DAYTUM – INTRODUCTION TO ENERGY MACHINE LEARNING

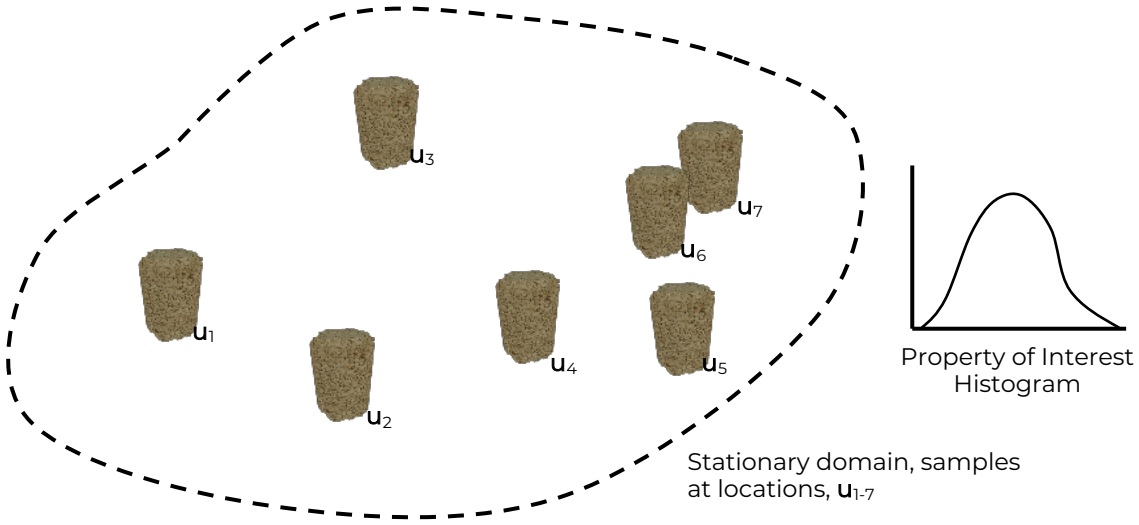
Spatial Data Analytics

Lecture outline . . .

- Stationarity
- Spatial Continuity
- Variogram Calculation
- Spatial Estimation

MOTIVATION

- The concepts of stationarity, spatial continuity and spatial estimation are central to spatial data analytics. We must integrate the spatial context.



STATIONARITY

STATIONARITY

Substituting Time for Space

- Any statistic requires replicates, repeated sampling (e.g. air or water samples from a monitoring station). In our geospatial problems repeated samples are not available at a location in the subsurface.

Extracted Sample at Location, u_1

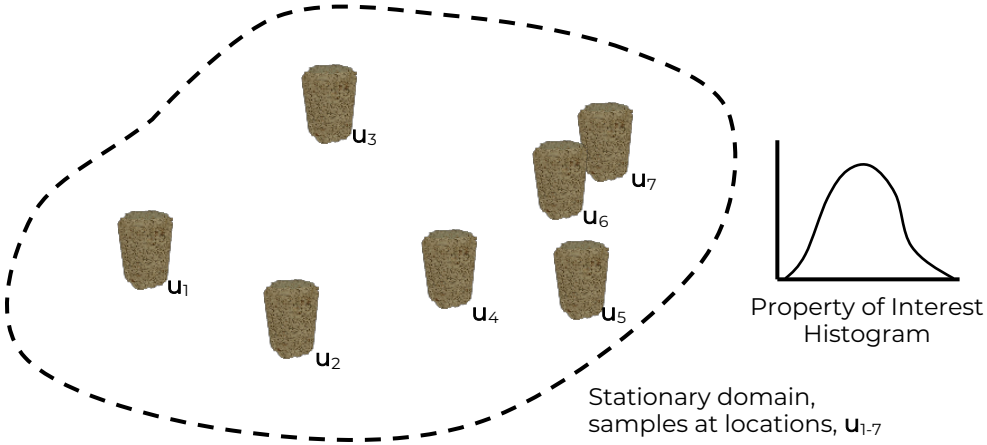
Hole Remains at Location, u_1

- Instead of time, **we must pool samples over space** to calculate our statistics. This decision to pool is the decision of stationarity. It is the decision that the subset of the subsurface is all the "same stuff".

STATIONARITY

Substituting Time for Space

- The decision of the stationary domain for sampling is an expert choice. Without it we are stuck in the “hole” and **cannot calculate any statistics** nor say anything about the behavior of the subsurface **between the sample data**.

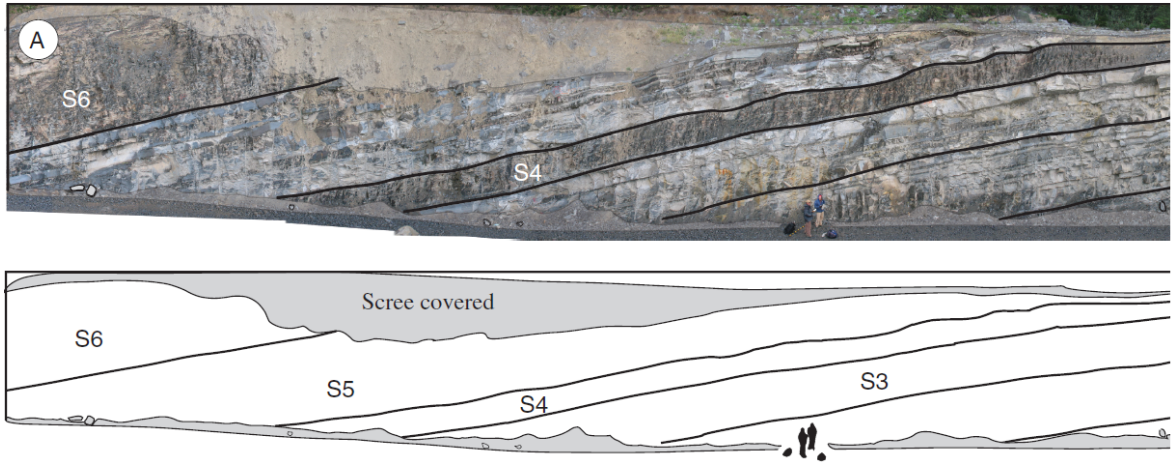


- **Import License**: choice to pool specific samples to evaluate a statistic.
- **Export License**: choice of where in the subsurface this statistic is applicable.

STATIONARITY

Definition 1: Geologic

- **Geological Definition**: e.g. 'The rock over the stationary domain is sourced, deposited, preserved, and postdepositionally altered in a similar manner, the domain is map-able and may be used for local prediction or as information for analogous locations within the subsurface; therefore, it is useful to pool information over this expert mapped volume of the subsurface.'



Photomosaic, line drawing Punta Barrosa Formation sheet complex (Fildani et al. (2009).

STATIONARITY

Definition 2: Statistical

- Statistical Definition: The metrics of interest are invariant under translation over the domain.

- For example, stationarity indicates the that histogram and associated statistics do not rely on location, $\mathbf{u}$. Statistical stationarity for some common statistics:

Stationary Mean: $E\{Z(\mathbf{u})\} = \text{m}, \ \forall \mathbf{u}$

Stationary Distribution: $F(\mathbf{u}; z) = F(z), \ \forall \mathbf{u}$

Stationary Semivariogram: $\gamma_z(\mathbf{u}; \mathbf{h}) = \gamma_z(\mathbf{h}), \ \forall \mathbf{u}$

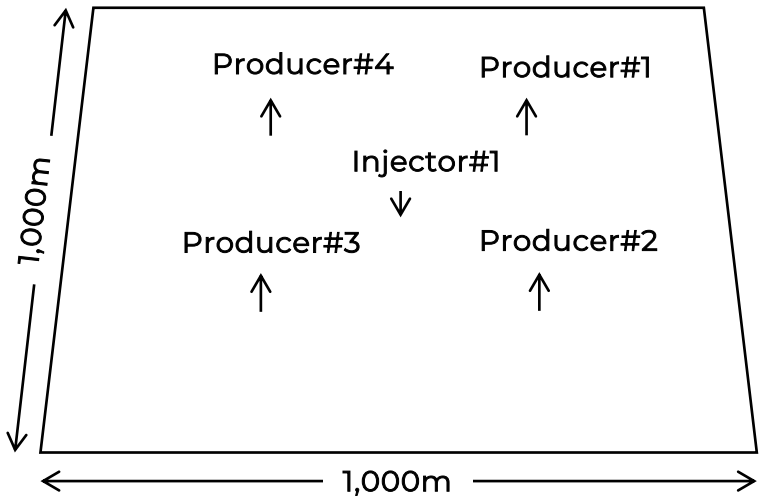
Stationarity: What metric / statistic? Over what volume?

- May be extended to any statistic of interest including, facies proportions, bivariate distributions and multiple point statistics.

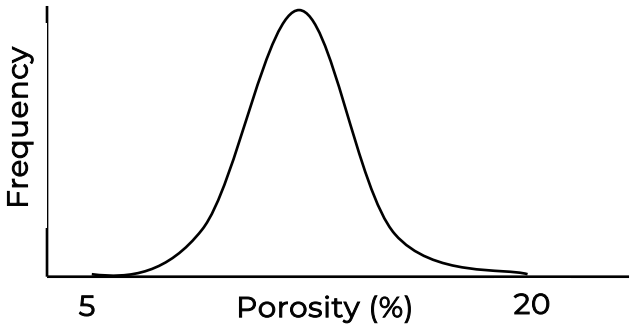
SPATIAL CONTINUITY

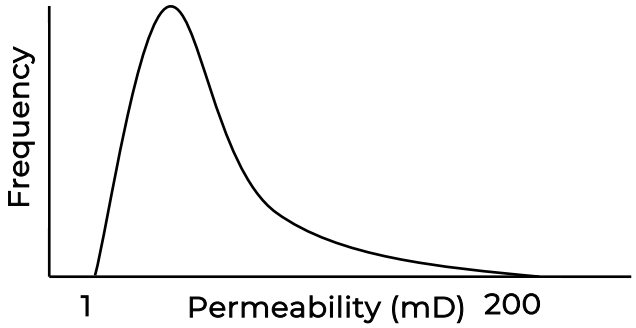
MOTIVATION FOR MEASURING SPATIAL CONTINUITY

- Simple Example
  - Area of interest
  - 1 Injector and 4 producers



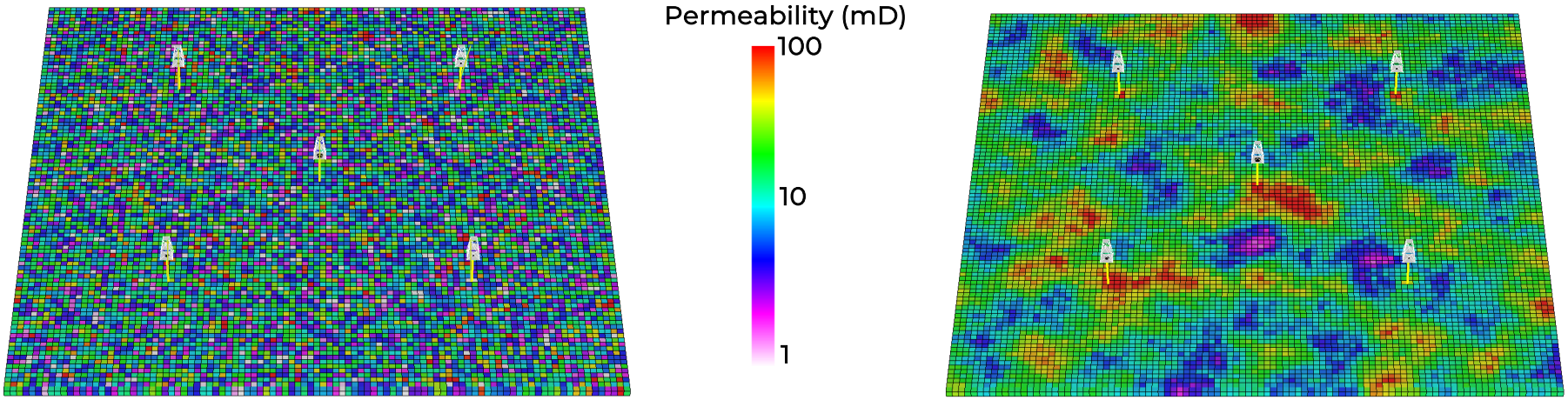
  - Porosity and permeability distributions (held constant for all cases)





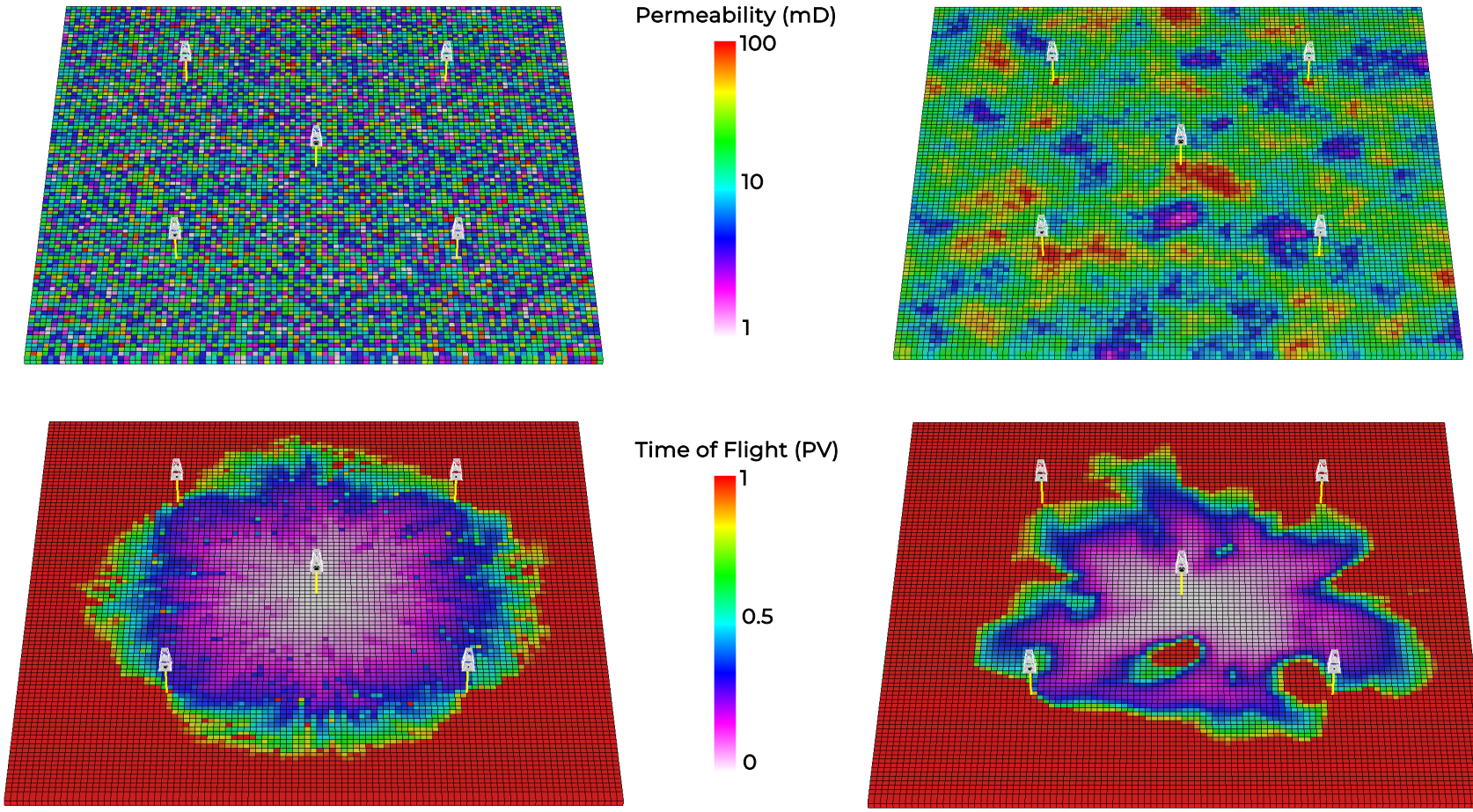
MOTIVATION FOR MEASURING SPATIAL CONTINUITY

- Does spatial continuity of reservoir properties matter?
- Consider these models of permeability



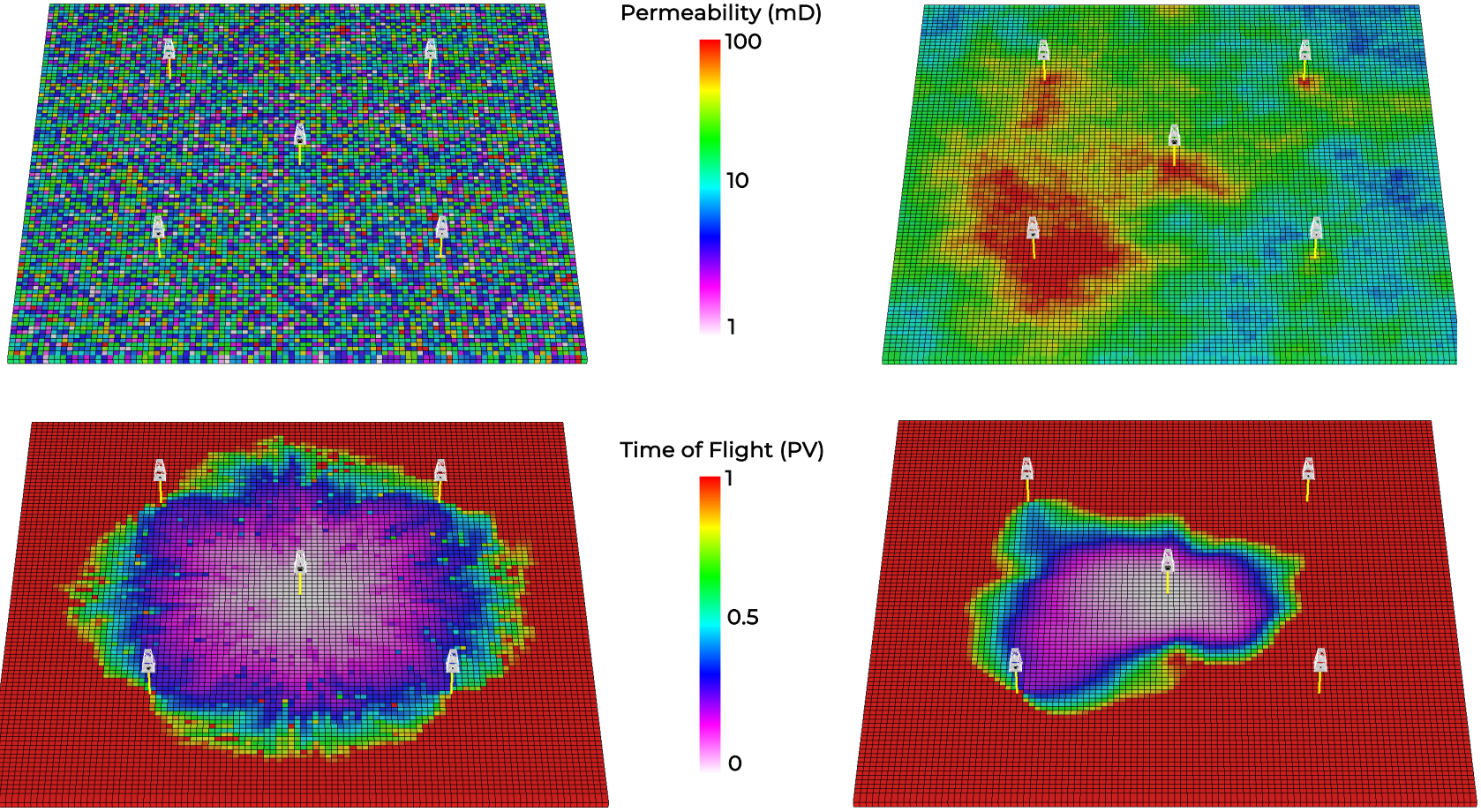
- Recall – all models have the same porosity and permeability distributions
  - Mean, variance, P10, P90 …
  - Same static oil in place!

