PGE 338 Data Analytics and Geostatistics

Lecture 9b: Spatial Bias

Lecture outline . . .

- Sampling Bias
- Declustering
- Debiasing with Secondary Data

Introduction

General Concepts

Univariate

Bivariate

Correlation

Regression

Model Checking

Time Series Analysis

Spatial Analysis

Machine Learning

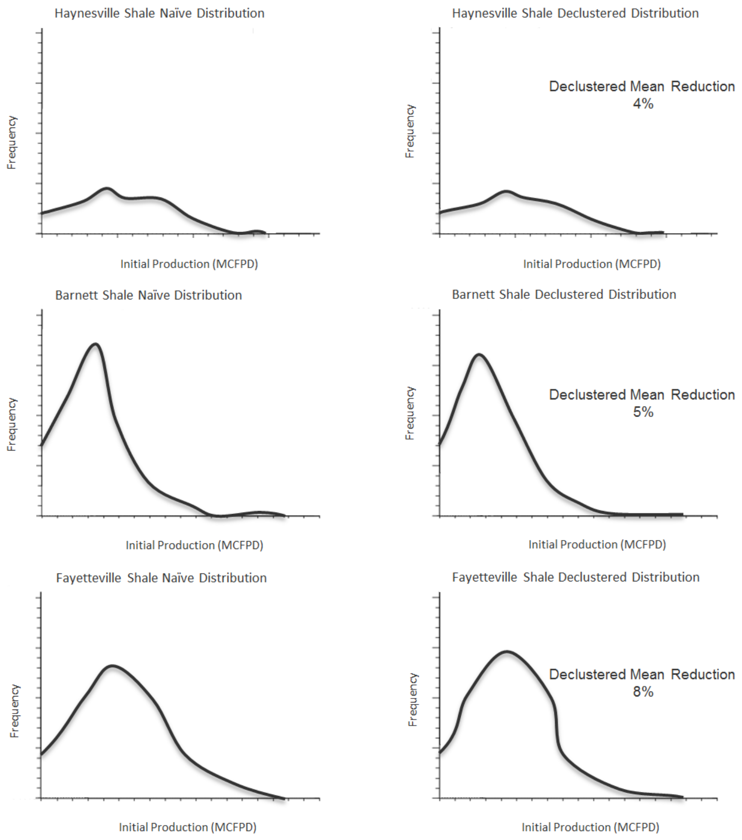
Uncertainty Analysis

Michael Pyrcz, The University of Texas at Austin

Motivation

Virtually all subsurface spatial data is sampled in a biased manner.

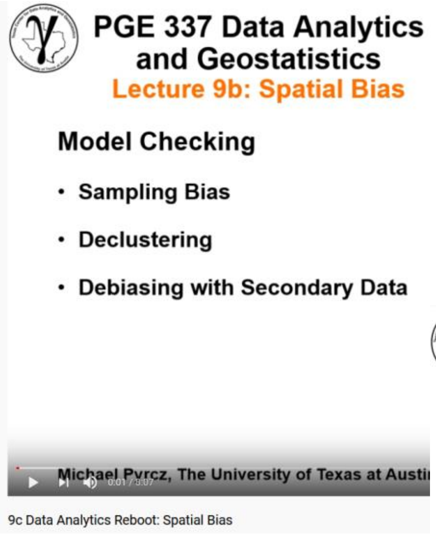
- We **CANNOT** use the raw statistics from these datasets to build models and support decision making.



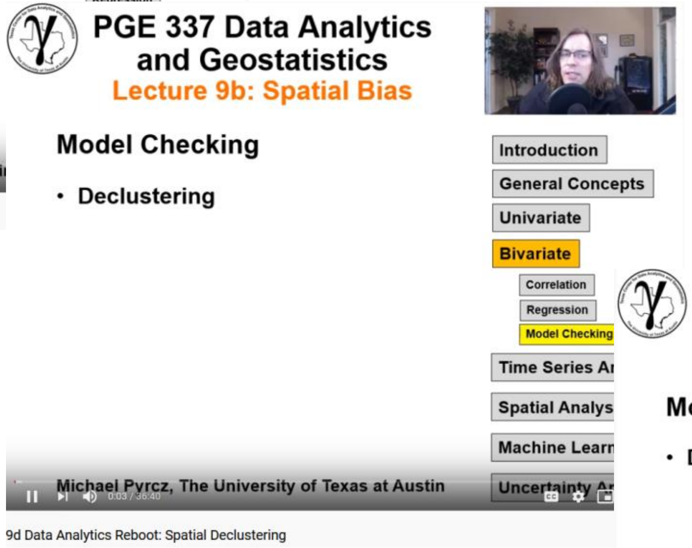
Naïve and Declustered distributions
for some US unconventional reservoirs.

