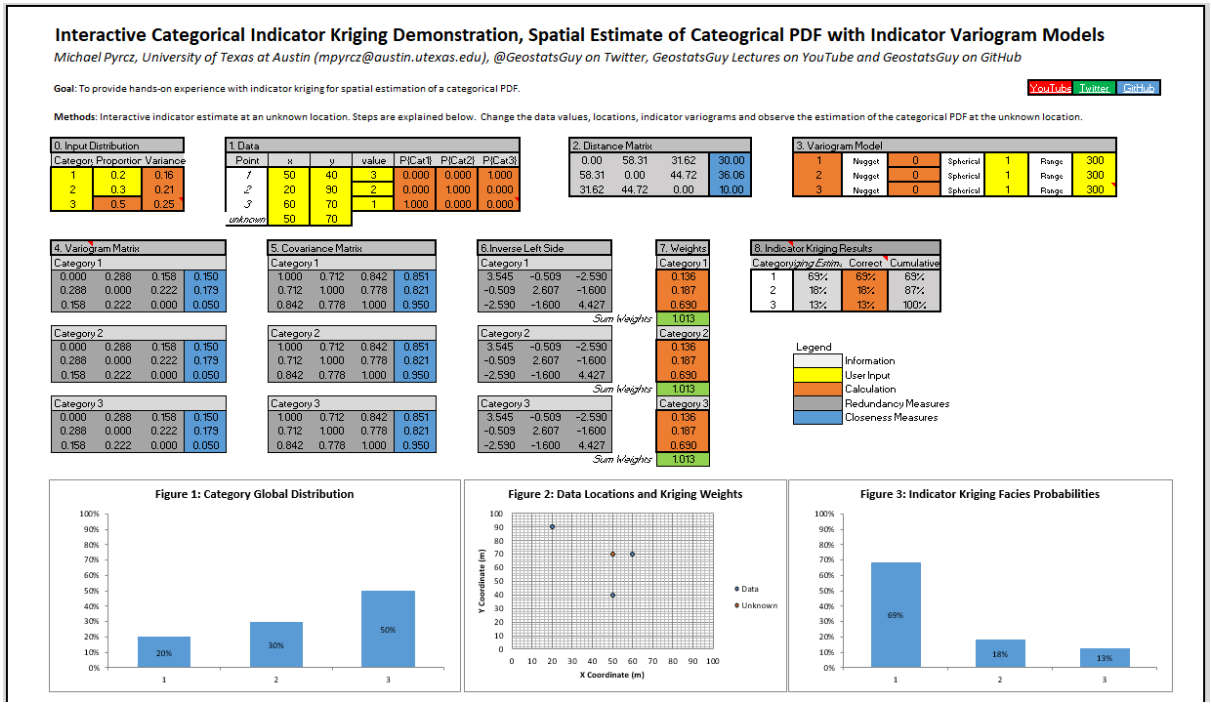
Here's an opportunity for experiential learning with Indicator Kriging.

- **Things to try:**

Pay attention to the kriging weights, kriging estimate and kriging variance while you:

1. Set the ranges all very small.
2. Set points 1 and 2 closer together.
3. Make the indicator variograms very different.

Observed the impact on the estimated PDF.



File Name: Indicator_Kriging_Demo.xlsx File is at: https://git.io/fpnY2

Summary of Indicators

Indicator Transforms

- Probability coding
- Indicator transform of continuous variable with thresholds
- Indicator transform of categorical variable, by-category

Indicator for Spatial Estimation

- Applied to estimate local CDF without assuming Gaussian distribution

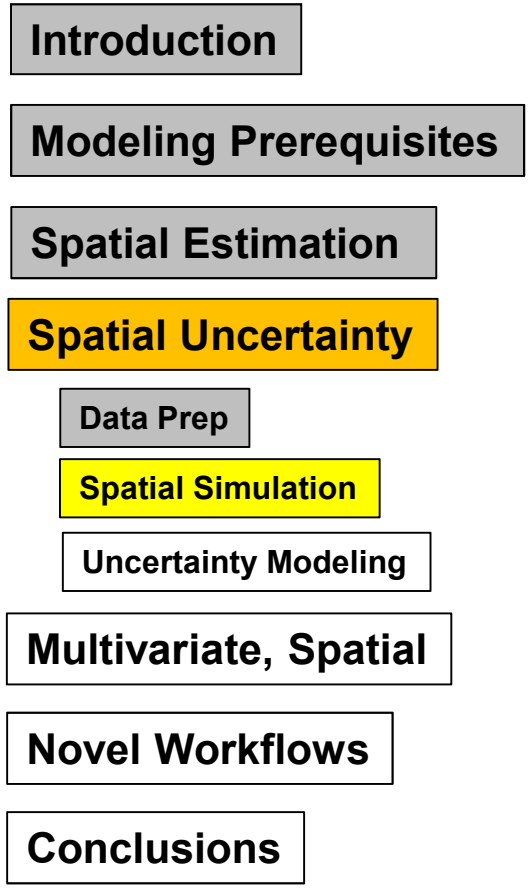
Indicator for Spatial Simulation

- Applied more often for categorical simulation
  - Replace simple / ordinary kriging with indicator kriging in the sequential context
  - i.e. Monte Carlo from indicator estimated CDF and transform simulated value for each threshold and use as data (sequential approach).

Data Analytics and Geostatistics:
Data Preparation

Lecture outline . . .

- **Gaussian Simulation**
- **Indicator Simulation**



Instructor: Michael Pyrcz, the University of Texas at Austin