MOTIVATION FOR MEASURING SPATIAL CONTINUITY



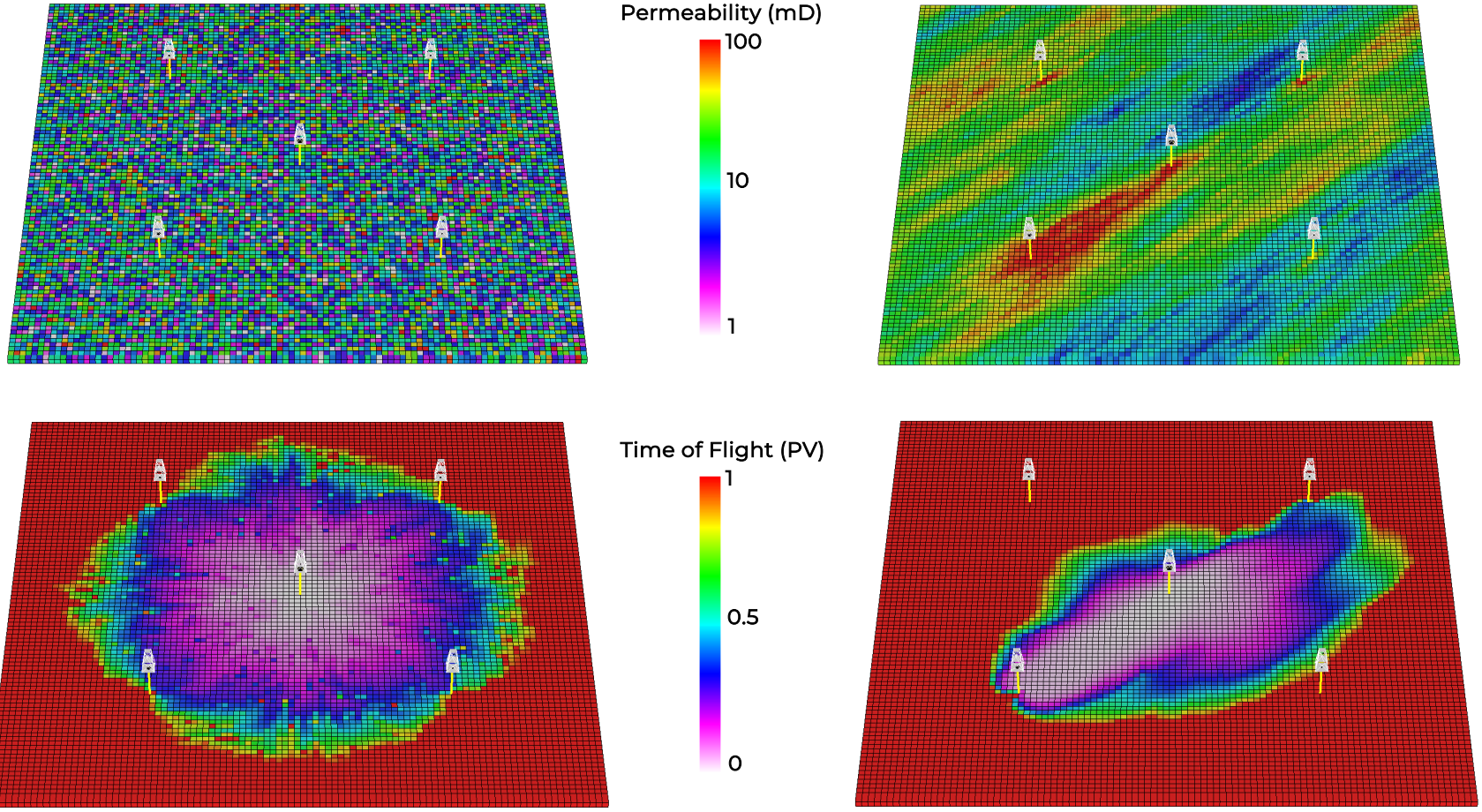
- How does heterogeneity impact recovery factor?

MOTIVATION FOR MEASURING SPATIAL CONTINUITY



- How does heterogeneity impact recovery factor?

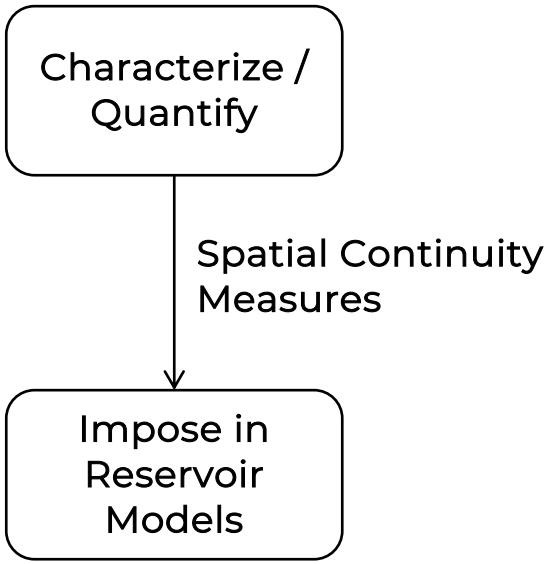
MOTIVATION FOR MEASURING SPATIAL CONTINUITY



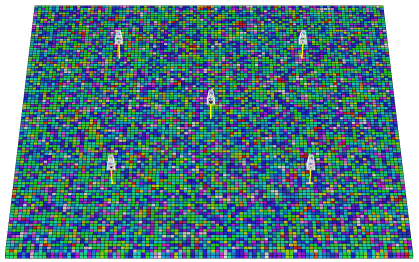
- How does heterogeneity impact recovery factor?

MOTIVATION FOR MEASURING SPATIAL CONTINUITY

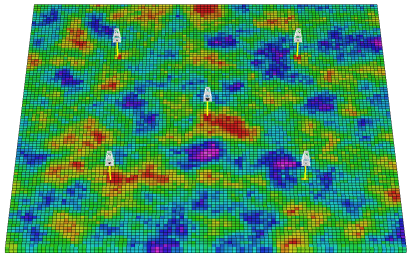
- For the same reservoir property distributions a wide range of spatial continuities are possible
- Spatial continuity often impacts reservoir forecasts
- Need to be able to:



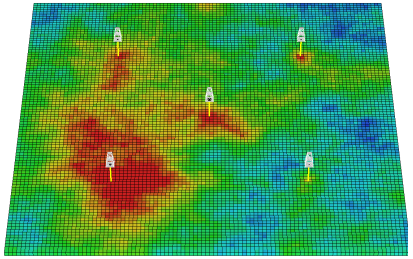
Spatial Continuity



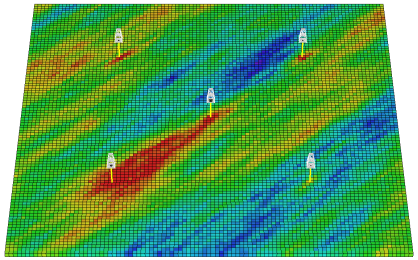
"Very Short"



"Medium"



"Long"



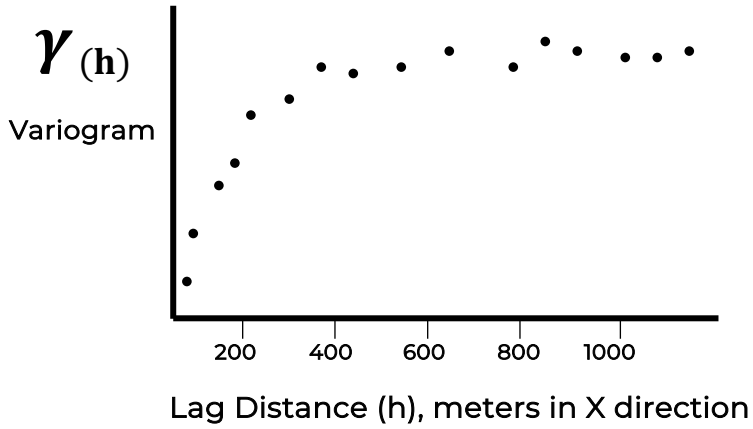
"Anisotropic "

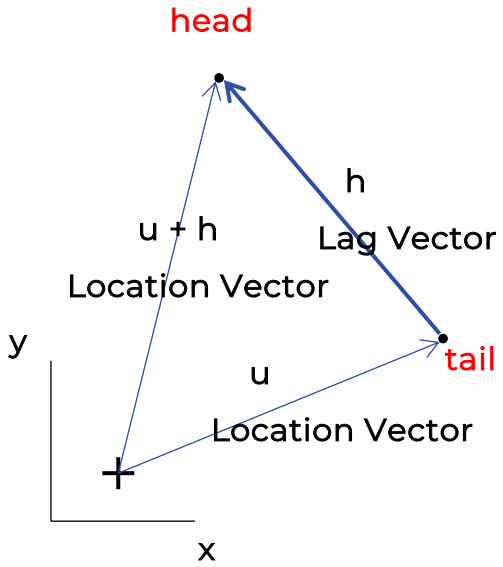
SPATIAL CONTINUITY DEFINITION

- Spatial Continuity – correlation between values over distance.
  - No spatial continuity – no correlation between values over distance, random values at each location in space regardless of separation distance.
  - Homogenous phenomenon have perfect spatial continuity, since all values as the same (or very similar) they are correlated.

MEASURING SPATIAL CONTINUITY

- We need a statistic to quantify spatial continuity!
- The Semivariogram:
  - Function of difference over distance





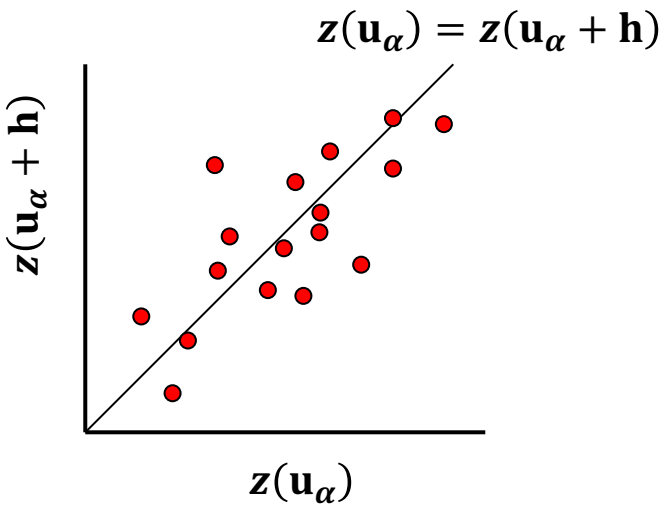
  - The equation:

$$\gamma(\mathbf{h}) = \frac{\mathbf{1}}{\mathbf{2}N(\mathbf{h})} \sum_{\alpha=1}^{N(\mathbf{h})} \left(z(\mathbf{u}_\alpha) - z(\mathbf{u}_\alpha + \mathbf{h})\right)^2$$

($z(\mathbf{u}_\alpha)$: tail; $z(\mathbf{u}_\alpha + \mathbf{h})$: head)

- One half the average squared difference over lag distance, **h**, over all possible pairs of data, $N(\mathbf{h})$

"H" SCATTERPLOT



- The variogram calculated for lag distance, $\mathbf{h}$, corresponds to the expected value of squared difference:

$$\gamma(\mathbf{h}) = \frac{1}{2N(\mathbf{h})} \sum_{\alpha=1}^{N(\mathbf{h})} \left(z(\mathbf{u}_\alpha) - z(\mathbf{u}_\alpha + \mathbf{h})\right)^2$$

- Calculate for a suite of lag distances to obtain a continuous function

VARIOGRAM DEFINITION

- Variogram – a measure of dissimilarity vs. distance. Calculated as ½ the average squared difference of values separated by a lag vector

$$\gamma(\mathbf{h}) = \frac{1}{2N(\mathbf{h})} \sum_{\alpha=1}^{N(\mathbf{h})} \left(z(\mathbf{u}_\alpha) - z(\mathbf{u}_\alpha + \mathbf{h}) \right)^2$$

  - The precise term is semivariogram (variogram if you remove the 1/2), but in practice the term variogram is used
  - The ½ is used so that the covariance function and variogram may be related directly:

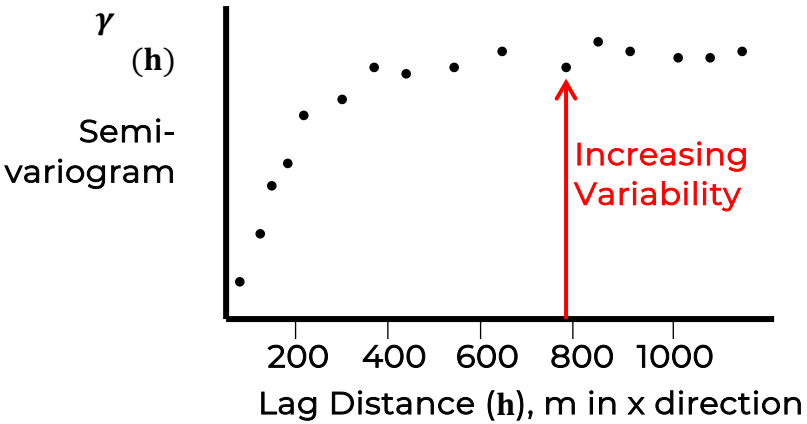
$$C_x(h) = \sigma_x^2 - \gamma_x(\mathbf{h})$$

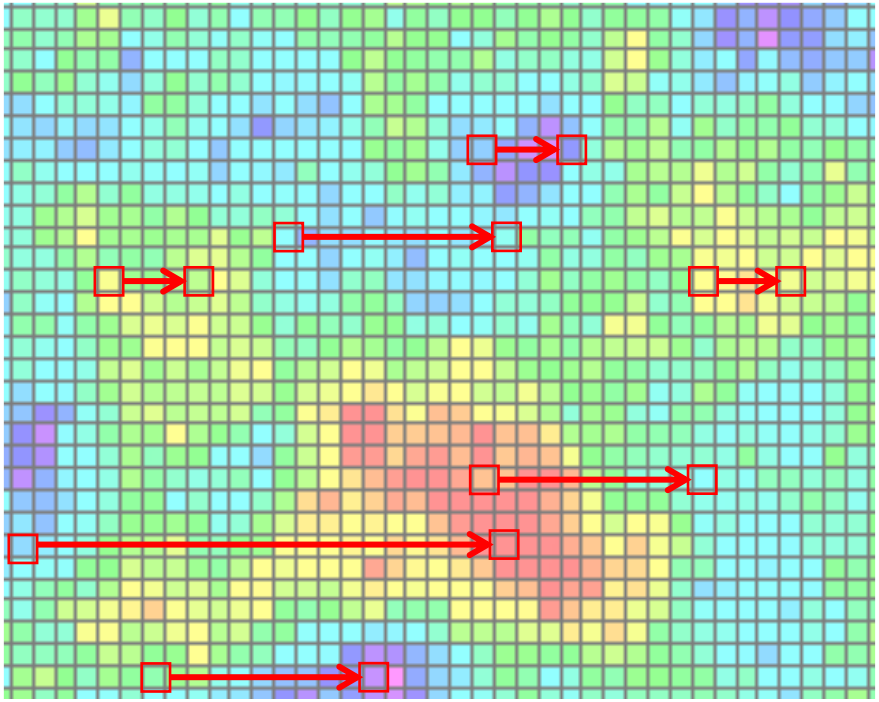
  - Note the correlogram is related to the covariance function as:

$\rho_x(h) = \frac{C_x(h)}{\sigma_x^2}$, h-scatter plot correlation vs. lag distance

VARIOGRAM OBSERVATIONS

- Observation #1
  - As distance increases, variability increase (in general)