Recorded Lectures

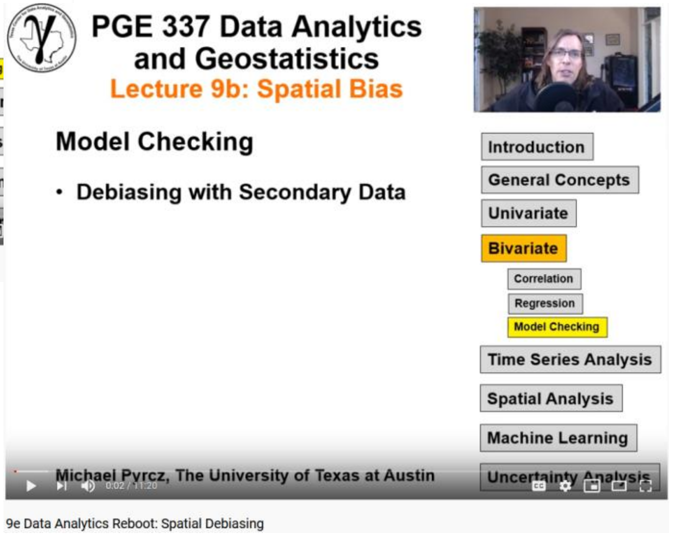
Intro



Declustering

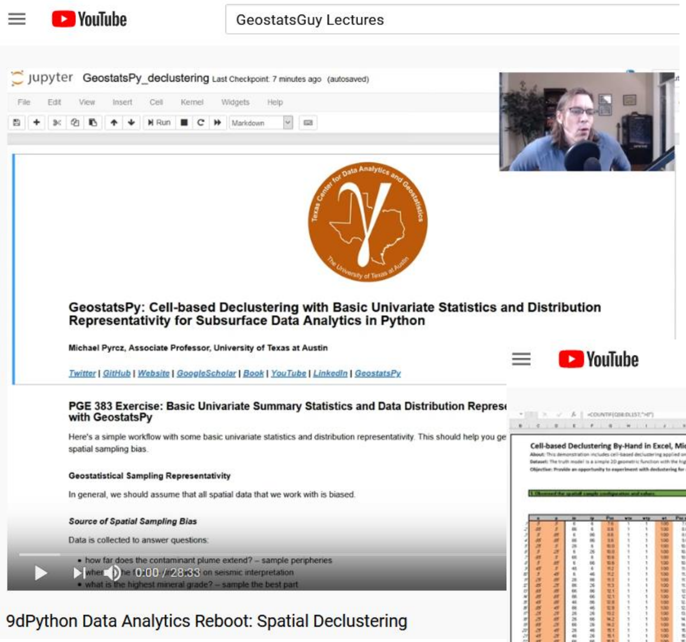


Debiasing



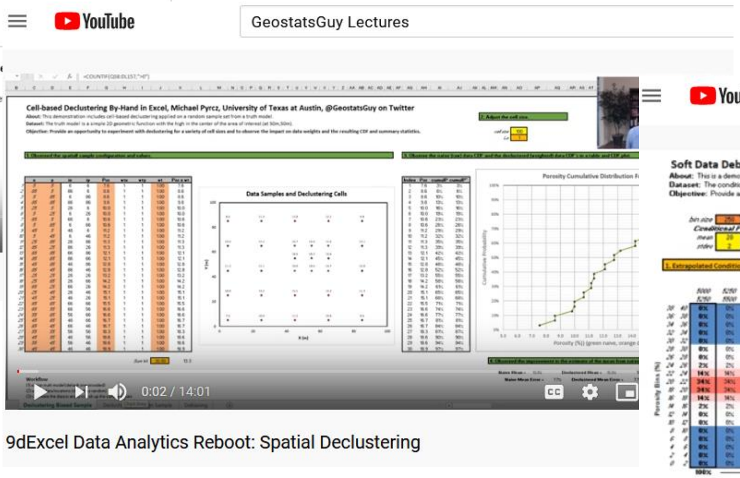
Updated the previous recorded lecture. Some improvements in content, improved sound, shorter, topical videos.