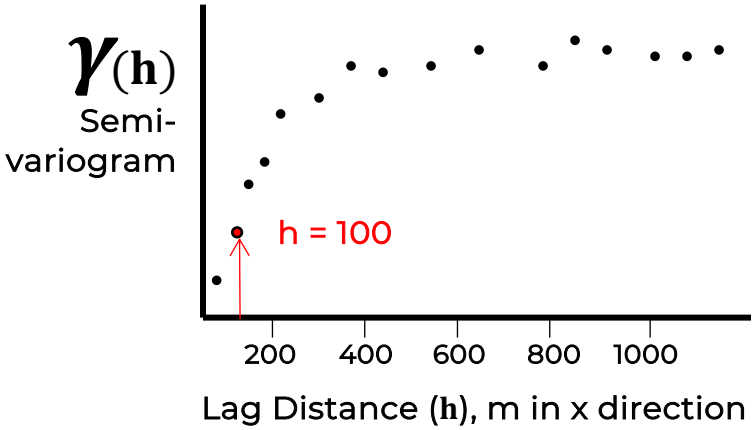


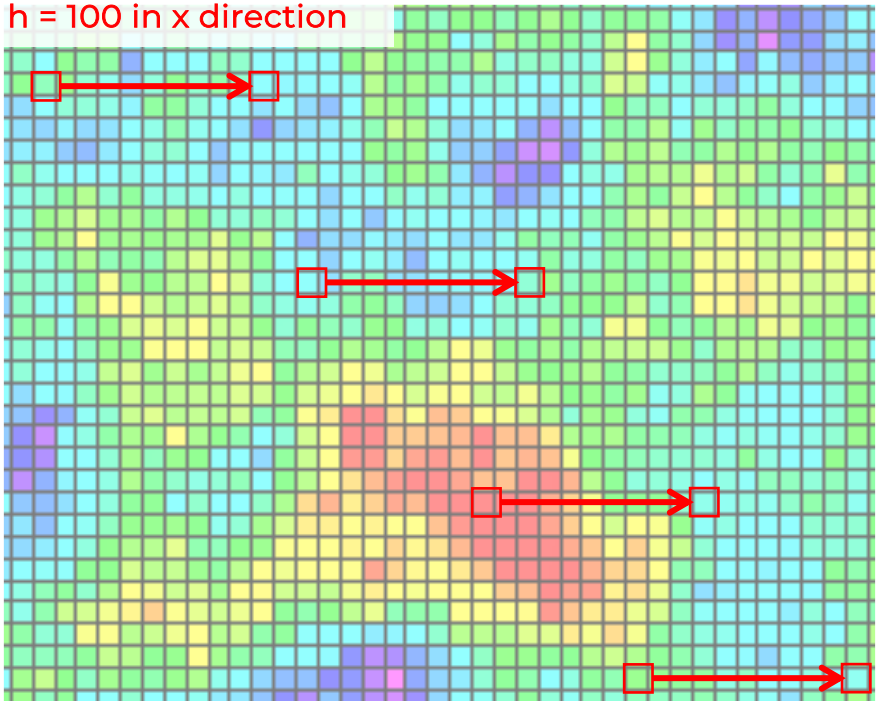


Increase in variability with increase in lag distance (zoomed in)

VARIOGRAM OBSERVATIONS

- Observation #2
  - Calculated with over all possible pairs separated by lag vector, $\mathbf{h}$



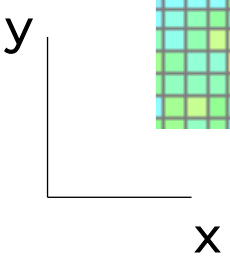


Scan of all possible sets of pairs (zoomed in)

- The variogram:

$$\gamma(\mathbf{h}) = \frac{1}{2N(\mathbf{h})} \sum_{\alpha=1}^{N(\mathbf{h})} \left(z(\mathbf{u}_\alpha) - z(\mathbf{u}_\alpha + \mathbf{h}) \right)^2$$

Given the number of pairs available N(h)

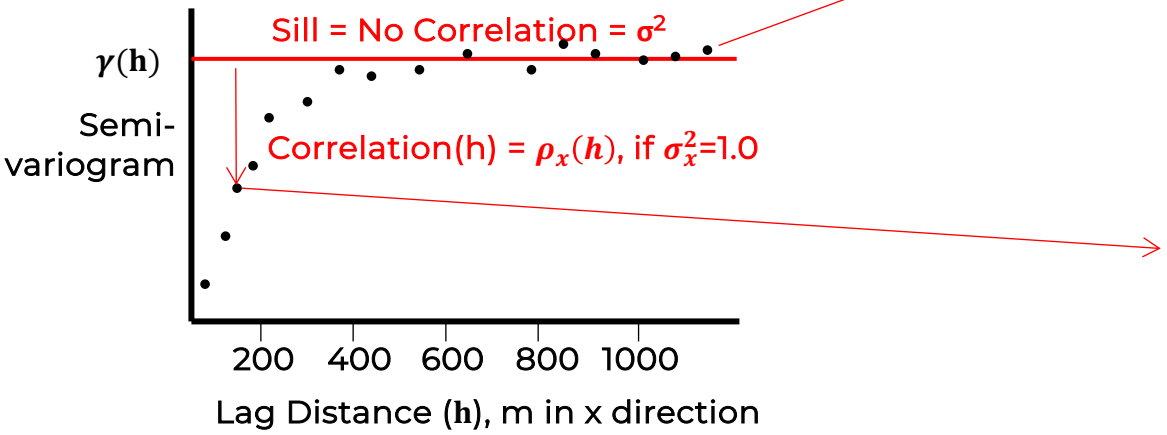


VARIOGRAM OBSERVATIONS

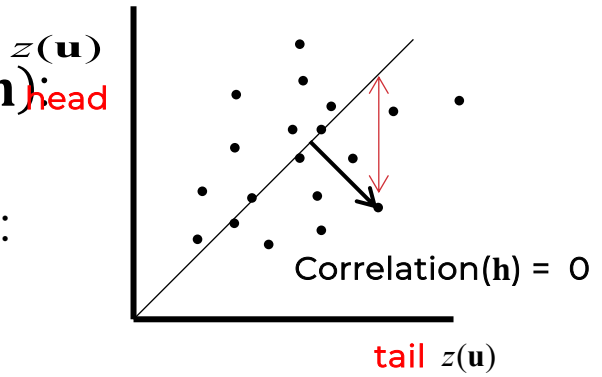
- Observation #3
  - Need to plot the sill to know the degree of correlation
  - Sill is the Variance, σ^2
  - Given stationarity of the variance and $\boldsymbol{\gamma_x(\mathbf{h})}$:

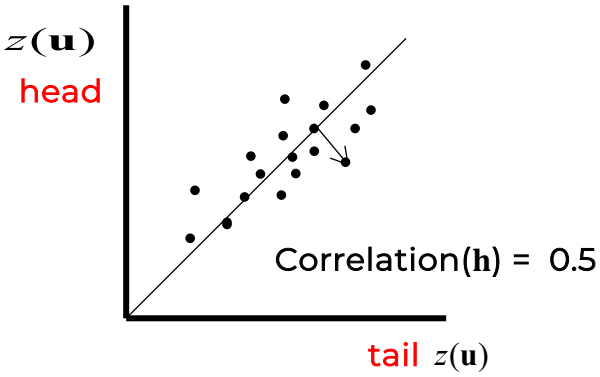
    Covariance Function: $\boldsymbol{C_x(h) = \sigma_x^2 - \gamma_x(\mathbf{h})}$
  - Given a standardized distribution $\boldsymbol{\sigma_x^2 = 1.0}$:

    Correlogram: $\boldsymbol{\rho_x(h) = \sigma_x^2 - \gamma_x(\mathbf{h})}$



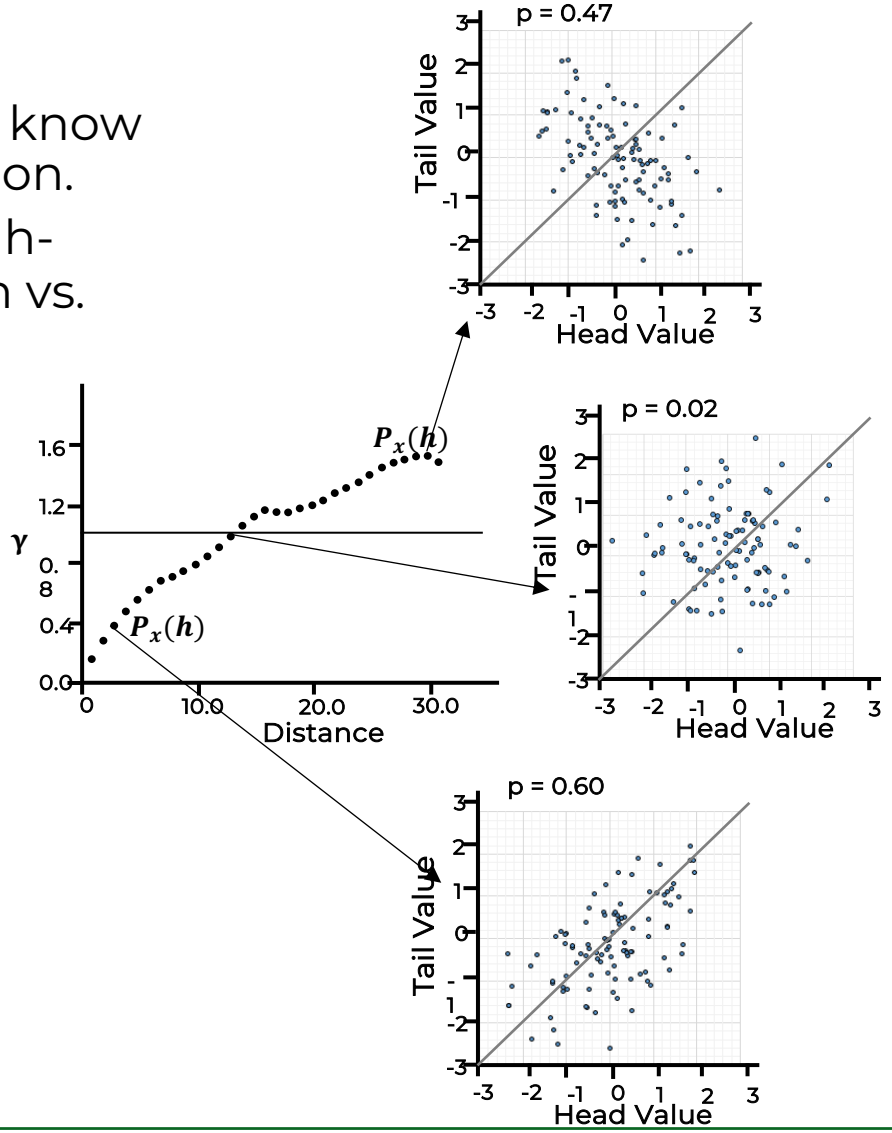
h – lag scatter plots





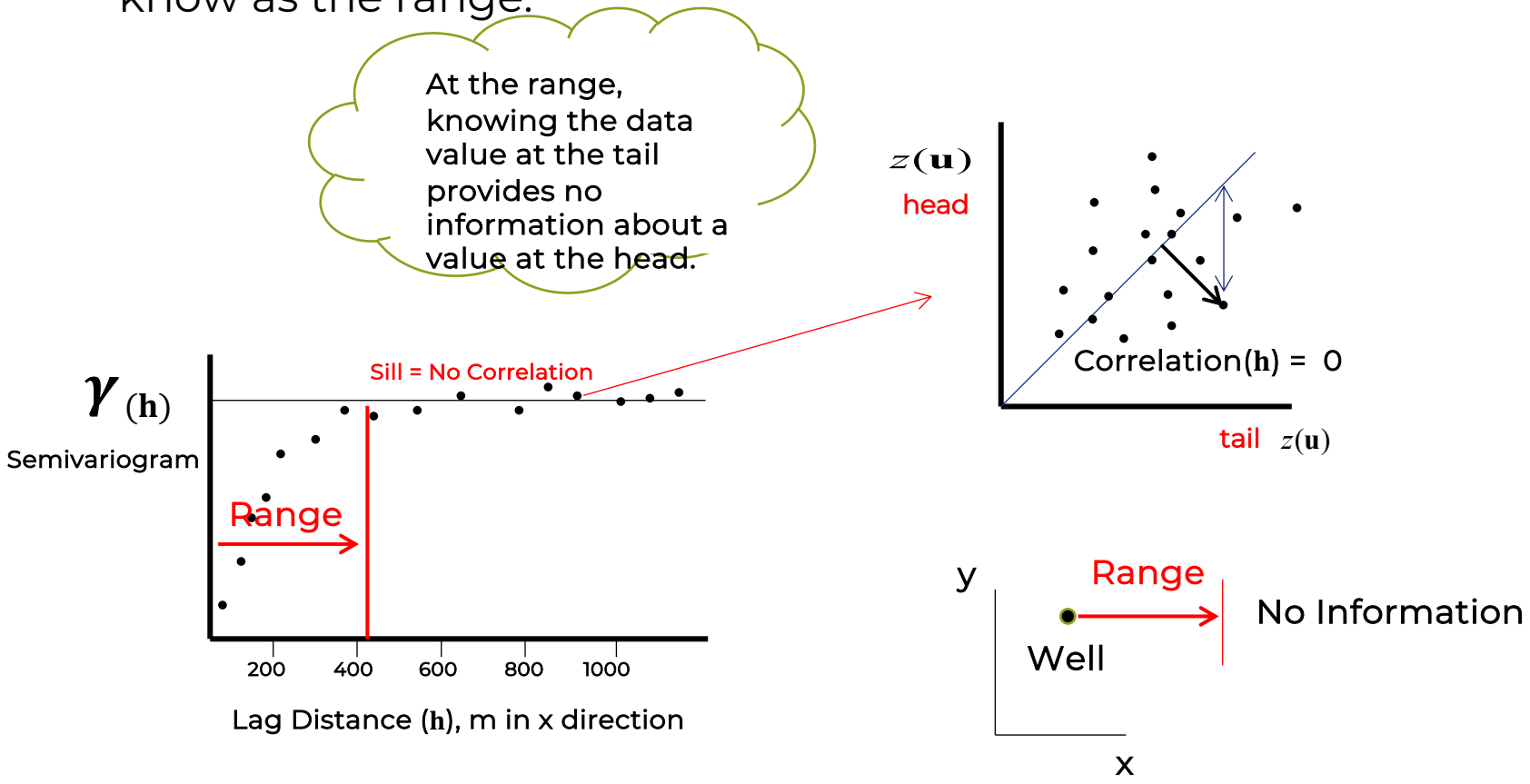
VARIOGRAM INTERPRETATION

- Observation #3
  - Need to plot the sill to know the degree of correlation.
  - Another illustration of h-scatter plot correlation vs. lag distance.



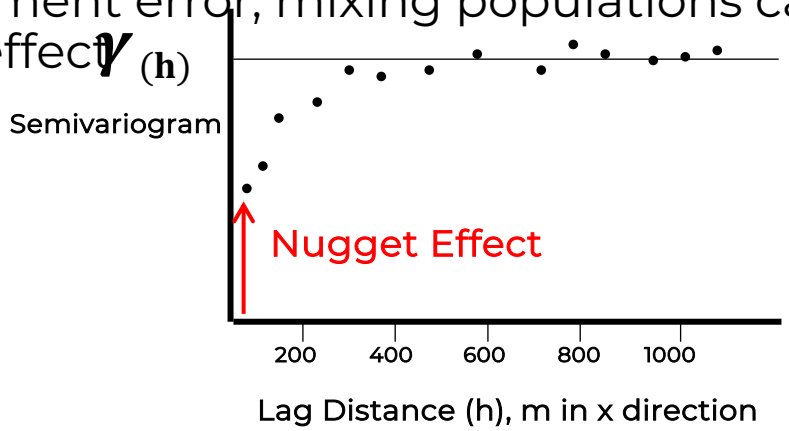
VARIOGRAM OBSERVATIONS

- Observation #4
  - The lag distance at which the variogram reaches the sill is know as the range.



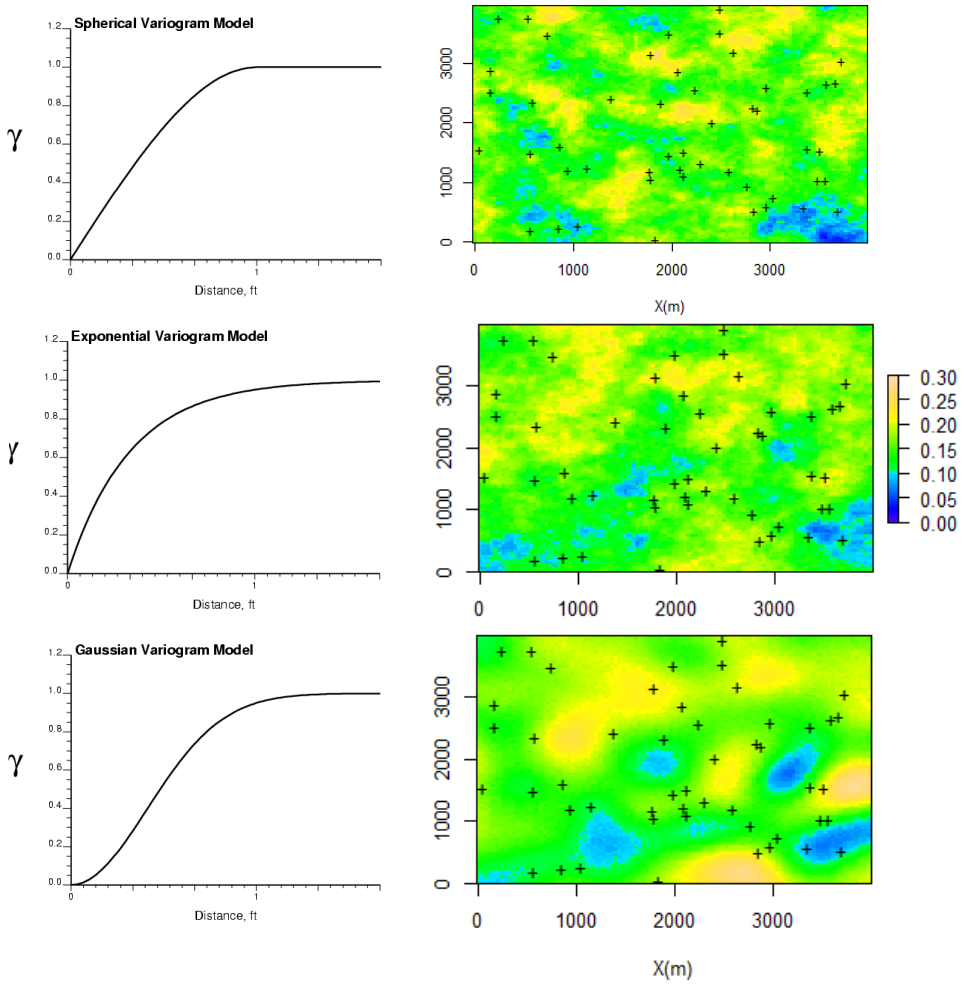
VARIOGRAM OBSERVATIONS

- Observation #5
  - Sometimes there is a discontinuity in the variogram at distances less than the minimum data spacing. This is known as nugget effect.
    - As a ratio of nugget / sill, is known as relative nugget effect (%)
    - Modeled as a no correlation structure that at lags, h>ε, an infinitesimal distance
    - Measurement error, mixing populations cause apparent nugget effect



SPATIAL VARIABILITY

- The three maps are remarkably similar: all three have the same 50 data, same histograms and same range of correlation, and yet their spatial variability/continuity is quite different
- The spatial variability/continuity depends on the detailed distribution of the petrophysical attribute (ϕ,K)
- The charts on the left are "variograms"
- Our map-making efforts should consider the spatial variability/continuity of the variable we are mapping:
  - Variability
  - Uncertainty



Porosity Realizations with 3 isotropic variograms

VARIOGRAM CALCULATION

VARIOGRAM DEFINITION

- **Variogram** – a measure of dissimilarity vs. distance. Calculated as ½ the average squared difference of values separated by a lag vector.

$$\gamma(\mathbf{h}) = \frac{1}{2N(\mathbf{h})} \sum_{\alpha=1}^{N(\mathbf{h})} \left(z(\mathbf{u}_\alpha) - z(\mathbf{u}_\alpha + \mathbf{h}) \right)^2$$

  - The precise term is semivariogram (variogram if you remove the 1/2), but in practice the term variogram is used
  - The ½ is used so that the covariance function and variogram may be related directly:

$$C_x(h) = \sigma_x^2 - \gamma_x(\mathbf{h})$$

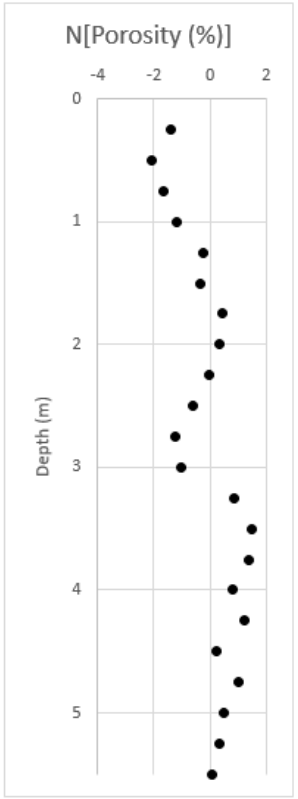
  - Note the correlogram is related to the covariance function as:

$\rho_x(h) = \frac{C_x(h)}{\sigma_x^2}$, h-scatter plot correlation vs. lag distance

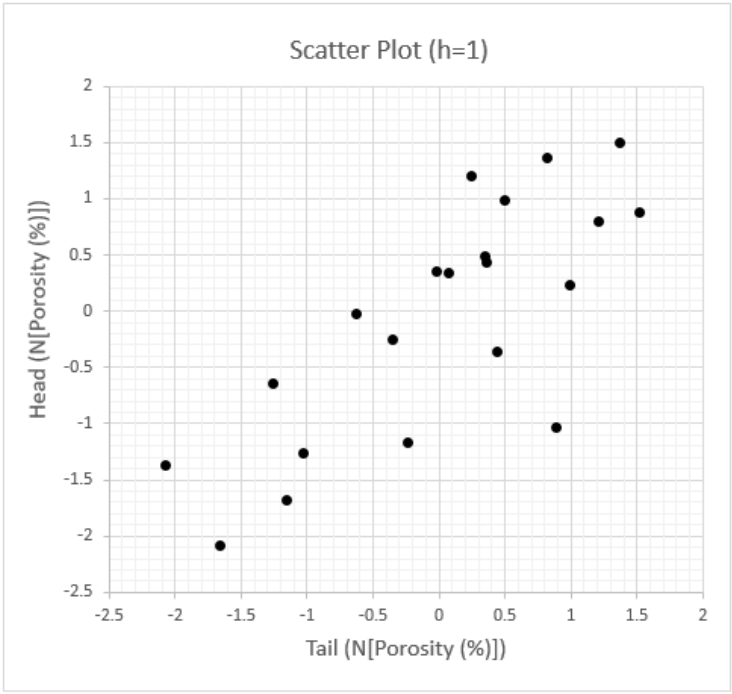
VARIOGRAM CALCULATION

- Consider data values separated by lag vectors (the h values)
- Here are two examples of a lag vector equal to the data spacing and then twice the data spacing:

Depth	N[Porosity]
0.25	-1.37
0.5	-2.08
0.75	-1.67
1	-1.16
1.25	-0.24
1.5	-0.36
1.75	0.44
2	0.36
2.25	-0.02
2.5	-0.63
2.75	-1.26
3	-1.03
3.25	0.88
3.5	1.51
3.75	1.37
4	0.81
4.25	1.21
4.5	0.24
4.75	0.99
5	0.49
5.25	0.34
5.5	0.07



Squared Difference		
	-1.37	
0.50	-2.08	-1.37
0.17	-1.67	-2.08
0.26	-1.16	-1.67
0.85	-0.24	-1.16
0.01	-0.36	-0.24
0.64	0.44	-0.36
0.01	0.36	0.44
0.14	-0.02	0.36
0.37	-0.63	-0.02
0.40	-1.26	-0.63
0.05	-1.03	-1.26
3.65	0.88	-1.03
0.40	1.51	0.88
0.02	1.37	1.51
0.31	0.81	1.37
0.16	1.21	0.81
0.94	0.24	1.21
0.56	0.99	0.24
0.25	0.49	0.99
0.02	0.34	0.49
0.07	0.07	0.34
Average / 2		0.07
0.23		

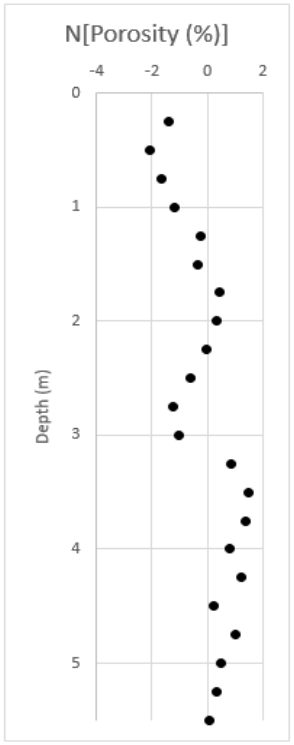


Correlation Coefficient (h=1)	0.77
Variogram (h=1)	0.23

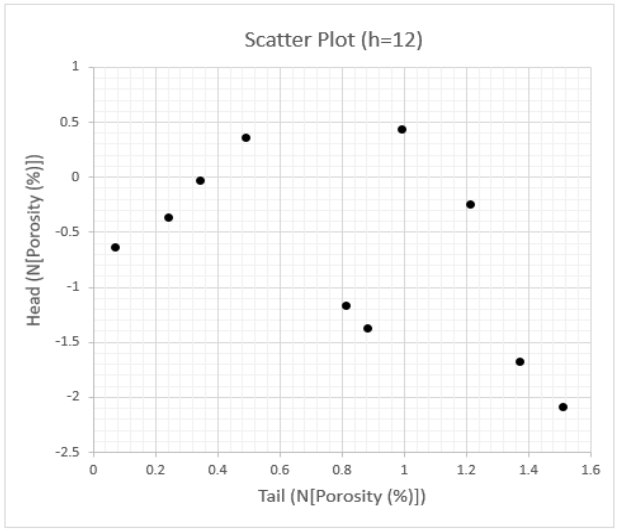
VARIOGRAM CALCULATION

- Consider data values separated by lag vectors (the h values)
- Here are two examples of a lag vector equal to the data spacing and then twice the data spacing:

Depth	N[Porosity]
0.25	-1.37
0.5	-2.08
0.75	-1.67
1	-1.16
1.25	-0.24
1.5	-0.36
1.75	0.44
2	0.36
2.25	-0.02
2.5	-0.63
2.75	-1.26
3	-1.03
3.25	0.88
3.5	1.51
3.75	1.37
4	0.81
4.25	1.21
4.5	0.24
4.75	0.99
5	0.49
5.25	0.34
5.5	0.07



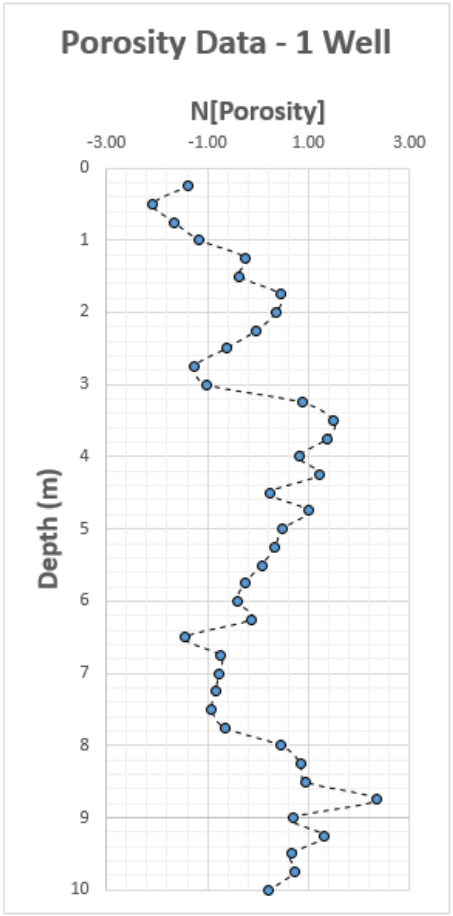
Squared Difference		
		-1.37
		-2.08
		-1.67
		-1.16
		-0.24
		-0.36
		0.44
		0.36
		-0.02
		-0.63
		-1.26
		-1.03
5.06	0.88	-1.37
12.89	1.51	-2.08
9.24	1.37	-1.67
3.88	0.81	-1.16
2.10	1.21	-0.24
0.36	0.24	-0.36
0.30	0.99	0.44
0.02	0.49	0.36
0.13	0.34	-0.02
0.49	0.07	-0.63
Average / 2		-1.26
1.72		-1.03
		0.88
		1.51
		1.37
		0.81
		1.21
		0.24
		0.99
		0.49
		0.34
		0.07



Correlation Coefficient (h=12)	-0.54
Variogram (h=12)	1.72

VARIOGRAM CALCULATION EXAMPLE

- e.g. we could calculate the variogram for this data set

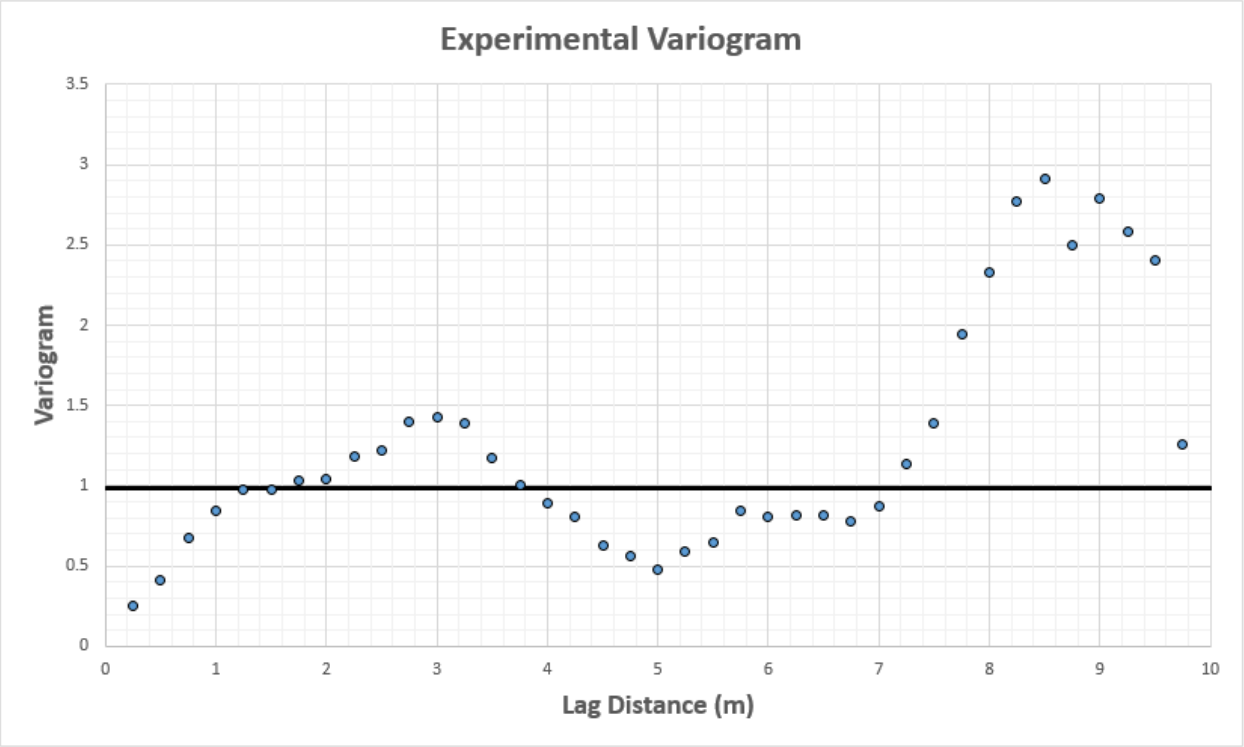


Depth	N[Porosity]
0.25	-1.37
0.5	-2.08
0.75	-1.67
1	-1.16
1.25	-0.24
1.5	-0.36
1.75	0.44
2	0.36
2.25	-0.02
2.5	-0.63
2.75	-1.26
3	-1.03
3.25	0.88
3.5	1.51
3.75	1.37
4	0.81
4.25	1.21
4.5	0.24
4.75	0.99
5	0.49
5.25	0.34
5.5	0.07
5.75	-0.26
6	-0.41
6.25	-0.14
6.5	-1.44
6.75	-0.75
7	-0.78
7.25	-0.85
7.5	-0.92
7.75	-0.66
8	0.47
8.25	0.85
8.5	0.95
8.75	2.35
9	0.69
9.25	1.31
9.5	0.66
9.75	0.72
10	0.21

VARIOGRAM CALCULATION EXAMPLE

- Could a lag distance and calculate the variogram for that one lag distance
- Here's all of them:

Depth	N[Porosity]
0.25	-1.37
0.5	-2.08
0.75	-1.67
1	-1.16
1.25	-0.24
1.5	-0.36
1.75	0.44
2	0.36
2.25	-0.02
2.5	-0.63
2.75	-1.26
3	-1.03
3.25	0.88
3.5	1.51
3.75	1.37
4	0.81
4.25	1.21
4.5	0.24
4.75	0.99
5	0.49
5.25	0.34
5.5	0.07
5.75	-0.26
6	-0.41
6.25	-0.14
6.5	-1.44
6.75	-0.75
7	-0.78
7.25	-0.85
7.5	-0.92
7.75	-0.66
8	0.47
8.25	0.85
8.5	0.95
8.75	2.35
9	0.69
9.25	1.31
9.5	0.66
9.75	0.72
10	0.21



THE VARIOGRAM AND COVARIANCE FUNCTION

- The variogram, $\boldsymbol{\gamma_x}$, covariance function, $\boldsymbol{C_x}$, and correlation coefficient, $\boldsymbol{\rho_x}$, are equivalent tools for characterizing spatial two-point correlation (assuming stationarity):

$$\boldsymbol{\gamma_x(\mathbf{h}) = \sigma_x^2 - C_x(h)} \\ \boldsymbol{= \sigma_x^2\ (1 - \rho_x(h))}$$

$$\boldsymbol{\rho_x(h) = \frac{C_x(h)}{\sigma_x^2}}$$

$$C_x(0) = \sigma_x^2$$

- where:

$$\boldsymbol{C_x(\mathbf{h}) = E\{X(\mathbf{u}) \cdot X(\mathbf{u}+\mathbf{h})\} - [E\{X(\mathbf{u})\} \cdot E\{X(\mathbf{u}+\mathbf{h})\}], \forall\, \mathbf{u}, \mathbf{u}+\mathbf{h} \in A}$$

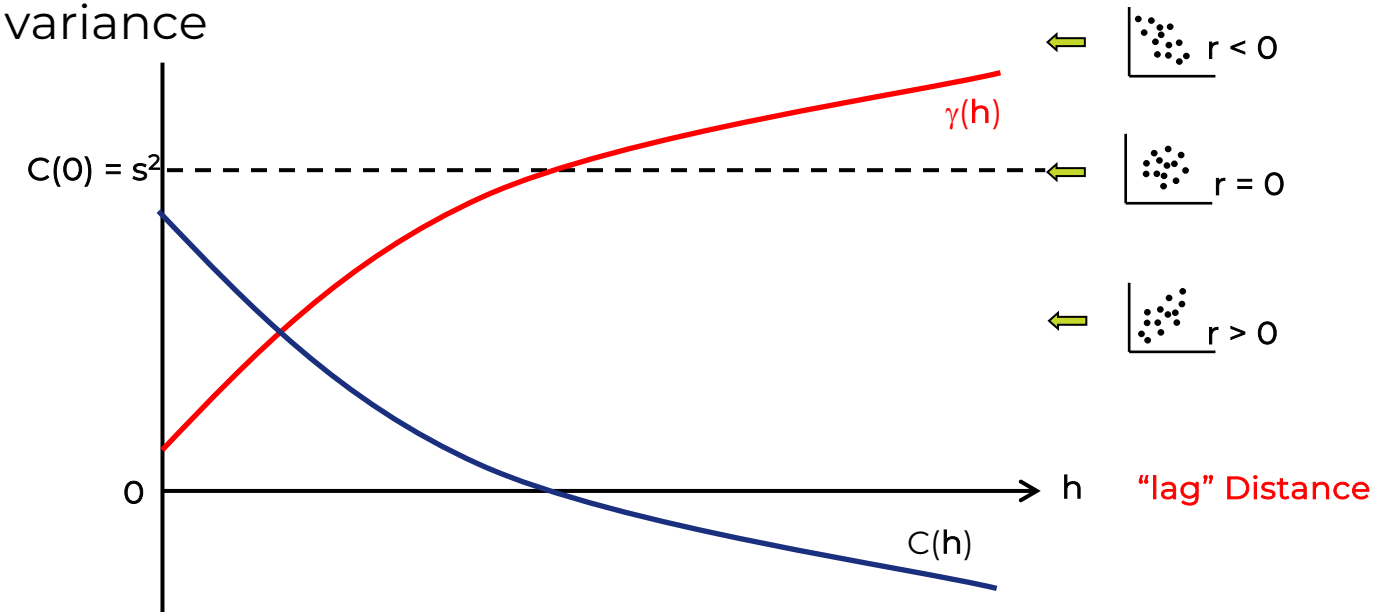
$\boldsymbol{C_x(\mathbf{h}) = \frac{\sum_{\alpha=1}^{n} x(\mathbf{u}_\alpha) \cdot x(\mathbf{u}_\alpha+\mathbf{h})}{n} - (\overline{x})^2}$, if stationary mean