Recorded Demos

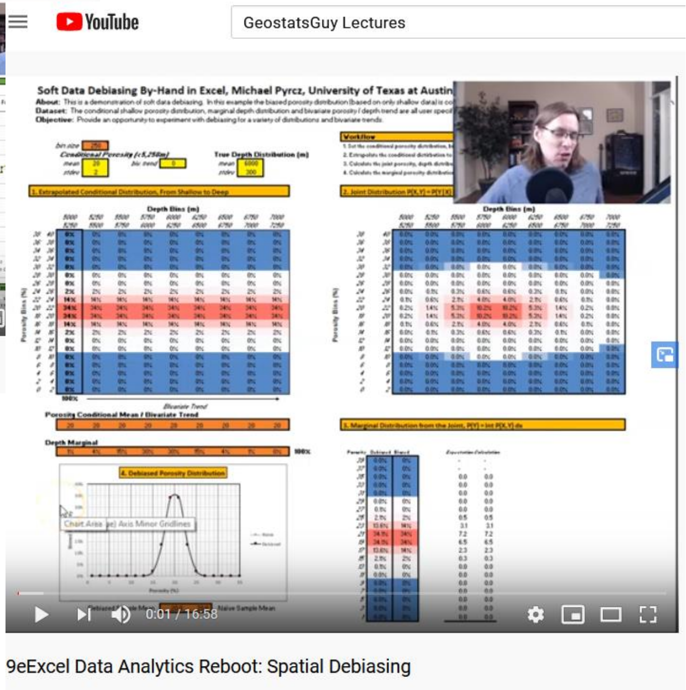


Intro

Declustering in Excel



Debiasing in Excel



Declustering in Python

I also recorded demonstrations that should help with understanding and to complete homework assignment.