- Stationarity entails that:

$$m(\mathbf{u}) = m(\mathbf{u}+\mathbf{h}) = m = E\{Z\}, \forall \mathbf{u} \in A$$
$$Var(\mathbf{u}) = Var(\mathbf{u}+\mathbf{h}) = \sigma^2 = Var\{Z\}, \forall \mathbf{u} \in A$$

THE VARIOGRAM AND COVARIANCE FUNCTION

- Must plot variance to interpret variogram:
  - Positive correlation when semivariogram less than variance
  - No correlation when the semivariogram is equal to the variance
  - Negative correlation when the semivariogram points above variance



COVARIANCE FUNCTION DEFINITION

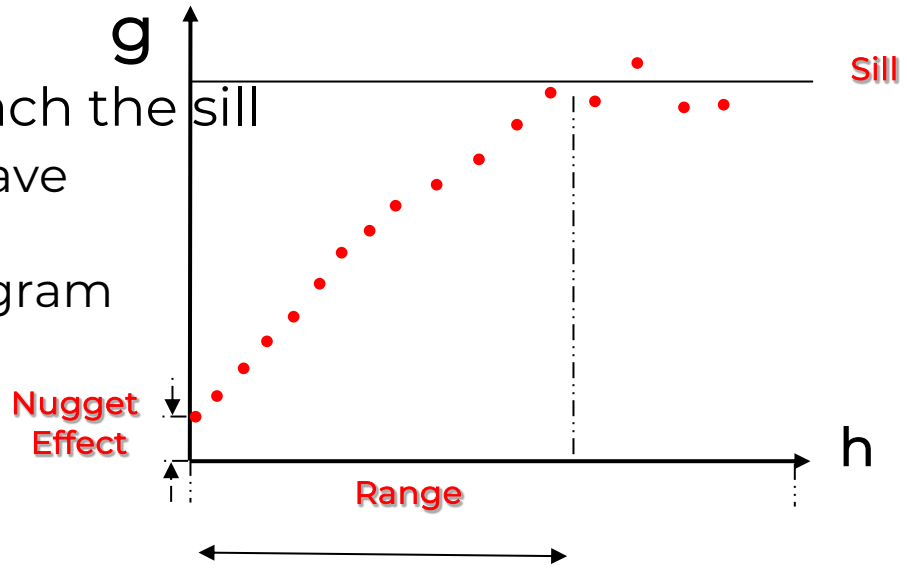
- Covariance Function – a measure of similarity vs. distance. Calculated as the average product of values separated by a lag vector centered by the square of the mean.

$$\boldsymbol{C_x(\mathbf{h}) = \frac{\sum_{\alpha=1}^{n} x(\mathbf{u}_\alpha)\cdot x(\mathbf{u}_\alpha+\mathbf{h})}{n} - (\overline{x})^2}, \text{ if stationary mean}$$

  - The covariance function is the variogram upside down. $\boldsymbol{\gamma_x(\mathbf{h}) = \sigma_x^2 - C_x(h)}$
  - We model variograms, but inside the kriging and simulation methods they are converted to covariance values for numerical convenience.

VARIOGRAM COMPONENTS DEFINITION

- Nugget Effect – discontinuity in the variogram at distances less than the minimum data spacing
  - As a ratio of nugget / sill, is known as relative nugget effect (%)
  - Measurement error, mixing populations cause apparent nugget effect
- Sill – the sample variance
  - Interpret spatial correlation relative to the sill, level of no correlation
- Range – lag distance to reach the sill
  - Up to that distance you have information
  - parameterization of variogram models



SPOILER ALERT

- We need to practically calculate and model spatial continuity. From the available and often sparse subsurface data.

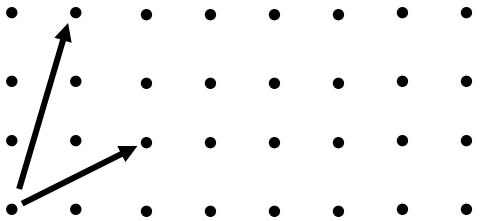
1. Calculate variogram with irregularly spaced data
   - Search templates with parameters
2. Valid spatial model
   - Fit with a couple different, nest (additive) spatial continuity models e.g. nugget, spherical, exponential and Gaussian
3. Full 3D spatial continuity model
   - Model primary directions, i.e. major horizontal, minor horizontal and vertical and combine together with assumption of geometric anisotropy

- We will use this model for spatial estimation and simulation.

Note: We will not cover variogram modeling.

CALCULATING EXPERIMENTAL VARIOGRAMS

- How do we get pairs separated by lag vector?
- Regular spaced data:
  - Specify as offsets of grid units
  - Fast calculation
  - Diagonal directions are awkward
- Irregular spaced data:
  - Nominal distance for each lag
  - Distance tolerance
  - Azimuth direction and tolerance
  - Dip direction and tolerance
  - Bandwidth (maximum deviation) in originally horizontal plane
  - Bandwidth in originally vertical plane



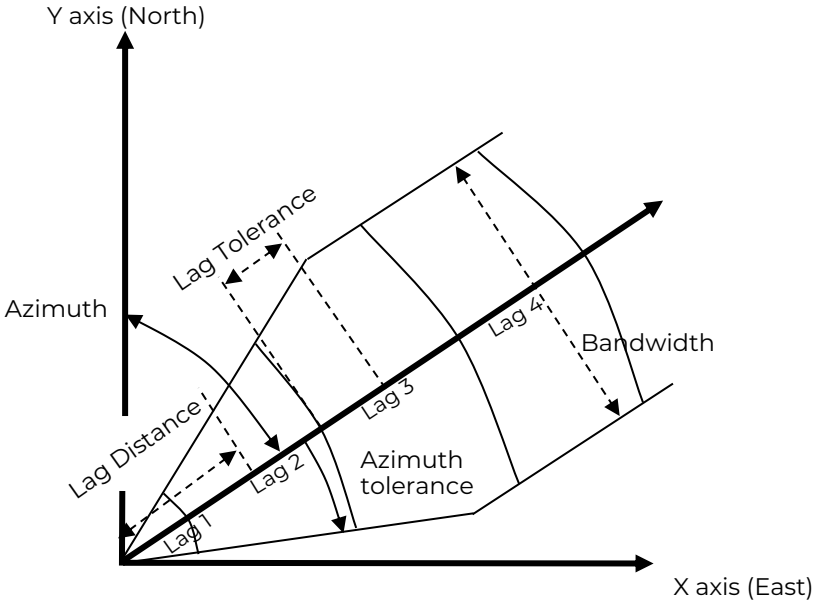


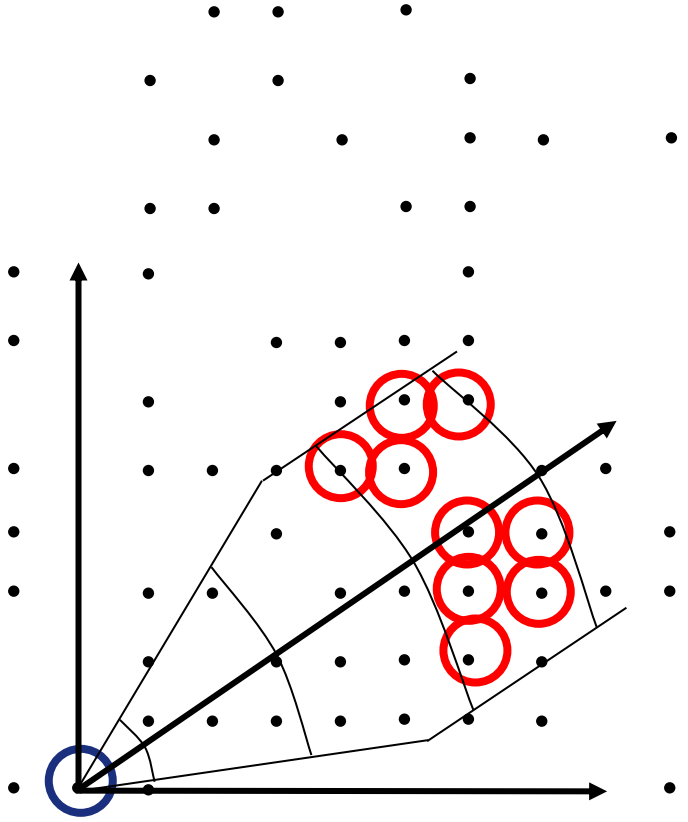
Image from Pyrcz and Deutsch, 2014

CALCULATING EXPERIMENTAL VARIOGRAMS

- Example: Starting With One Lag (i.e. #4)

$$2\gamma(h) = \frac{1}{N(h)} \sum_{N(h)} [z(u) - z(u+h)]^2$$

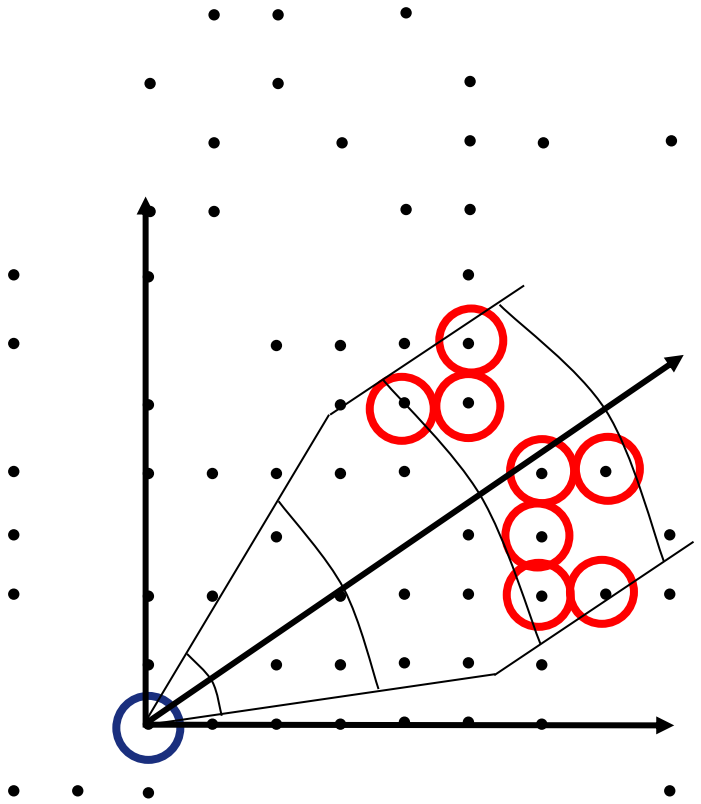
- Start at a node, and compare value to all nodes which fall in the lag and angle tolerance



CALCULATING EXPERIMENTAL VARIOGRAMS

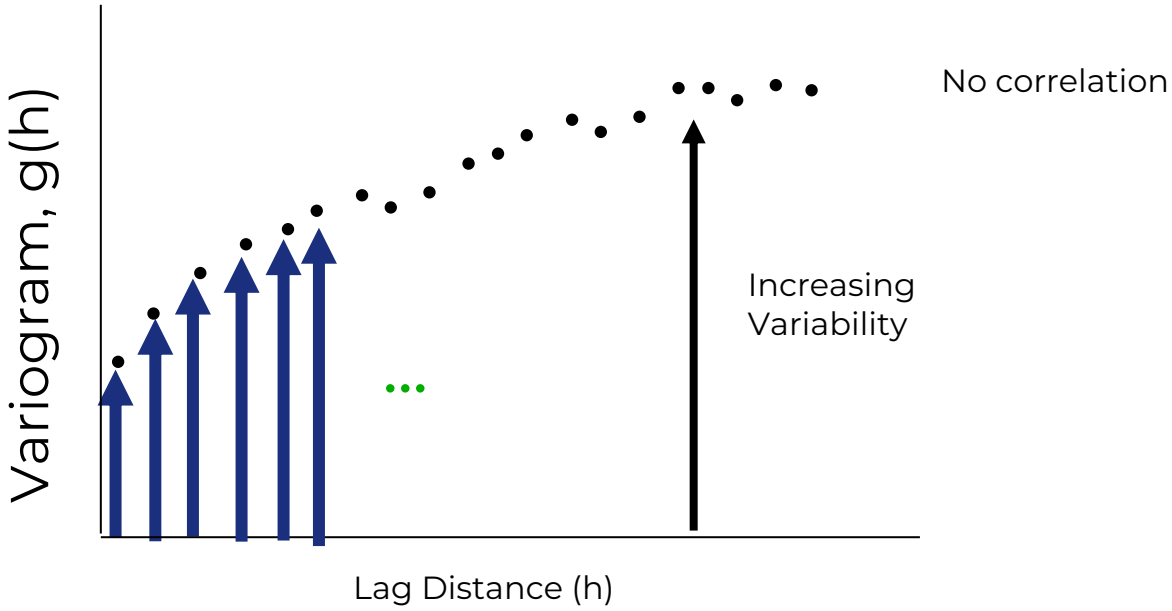
$$2\gamma(h) = \frac{1}{N(h)} \sum_{N(h)} [z(u) - z(u+h)]^2$$

- Move to next node



CALCULATING EXPERIMENTAL VARIOGRAMS

- Repeat for all nodes
- Repeat for all lag distances



CHOOSING THE DIRECTIONS

- Inspect the data and interpretations, sections, plan views, …
- Azimuth angles in degrees clockwise from north
- Review multiple directions before choosing a set of 3 perpendicular directions
  - Omnidirectional: all directions taken together → often yields the most well-behaved variograms.
  - Major horizontal direction & two perpendicular to major direction
  - All anisotropy in geostatistics is geometric – three mutually orthogonal directions with ellipsoidal change in the other directions:

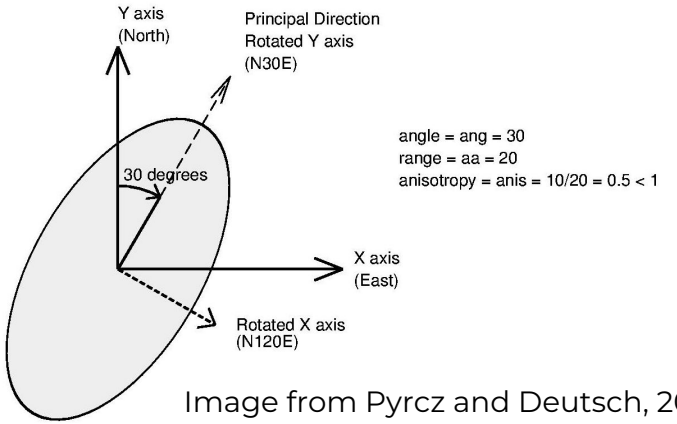


Image from Pyrcz and Deutsch, 2014

CHOOSING THE LAG DISTANCES AND TOLERANCES

- Guidance for Variogram Calculation Parameters:
- Lag separation distance should coincide with data spacing
- Lag tolerance typically chosen to be ½ lag separation distance
  - in cases of erratic variograms, may choose to overlap calculations so lag tolerance > ½ lag separation, results in more pairs.
- The variogram is only valid for a distance one half of the field size: start leaving data out of calculations with larger distances

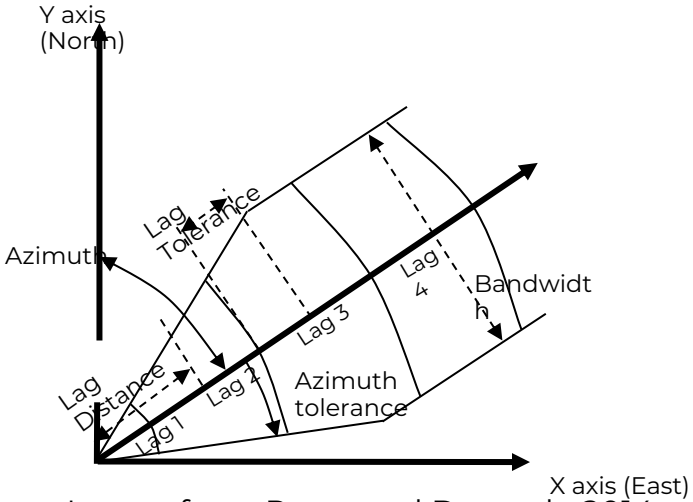


Image from Pyrcz and Deutsch, 2014

SPATIAL CALCULATION IN DEMO IN PYTHON

- Experiment with Variogram Calculation:
- Things to Try:
- Variogram maps
  - Relate to the data locatior maps
- Directional variograms
  - Change lag tolerance
  - Change lag distance
- Python notebook file: GeostatsPy_spatial_continuity_directions.ipynb

Daytum +2 Course: Data Analytics, Geostatistics and Machine Learning Deep Dive

Spatial Continuity Calculation Demonstration and Exercise ¶

Goal

Calculate spatial continuity for a spatial dataset.

Description

Here's a simple, documented workflow, demonstration of spatial continuity calculation for subsurface modeling workflows. This should help you get started with building subsurface models that integrate spatial continuity.

Spatial Continuity

Spatial Continuity is the correlation between values over distance.

- No spatial continuity – no correlation between values over distance, random values at each location in space regardless of separation distance.
- Homogenous phenomenon have perfect spatial continuity, since all values as the same (or very similar) they are correlated.

We need a statistic to quantify spatial continuity! A convenient method is the Semivariogram.

The Semivariogram

Function of difference over distance.

- The expected (average) squared difference between values separated by a lag distance vector (distance and direction), h:

$$\gamma(\mathbf{h}) = \frac{1}{2\mathrm{N}(\mathbf{h})} \sum_{\alpha=1}^{\mathrm{N}(\mathbf{h})} (z(\mathbf{u}_\alpha) - z(\mathbf{u}_\alpha + \mathbf{h}))^2$$

where $z(\mathbf{u}_\alpha)$ and $z(\mathbf{u}_\alpha + \mathbf{h})$ are the spatial sample values at tail and head locations of the lag vector respectively.

- Calculated over a suite of lag distances to obtain a continuous function.
- the $\frac{1}{2}$ term converts a variogram into a semivariogram, but in practice the term variogram is used instead of semivariogram.
- We prefer the semivariogram because it relates directly to the covariance function, $C_x(\mathbf{h})$ and univariate variance, σ_x^2:

$$C_x(\mathbf{h}) = \sigma_x^2 - \gamma(\mathbf{h})$$

Note the correlogram is related to the covariance function as:

$$\rho_x(\mathbf{h}) = \frac{\mathrm{C_x}(\mathbf{h})}{\sigma_\mathrm{x}^2}$$

The correlogram provides of function of the $\mathbf{h} - \mathbf{h}$ scatter plot correlation vs. lag offset $\mathbf{h}$.

$$-1.0 \leq \rho_x(\mathbf{h}) \leq 1.0$$

Variogram Observations

SPATIAL CONTINUITY

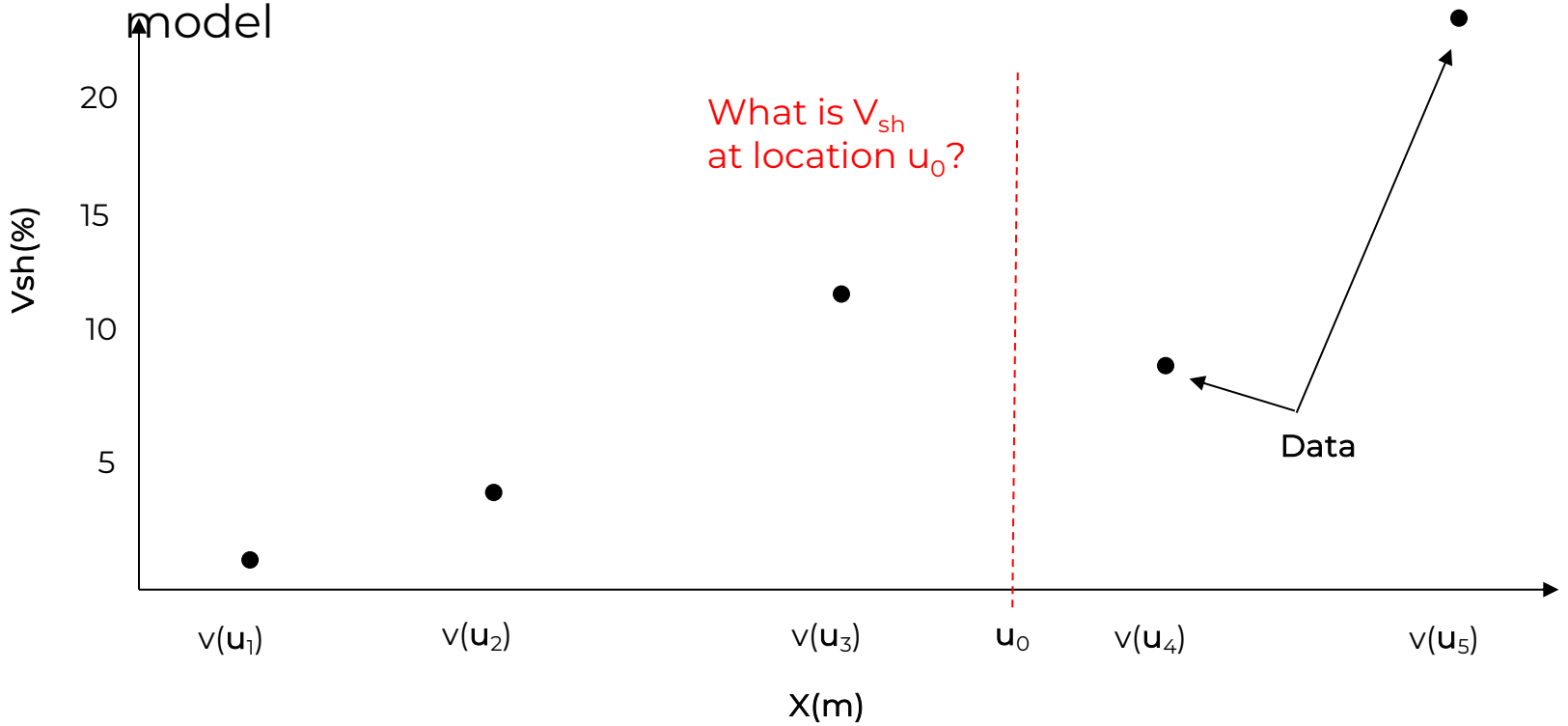
New Tools

Topic	Application to Subsurface Modeling
Stationarity	In the presence of multivariate relationships, must jointly model variables. *Summarize with bivariate statistics, and visualize and use conditional statistics to go beyond linear measures.*
Random Variables and Functions	Random variables and random functions are used to represent spatial uncertainty. *Porosity at a pre-drill location has the uncertainty model based on a random variable with Gaussian mean of 15% and standard deviation of 3%.*
Variogram Calculation	*Calculate spatial continuity from spatial data to use for spatial prediction.* *From the available wells the porosity spatial continuity range is 300 m in the 030 azimuth, we have no information beyond this spacing from existing wells.*

SPATIAL ESTIMATION

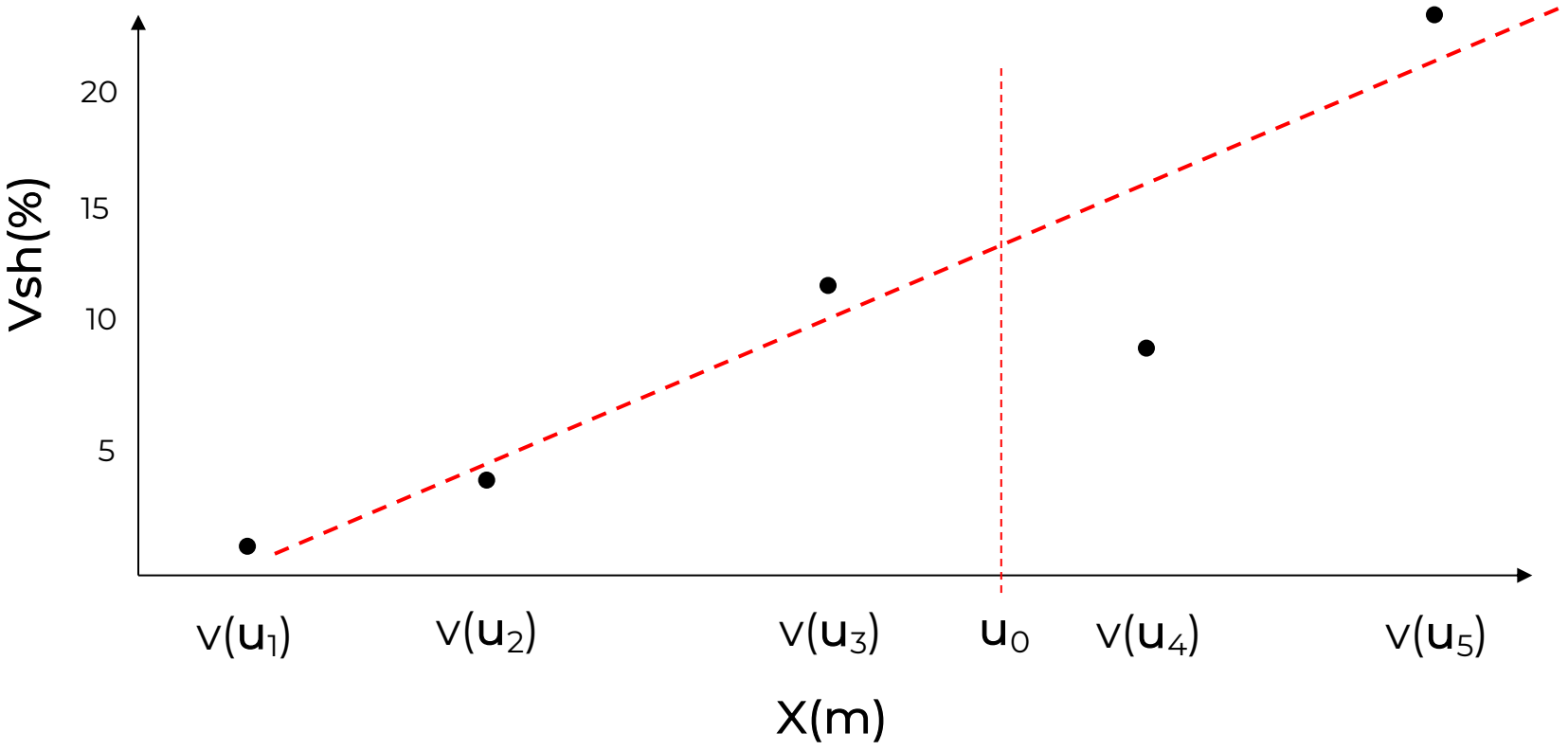
TREND MODELING

- We Must Start with Trend Modeling
- Geostatistical spatial estimation methods make an assumption concerning stationarity
  - In the presence of significant non-stationarity we would not rely 100% for spatial estimation on data + spatial continuity model



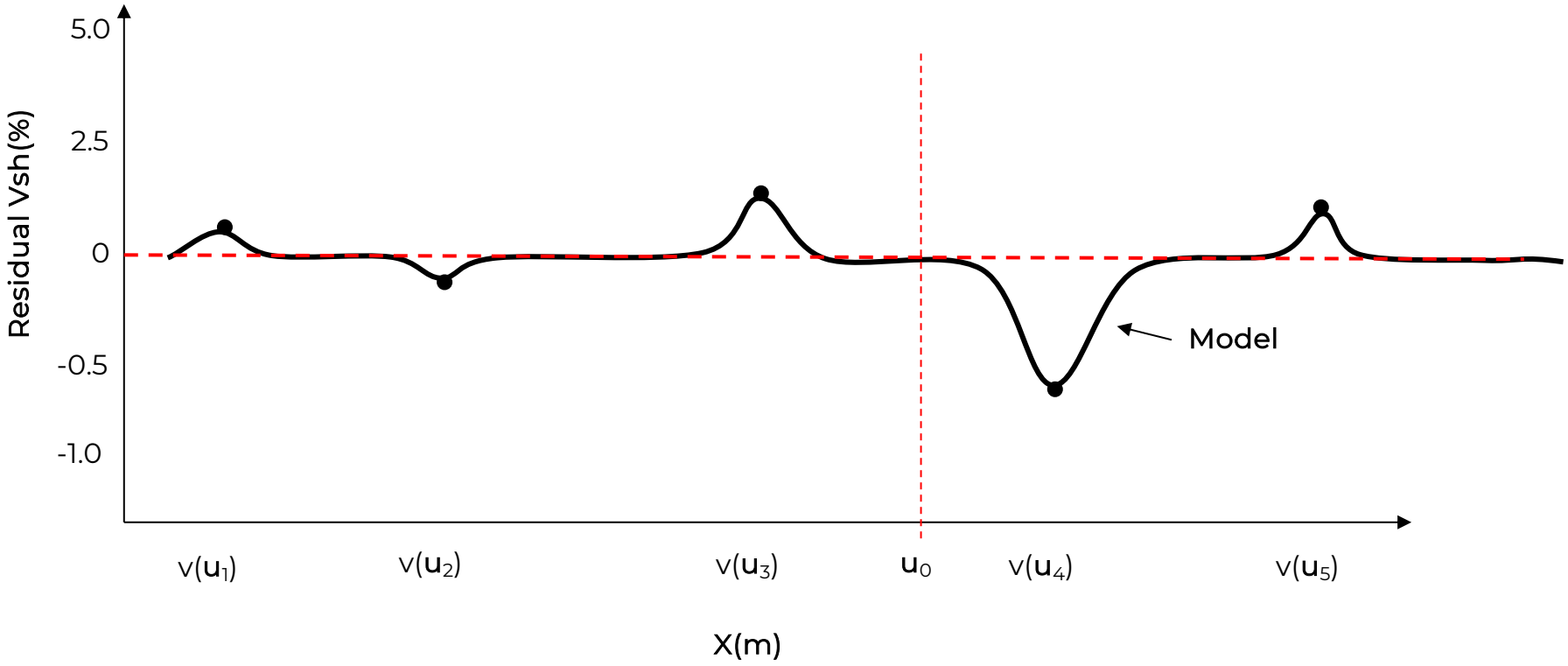
TREND MODELING

- Geostatistical spatial estimation methods will make an assumption concerning stationarity
  - If we observe a trend, we should model the trend



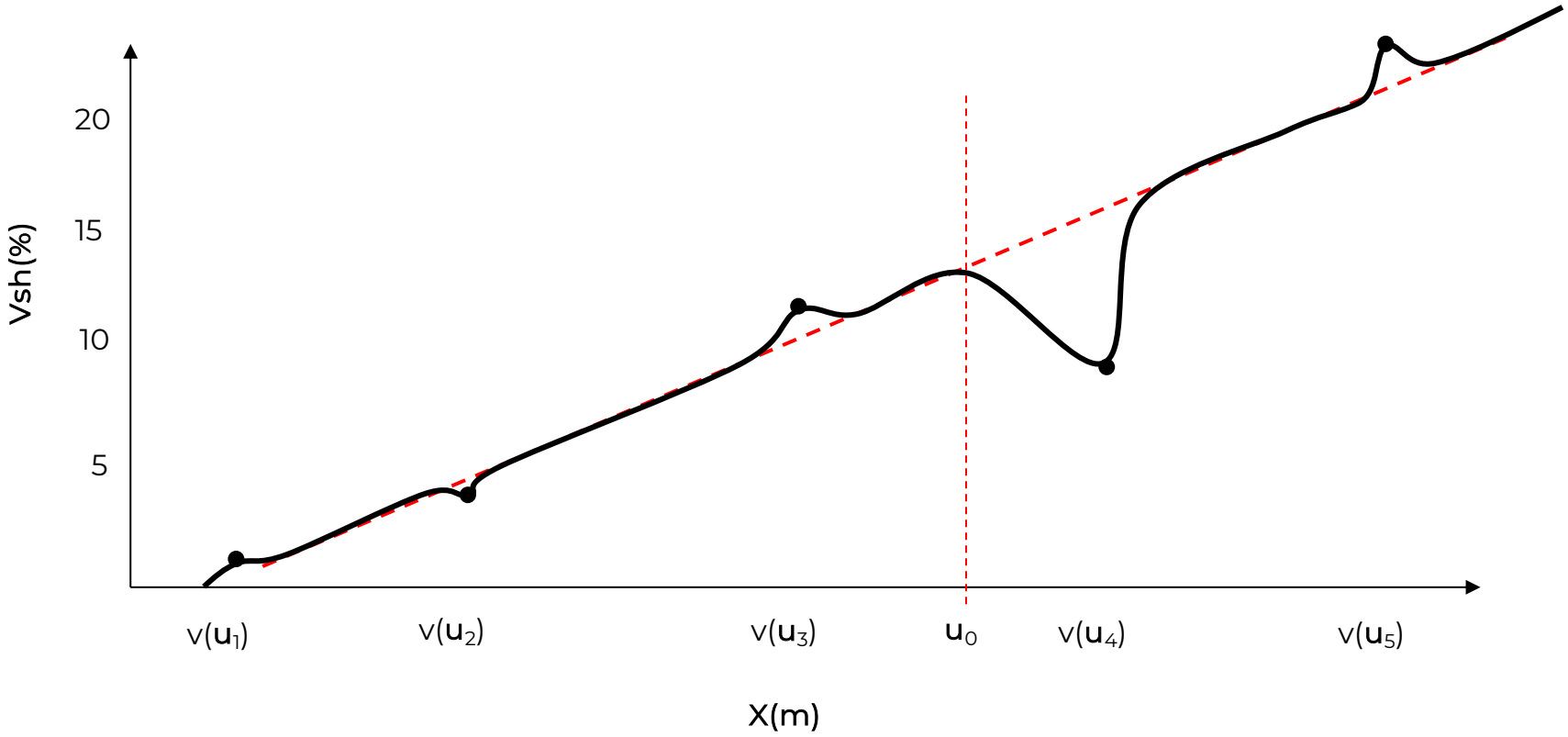
TREND MODELING

- Geostatistical spatial estimation methods will make an assumption concerning stationarity
  - Then model the residuals



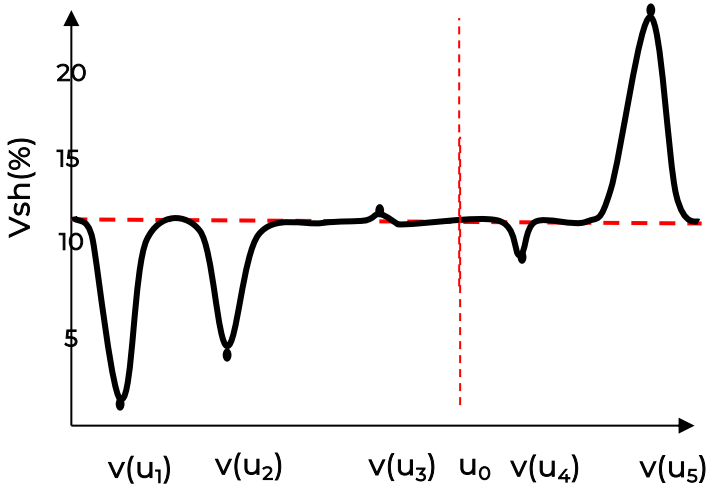
TREND MODELING

- Geostatistical spatial estimation methods will make an assumption concerning stationarity
  - After modeling, add the trend back to the modelled residuals

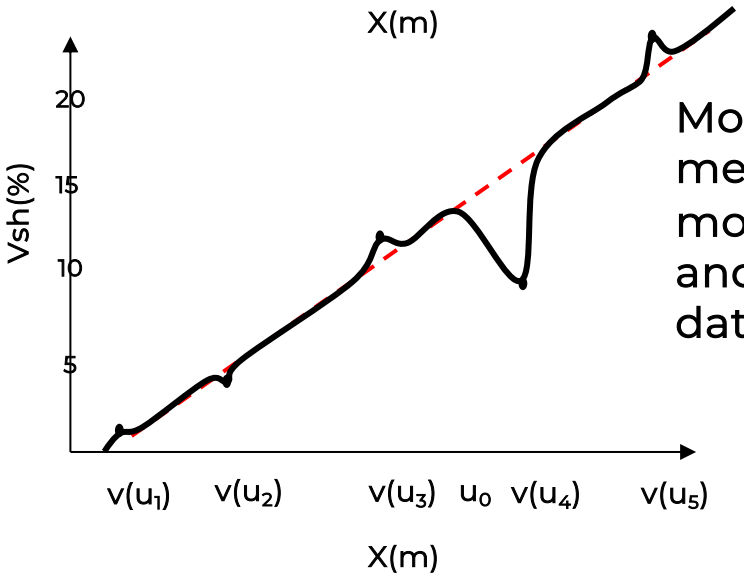


TREND MODELING

- How bad could it be if we did not model a trend?
- Geostatistical estimation would assume stationarity and away from data we would estimate with the global mean (simple kriging)!