PGE 338 Data Analytics and Geostatistics

Lecture 9b: Spatial Bias

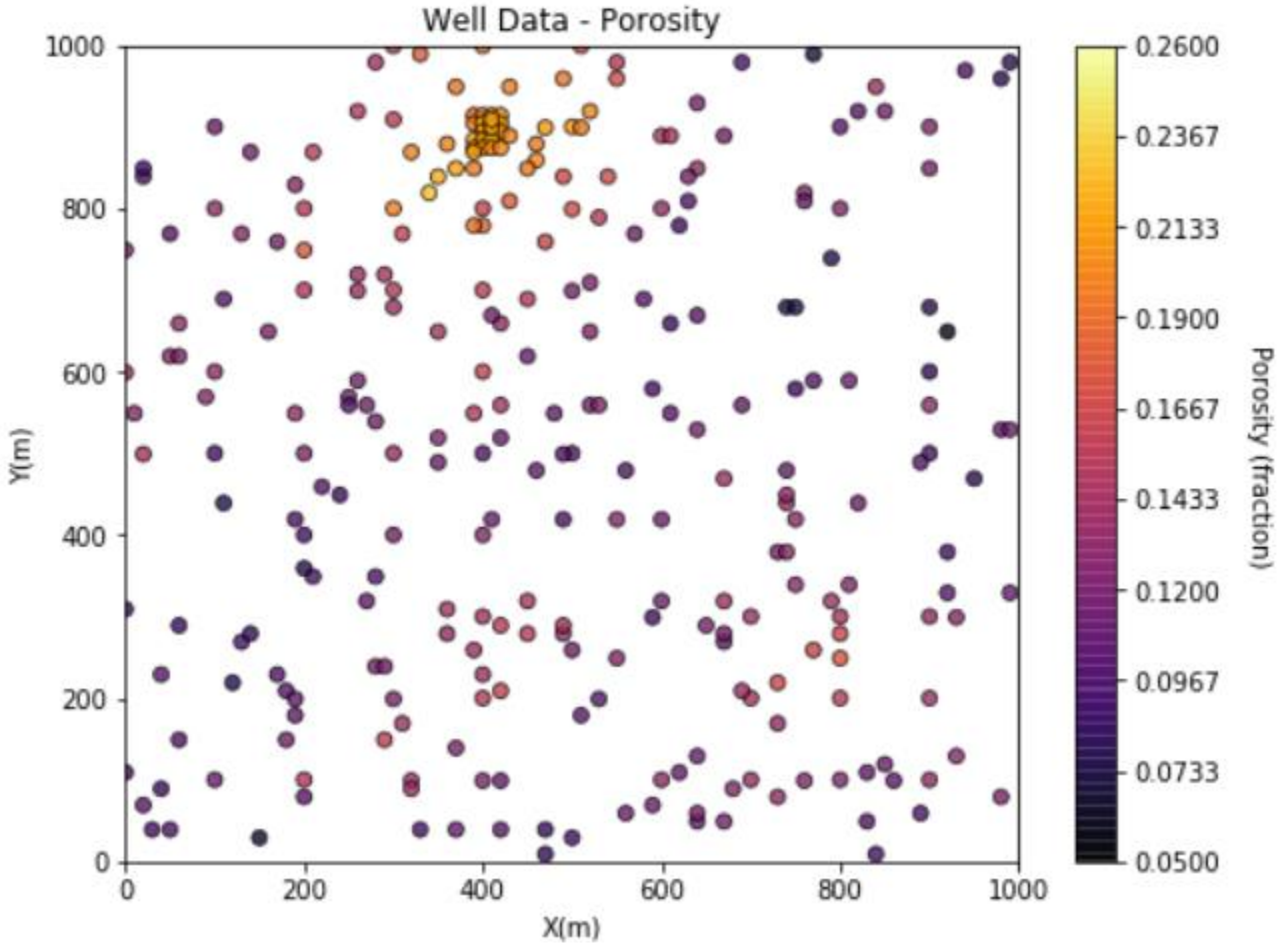
Lecture outline . . .

- Sampling Bias

Michael Pyrcz, The University of Texas at Austin

Sampling Bias

What is wrong with this sample set?