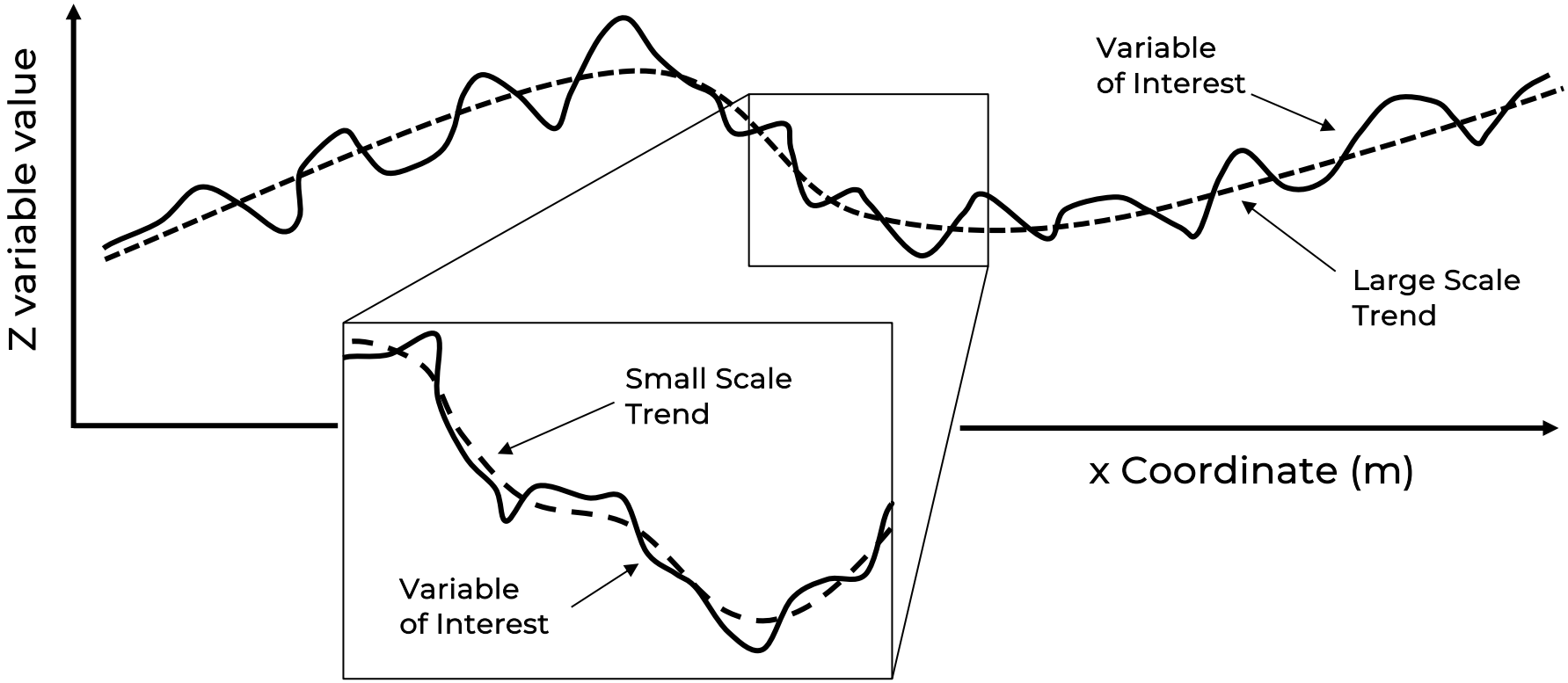
Model with stationary mean + data



Model with mean trend model and residual + data

TREND MODELING

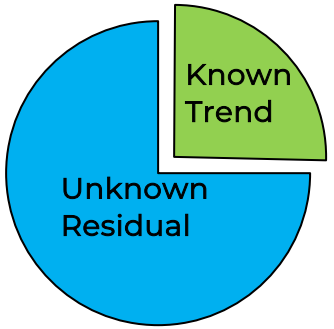
- Trend Modeling
  - We must identify and model trends / non-stationarities



TREND MODELING

- Any variance in the trend is removed from the residual:

$$\sigma_X^2 = \sigma_{X_t}^2 + \sigma_{X_r}^2 + 2C_{X_t,X_r}$$



- if the $X_t \perp\!\!\!\perp X_r, C_{X_t,X_r} = 0$,

$$\sigma_{X_r}^2 = \sigma_X^2 - \sigma_{X_t}^2$$

  - So if σ_X^2 is the total variance (variability), and $\sigma_{X_t}^2$ is the variability that is deterministically modelled, treated as known, and $\sigma_{X_r}^2$ is the component of the variability that is treated as unknown
  - Result: The more variability explained by the trend the less variability that remains as uncertain

DEFINITION DETERMINISTIC MODEL

- Model that assumes perfect knowledge, without uncertainty
- Based on knowledge of the phenomenon or trend fitting to data
- Most subsurface models have a deterministic component (trend) to capture expert knowledge and to provide a stationary residual for geostatistical modeling

TREND MODELING HANDS-ON

- Here's an opportunity for experiential learning with Trend Modeling
- Things to try:

1. Set the radius very large (50). How's the trend model performing? Try radius very small (1).
2. What do you think is the best radius to fit a trend to this spatial data?

- File Name: Daytum_Trends_v2.ipynb

Daytum +2 Course: Data Analytics, Geostatistics and Machine Learning Deep Dive

Trend Modeling Demonstration and Exercise

Goal

Calculate data-driven trend for a spatial dataset.

Description

Here's a simple, documented workflow, demonstration of trend calculation and diagnostics for subsurface modeling workflows. This should help you get started with building subsurface models that integrate trends.

Trend Modeling

Trend modeling is the modeling of local features, based on data and interpretation, that are deemed certain (known). The trend is substracted from the data, leaving a residual that is modeled stochastically with uncertainty (treated as unknown).

- geostatistical spatial estimation methods will make an assumption concerning stationarity
  - in the presence of significant nonstationarity we can not rely on spatial estimates based on data + spatial continuity model
- if we observe a trend, we should model the trend.
  - then model the residuals stochastically

Trend modeling significantly improves our models by accounting for the nonstationary map-able component of the spatial property. Consider this spatial model in 1D with and without trend.

INSERT FIGURE FROM 04_Spatial_Continuity.ppts SLIDE 50.

At a distance beyond the range of spatial continuity away from the data, the best estimate approaches the global mean or the trend model. WIthout the trend model we would systemally underestimate in the high regions and overestimate in the low regions.

Steps:

1. model trend consistent with data and intepretation at all locations within the area of itnerest, integrate all available information and expertise.

$$m(\mathbf{u}_\beta), \forall \beta \in \mathbf{AOI}$$

2. substract trend from data at the n data locations to formulate a residual at the data locations.

$$y(\mathbf{u}_\alpha) = \mathbf{z}(\mathbf{u}_\alpha) - \mathbf{m}(\mathbf{u}_\alpha), \forall \alpha = 1, \dots, \mathbf{n}$$

3. characterize the statistical behavoir of the residual $y(\mathbf{u}_\alpha)$ integrating any information sources and interpretations. For example the global cumulative distribution function and a measure of spatial continuity shown here.

$$F_y(y) \quad \gamma_y(\mathbf{h})$$

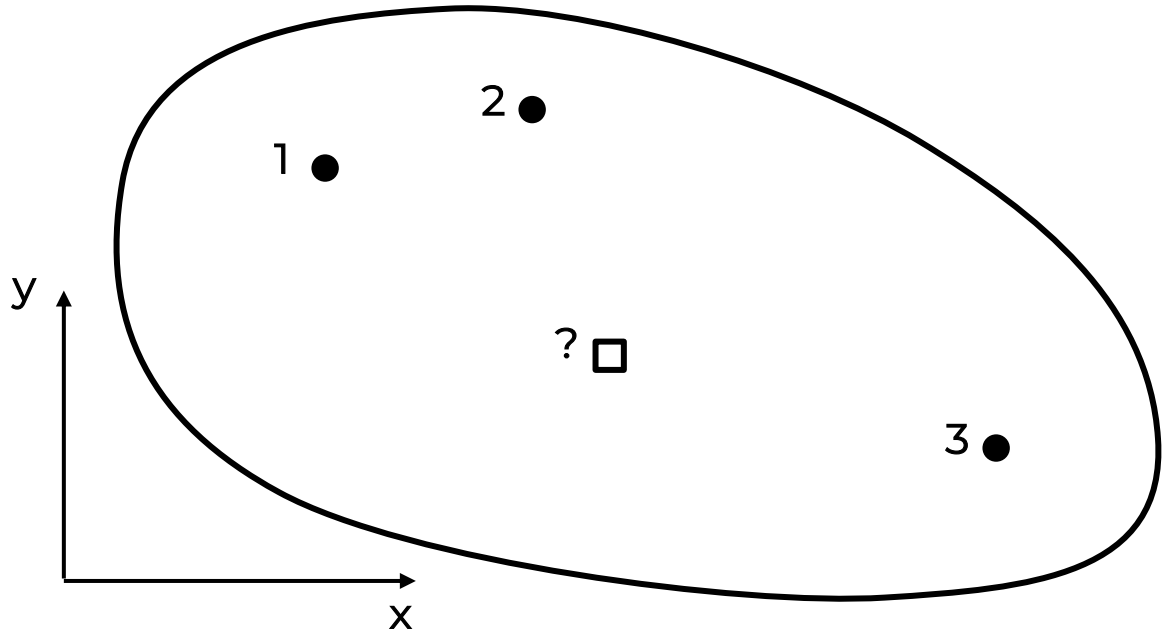
4. model the residual at all locations with L multiple realizations.

$$Y^\ell(\mathbf{u}_\beta), \forall \beta \in \mathbf{AOI}; \ell = 1, \dots, \mathbf{L}$$

5. add the trend back in to the stochastic residual realizations to calculate the multiple realizations, L, of the property of interest based on the composite model of known deterministic trend, $m(\mathbf{u}_\alpha)$ and unknown stochastic residual, $y(\mathbf{u}_\alpha)$

SPATIAL ESTIMATION

- Consider the case of estimating at some unsampled location:

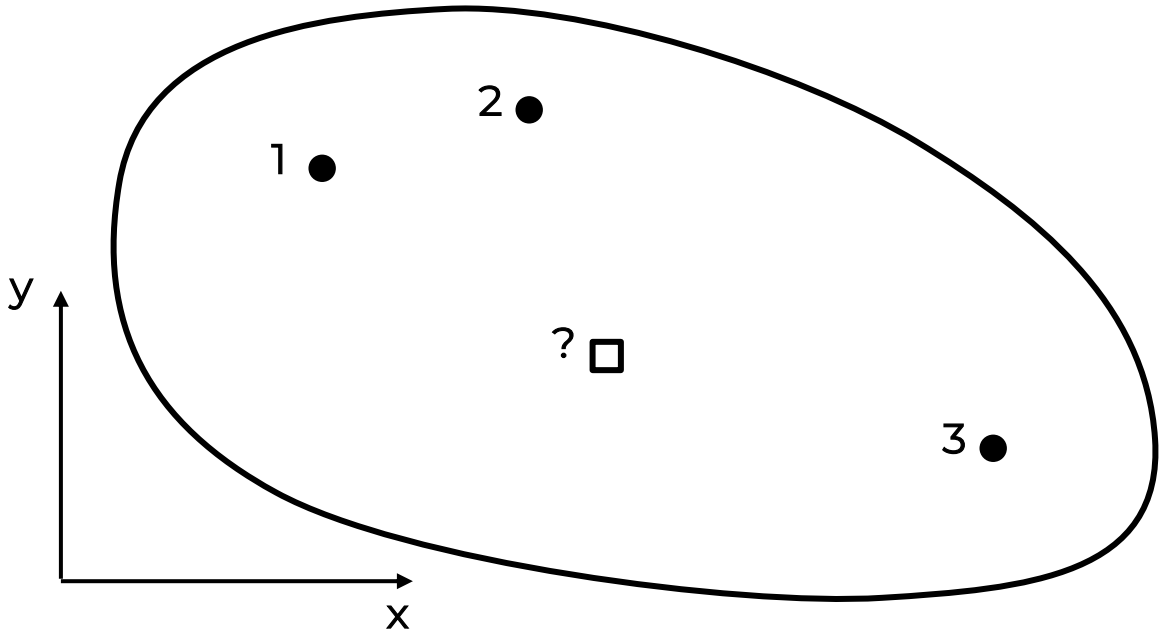


- How would you do this given data, $z(\mathbf{u}_1)$, $z(\mathbf{u}_2)$, and $z(\mathbf{u}_3)$?

Note: z is the variable of interest (e.g. porosity etc.) and u_i is the data locations.

SPATIAL ESTIMATION

- Consider the case of estimating at some unsampled location:



$\mathbf{z}(\mathbf{u}_\alpha)$ is the data values

$\boldsymbol{z}^*(\mathbf{u_0})$ is an estimate

$\boldsymbol{\lambda}_\alpha$ is the data weights

$\boldsymbol{m}_z$ is the global mean

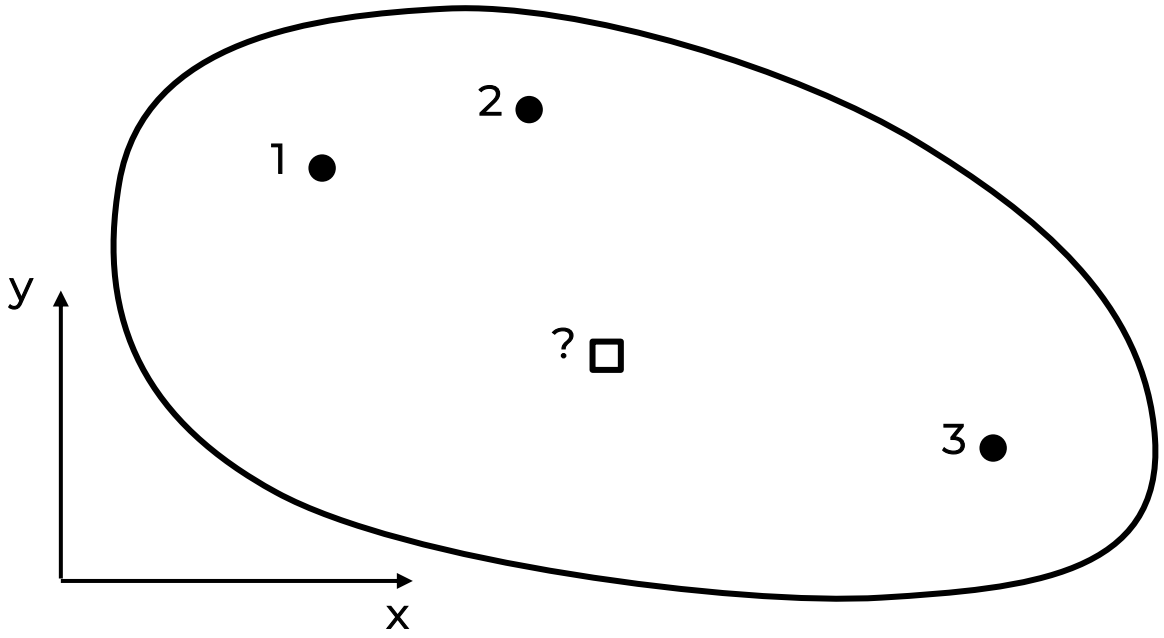
- How would you do this given data, $z(\mathbf{u}_1)$, $z(\mathbf{u}_2)$, and $z(\mathbf{u}_3)$?

$$\boldsymbol{z}^*(\mathbf{u_0}) = \sum_{\alpha=1}^{n} \boldsymbol{\lambda}_\alpha \mathbf{z}(\mathbf{u}_\alpha) + \left(1 - \sum_{\alpha=1}^{n} \boldsymbol{\lambda}_\alpha\right) \boldsymbol{m}_z$$

Unbiasedness Constraint
Weights sum to 1.0.