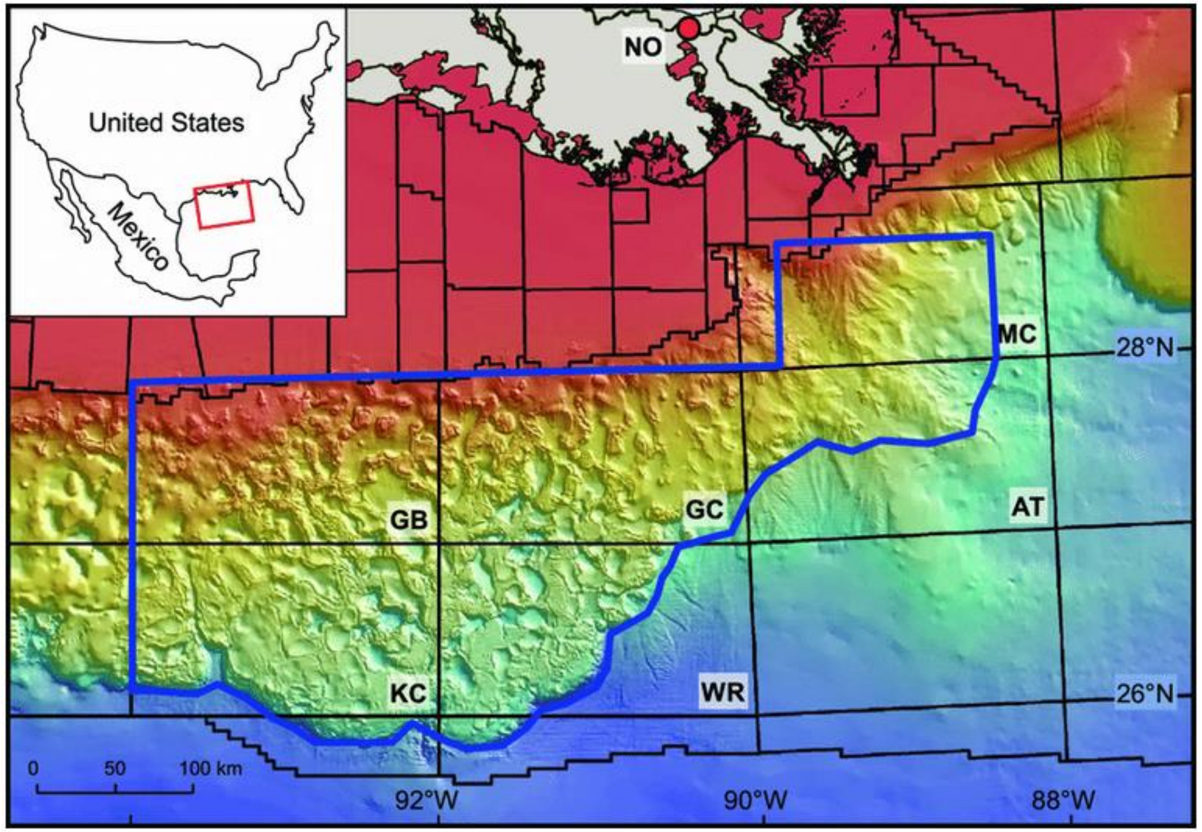


Map of average porosity at wells over a reservoir.

Spatial Data Collection

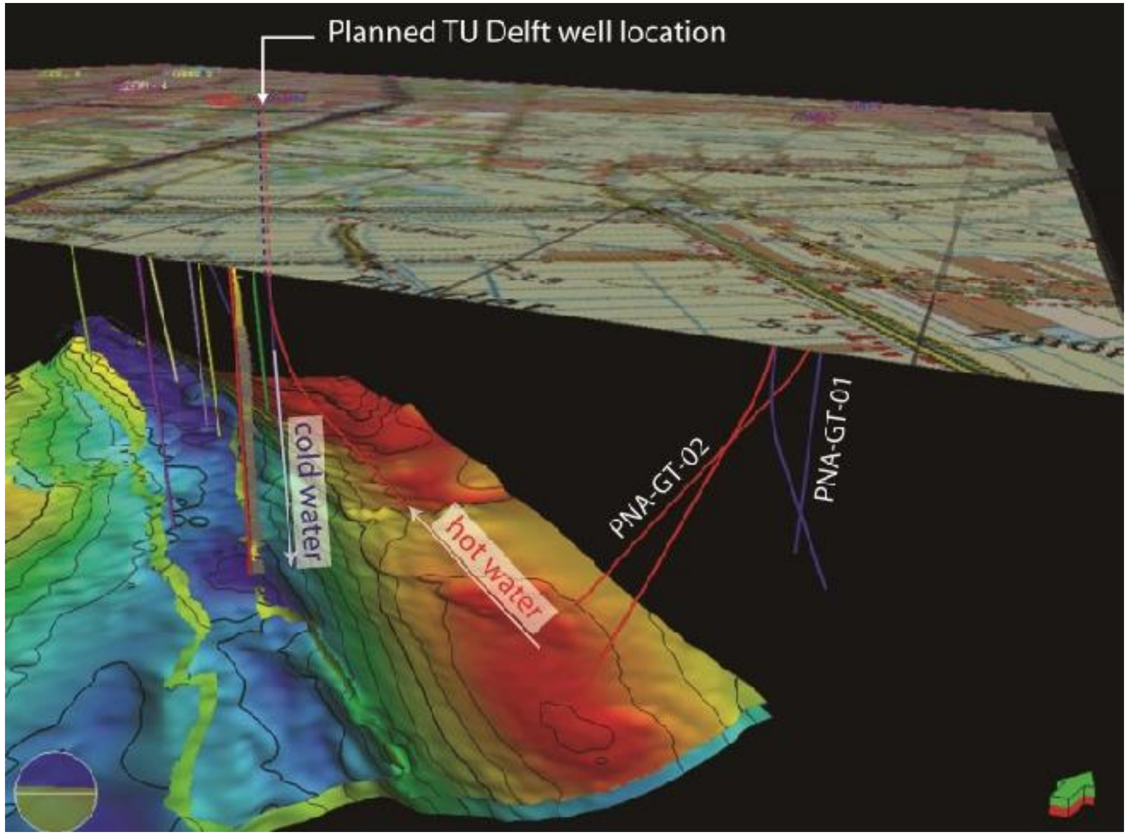
How do we decide where to drill?

Exploration Drilling



Bathymetric maps of Northern Gulf of Mexico (NOAA) with GB – Garden Banks, KC – Keathley Canyon, GC – Green Canyon, WR – Walker Ridge, MC - Mississippi Canyon, and AT - Atwater Valley (Kilsdonk, 2011).

Appraisal and Development Drilling



Structural model of Delft Sandstone Member in West Netherlands Basin with existing and planned wells (Donselaar et al., 2015).

Spatial Data Collection

Data is collected to:

1. answer questions

- how far does the contaminant plume extend? – *sample peripheries*
- where is the fault? – *drill based on seismic interpretation*
- what is the highest mineral grade? – *sample the best part*
- who far does the reservoir extend? – *offset drilling*

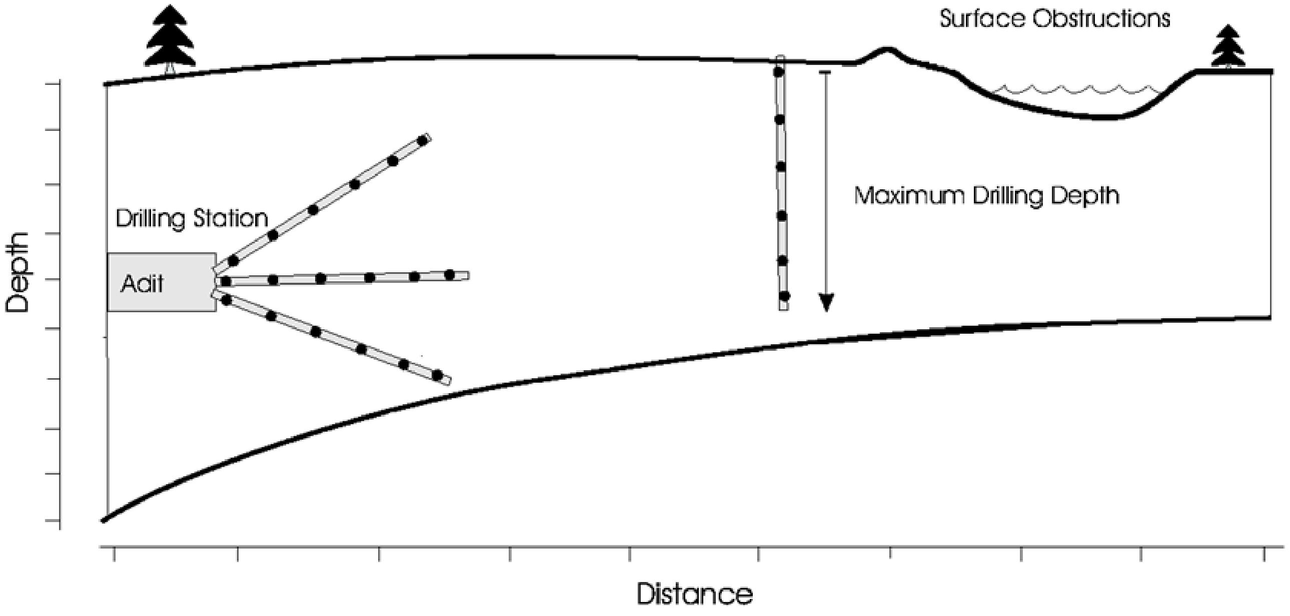
2. maximize NPV directly

- maximize production rates

Spatial Data Collection Limitations

There are also limits to our data collection:

- **limits in accessibility to the sample** – obstruction, reliable drilling, subsalt imaging limit where we can drill
- **limits of sample handling** – may not be able to recover shale core samples from depth
- **limits of measurements** - can't run permeability the on very low permeability rock



Schematic of subsurface data collection (Pyrcz and Deutsch, 2003).