SPATIAL ESTIMATION

- Consider the case of estimating at some unsampled location:



- How would you do this given data, $z(\mathbf{u}_1)$, $z(\mathbf{u}_2)$, and $z(\mathbf{u}_3)$?

$$z^*(\mathbf{u_0}) - m_z(\mathbf{u_0}) = \sum_{\alpha=1}^{n} \lambda_\alpha \left(\mathbf{z}(\mathbf{u}_\alpha) - m_z(\mathbf{u}_\alpha)\right)$$

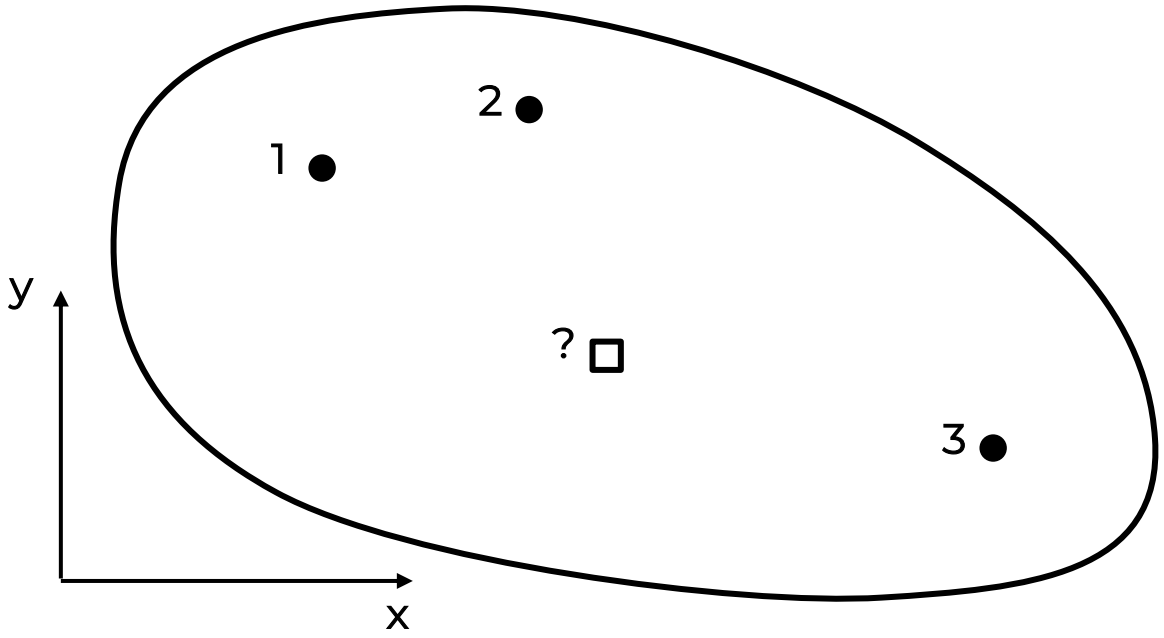
In the case where the mean is non-stationary.

Given y= $z - m$, $y^*(\mathbf{u_0}) = \sum_{\alpha=1}^{n} \lambda_\alpha \, \mathbf{y}(\mathbf{u}_\alpha)$

Simplified with residual, y.

SPATIAL ESTIMATION

- Consider the case of estimating at some unsampled location:



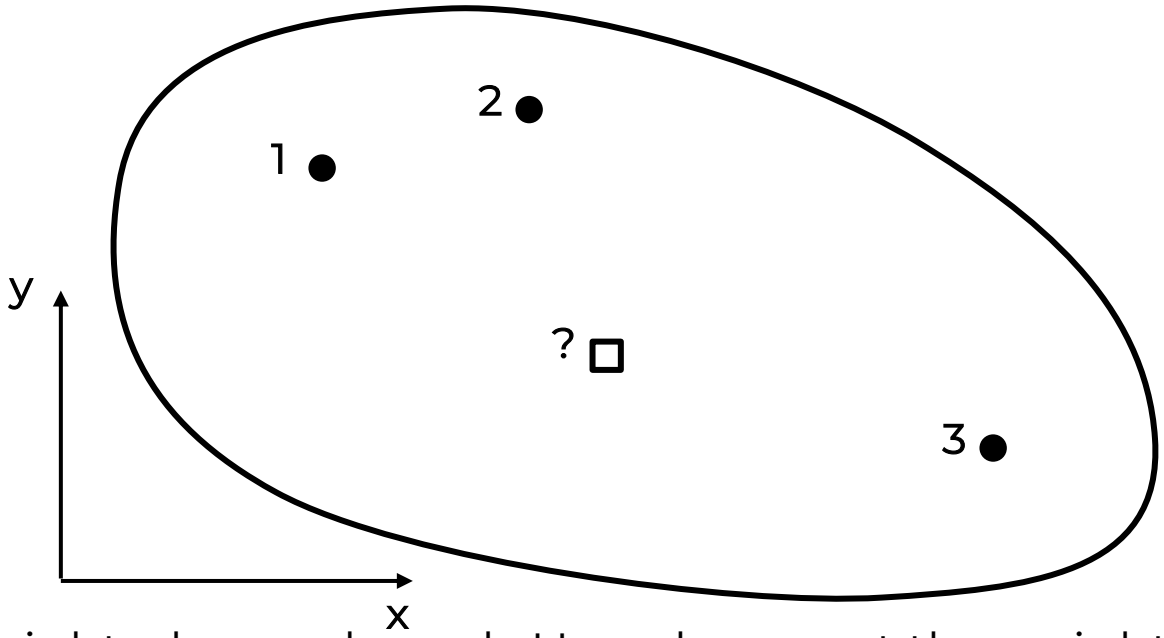
- Linear weighted, sound good. How do we get the weights? $\boldsymbol{\lambda_\alpha, \alpha = 1, \ldots, n}$

Simplified with residual, y.

$$\boldsymbol{y^*(\mathbf{u_0}) = \sum_{\alpha=1}^{n} \lambda_\alpha \, \mathbf{y}(\mathbf{u}_\alpha)}$$

SPATIAL ESTIMATION

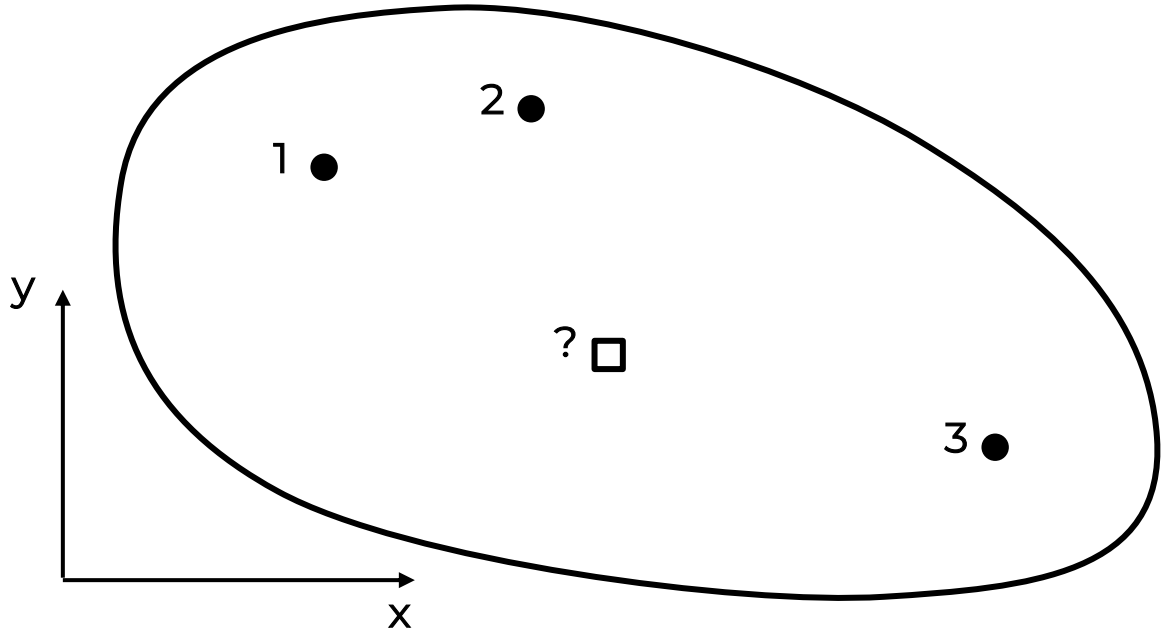
- Consider the case of estimating at some unsampled location:



- Linear weighted, sound good. How do we get the weights? $\boldsymbol{\lambda_\alpha, \alpha = 1, \ldots, n}$
- Equal weighted / average? $\boldsymbol{\lambda_\alpha = 1/n}$ Equal weight average of data
- What's wrong with that?

SPATIAL ESTIMATION

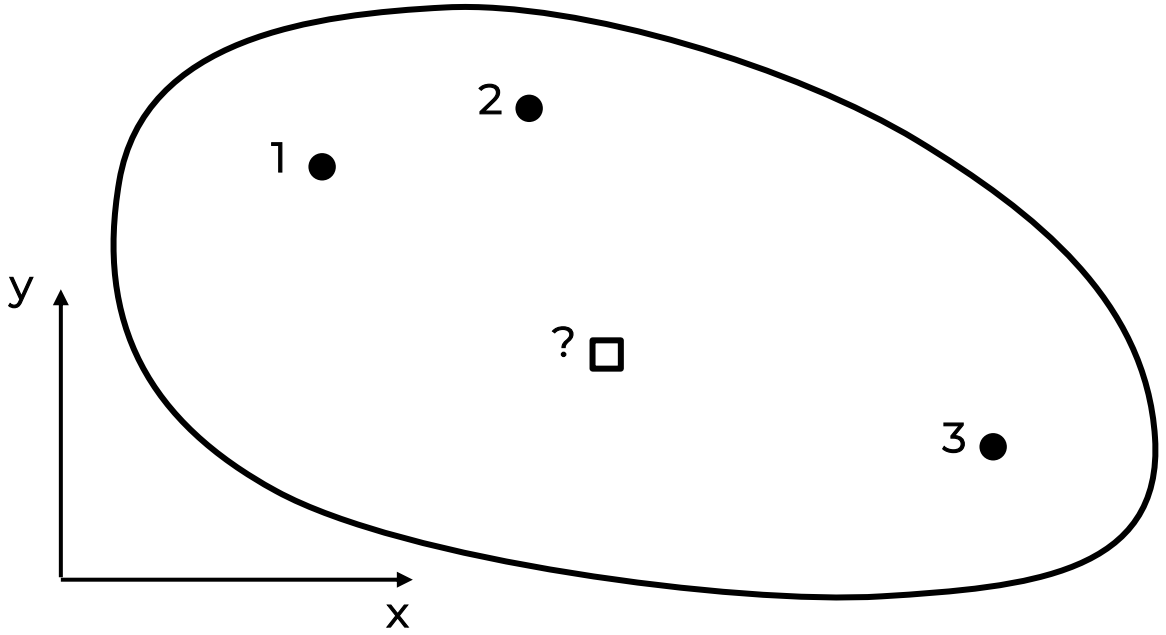
- Consider the case of estimating at some unsampled location:



- How do we get the weights? $\boldsymbol{\lambda_\alpha, \alpha = 1, \ldots, n}$
- Inverse distance? $\boldsymbol{\lambda_\alpha = \frac{\frac{1}{dist(\mathbf{u_0}, \mathbf{u_\alpha})^p}}{\sum_{\alpha=1}^{n} \lambda_\alpha}}$ Inverse distance to power standardized so weights sum to 1.0.
- What's wrong with that?

SPATIAL ESTIMATION

- Consider the case of estimating at some unsampled location:



- How do we get the weights? $\boldsymbol{\lambda_\alpha, \alpha = 1, \ldots, n}$
- It would be great to use weight that account for closeness (spatial correlation > distance alone), redundancy (once again with spatial correlation).
- How can we do that?

SPATIAL ESTIMATION

Kriging Derivations Removed

The results is a very useful and interpretable linear system of equations

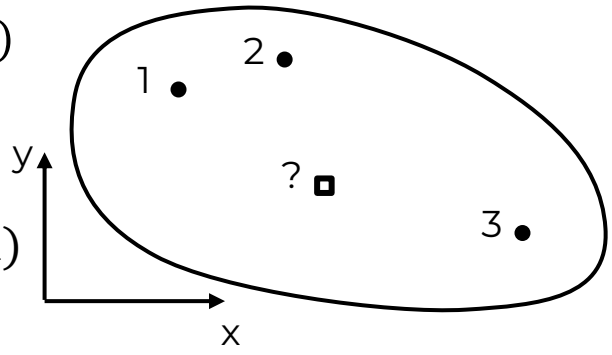
SIMPLE KRIGING: SOME DETAILS

- There are three equations to determine the three weights:

$$\lambda_1 \cdot C(\mathbf{u}_1, \mathbf{u}_1) + \lambda_2 \cdot C(\mathbf{u}_1, \mathbf{u}_2) + \lambda_3 \cdot C(\mathbf{u}_1, \mathbf{u}_3) = C(\mathbf{u}_0, \mathbf{u}_1)$$

$$\lambda_1 \cdot C(\mathbf{u}_2, \mathbf{u}_1) + \lambda_2 \cdot C(\mathbf{u}_2, \mathbf{u}_2) + \lambda_3 \cdot C(\mathbf{u}_2, \mathbf{u}_3) = C(\mathbf{u}_0, \mathbf{u}_2)$$

$$\lambda_1 \cdot C(\mathbf{u}_3, \mathbf{u}_1) + \lambda_2 \cdot C(\mathbf{u}_1, \mathbf{u}_2) + \lambda_3 \cdot C(\mathbf{u}_1, \mathbf{u}_3) = C(\mathbf{u}_0, \mathbf{u}_1)$$



- In matrix notation: Recall that $C(\mathbf{h}) = C(0) - \gamma(\mathbf{h})$
- $$\begin{bmatrix} C(\mathbf{u}_1, \mathbf{u}_1) & C(\mathbf{u}_1, \mathbf{u}_2) & C(\mathbf{u}_1, \mathbf{u}_3) \\ C(\mathbf{u}_2, \mathbf{u}_1) & C(\mathbf{u}_2, \mathbf{u}_2) & C(\mathbf{u}_2, \mathbf{u}_3) \\ C(\mathbf{u}_3, \mathbf{u}_1) & C(\mathbf{u}_3, \mathbf{u}_2) & C(\mathbf{u}_3, \mathbf{u}_3) \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \end{bmatrix} = \begin{bmatrix} C(\mathbf{u}_o, \mathbf{u}_1) \\ C(\mathbf{u}_o, \mathbf{u}_2) \\ C(\mathbf{u}_o, \mathbf{u}_3) \end{bmatrix}$$

redundancy

weights

closeness

PROPERTIES OF SIMPLE KRIGING

- What Does Kriging Provide?:
  - **Best Estimate**: Minimum error estimator (just try to pick weights, you won't bet it)
  - **Estimation Variance**: Provides a measure of the estimation (or kriging) variance (uncertainty in the estimate):

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

$$\boldsymbol{\sigma_E^2} \rightarrow [\mathbf{0}, \boldsymbol{\sigma_x^2}]$$

Estimation Variance

Variance

Kriging Weights

Covariance Between Data and Estimate Location

MORE PROPERTIES

- Exact interpolator: at data location
- Kriging variance can be calculated before getting the sample information, homoscedastic!
- Kriging takes into account:
  - Distance of the information: $C(\boldsymbol{u}, \boldsymbol{u_i})$
  - Configuration of the data: $C(\boldsymbol{u_i}, \boldsymbol{u_j})$
  - Structural continuity of the variable being considered: $C(\mathbf{h})$
- The smoothing effect of kriging can be forecast – we will return to this with simulation

SIMPLE KRIGING HANDS-ON

- Here's an opportunity for experiential learning with Simple Kriging
- Walkthrough together
- Things to try:

1. Decrease the variogram range
2. Increase the geometric anisotropy

- File Name: Daytum_Kriging.ipynb

Daytum +2 Course: Data Analytics, Geostatistics and Machine Learning Deep Dive

Spatial Estimation / Kriging Demonstration and Exercise ¶

Goal

Calculate spatial estimates away from spatial samples.

Description

Here's a simple, documented workflow, demonstration of spaital estimation for subsurface modeling workflows. This should help you get started with building subsurface models that predict away from available data.

Here's a simple workflow for spatial estimation with kriging and indicator kriging. This step is critical for:

1. Prediction away from wells, e.g. pre-drill assessments.
2. Spatial cross validation.
3. Spatial uncertainty modeling.

First let's explain the concept of spatial estimation.

Spatial Estimation

Consider the case of making an estimate at some unsampled location, $z(\mathbf{u_0})$, where z is the property of interest (e.g. porosity etc.) and $\mathbf{u}_0$ is a location vector describing the unsampled location.

How would you do this given data, $z(\mathbf{u_1})$, $z(\mathbf{u_2})$, and $z(\mathbf{u_3})$?

It would be natural to use a set of linear weights to formulate the estimator given the available data.

$$z^*(\mathbf{u}) = \sum_{\alpha=1}^{n} \lambda_\alpha \mathbf{z}(\mathbf{u}_\alpha)$$

We could add an unbiasedness constraint to impose the sum of the weights equal to one. What we will do is assign the remainder of the weight (one minus the sum of weights) to the global average; therefore, if we have no informative data we will estimate with the global average of the property of interest.

$$z^*(\mathbf{u}) = \sum_{\alpha=1}^{n} \lambda_\alpha \mathbf{z}(\mathbf{u}_\alpha) + \left(1 - \sum_{\alpha=1}^{n} \lambda_\alpha\right) \bar{\mathbf{z}}$$

We will make a stationarity assumption, so let's assume that we are working with residuals, y.

$$y^*(\mathbf{u}) = \mathbf{z}^*(\mathbf{u}) - \bar{\mathbf{z}}(\mathbf{u})$$

If we substitute this form into our estimator the estimator simplifies, since the mean of the residual is zero.

$$y^*(\mathbf{u}) = \sum_{\alpha=1}^{n} \lambda_\alpha \mathbf{y}(\mathbf{u}_\alpha)$$

SPATIAL ESTIMATION

New Tools

Topic	Application to Subsurface Modeling
Trend Modeling	Decompose variance into deterministic trend and stochastic residual. *30% of porosity variance is described by a linear depth trend and 70% is described by a 3D variogram model.*
Kriging Estimates and Kriging Variances	Kriging provides the best estimate and a measure of estimation variance. *Given a kriging estimate of 13% and kriging variance of 9% and the assumption of a Gaussian distribution we have a complete local distribution of uncertainty for pre-drill porosity.*

Spatial Data Analytics

Lecture Recap

- Stationarity
- Spatial Continuity
- Variogram Calculation
- Spatial Estimation