Spatial Data Collection

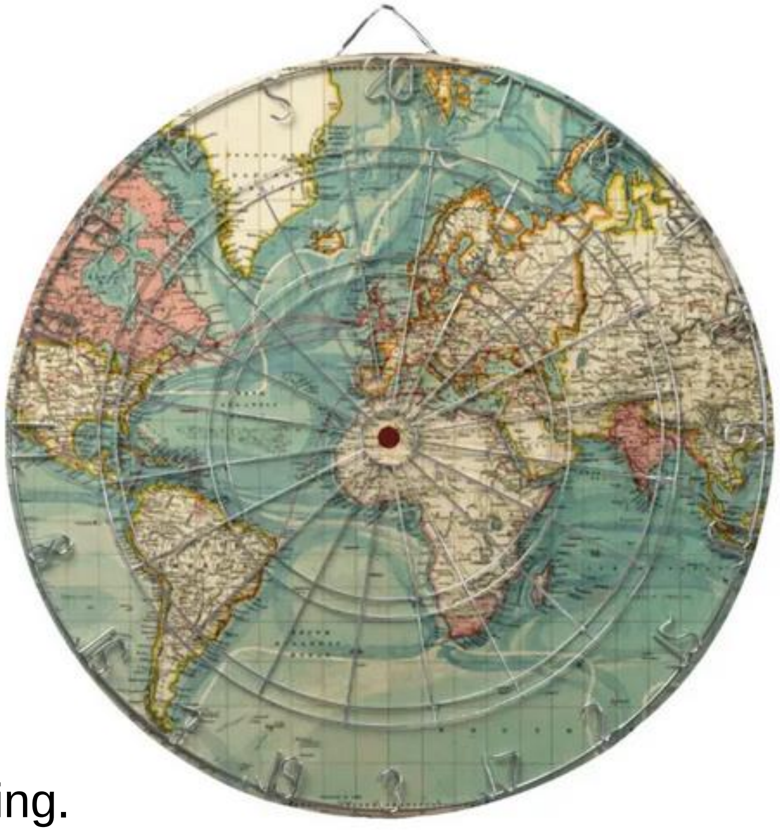
If we were sampling for representativity of the sample set and resulting sample statistics, by theory we have 2 options:

1. random sampling
2. regular sampling (as long as we don't align with natural periodicity)

What would happen if you proposed random sampling (well location) in the Gulf of Mexico at $150M per well?

We should not change our current sampling methods! Sampling to maximize profit and minimize uncertainty has the best economics, we should address sampling bias in the data.

Therefore, never use raw spatial data without access sampling bias / correcting.



Vintage world map dart board.

Image from https://www.zazzle.com/vintage_world_map_dartboard_with_darts

Representative Sampling Definition

Sampling that avoids bias, or preselection, when selecting from the population.

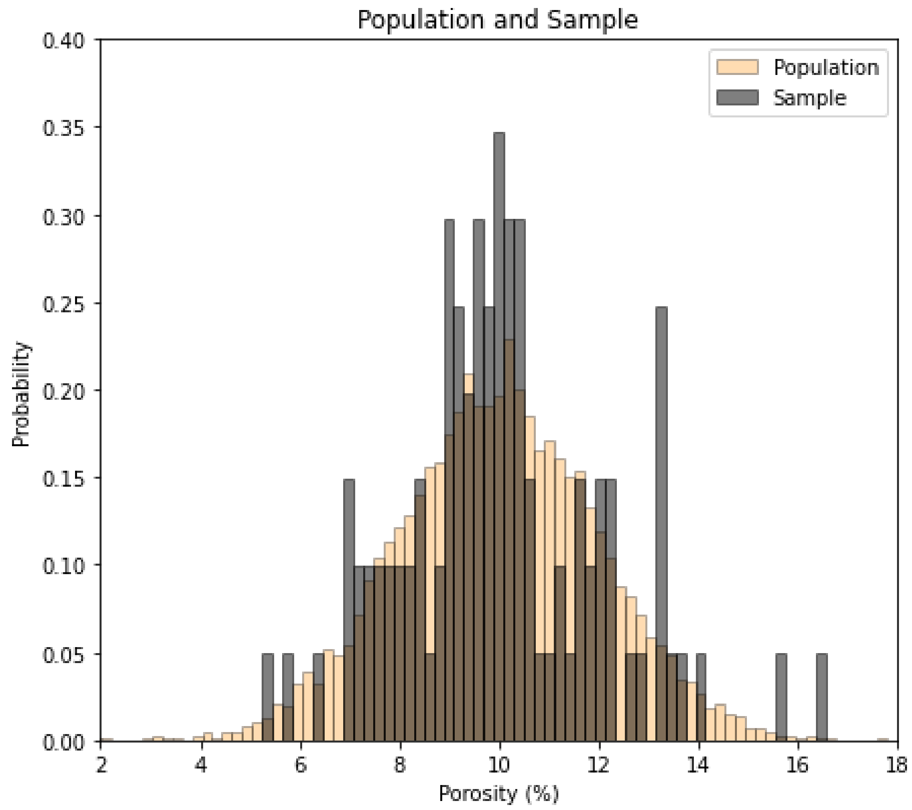
Sampling that results in statistics that match the population parameters in expectation.

For example, given z^s is a sample set and Z is the population.

mean: $E\{z^s\} = E\{Z\}$,

13th percentile: $E\{F_{z^s}^{-1}(0.13)\} = F_Z^{-1}(0.13)$

and so on…



Example population and 100 samples.

Simple Random Sampling Definition

Recall the population is subsurface as an exhaustive set of mutually exclusive volumes at the scale of the measurement tool.

Each potential sample from the population is equally likely to be sampled at each step.

- Each location is the subsurface is just as likely to be sampled.
- Selecting a specific location has no impact on the selection of subsequent locations.

Assumes the population size that is much larger than the sample size:

- Therefore, there is not significant correlation imposed due to without replacement sampling (the constraint that you can only sample a location once).
- Generally not an issue for the subsurface, massive populations sparsely sampled

Other Common Sampling Issues

Preselection Bias / Survivorship Bias

- e.g. any study that focusses on success cases

Sample Design Framework

- traditional statistical analysis requires careful sample design vs. we typically work with the data we get!

Spatial Sample Bias

- typically significant and we will cover mitigation methods later

We should assume all of our spatial data sets are biased.

Cognitive Biases

In any modeling there will be choices. We must understand and mitigate our own biases.

Example of Cognitive Biases:

1. **Anchoring Bias**: too much emphasis on first piece of information. Studies have shown that first piece of information could be completely irrelevant!
2. **Availability Heuristic**: overestimate importance of information available to them. “My grandpa smoked 3 packs a day and lived to 100”.
3. **Bandwagon Effect**: probability increases with the number of people holding the belief.
4. **Blind-spot Effect**: fail to see your own cognitive biases.
5. **Choice-supportive Bias**: probability increases after a commitment, decision is made.
6. **Clustering Illusion**: seeing patterns in random events.
7. **Confirmation Bias**: only consider new information that supports current model.
8. **Conservatism Bias**: favor old data to newly collected data.
9. **Recency Bias**: favor the most recently collected data.
10. **Survivorship Bias**: focus on success cases only.

Reference Only – Not on Test

Solutions to Biased Spatial Data

There is a need, however, to adjust the histograms and summary statistics to be representative of the entire volume of interest. We use statistics to make decisions!

1. **Declustering techniques assign each datum a weight** based on closeness to surrounding data
   - $w_i, i = 1, \ldots, n$ (weights are greater than 0 and sum to n)
   - Histogram and cumulative histogram use $w_i, i = 1, \ldots, n$ instead of equal weighted, $w_i = 1.0$.

2. **Debiasing techniques** derive an entirely new distribution based on a secondary data source such as geophysical measurements or expert interpretation

PGE 338 Data Analytics and Geostatistics

Lecture 9b: Spatial Bias

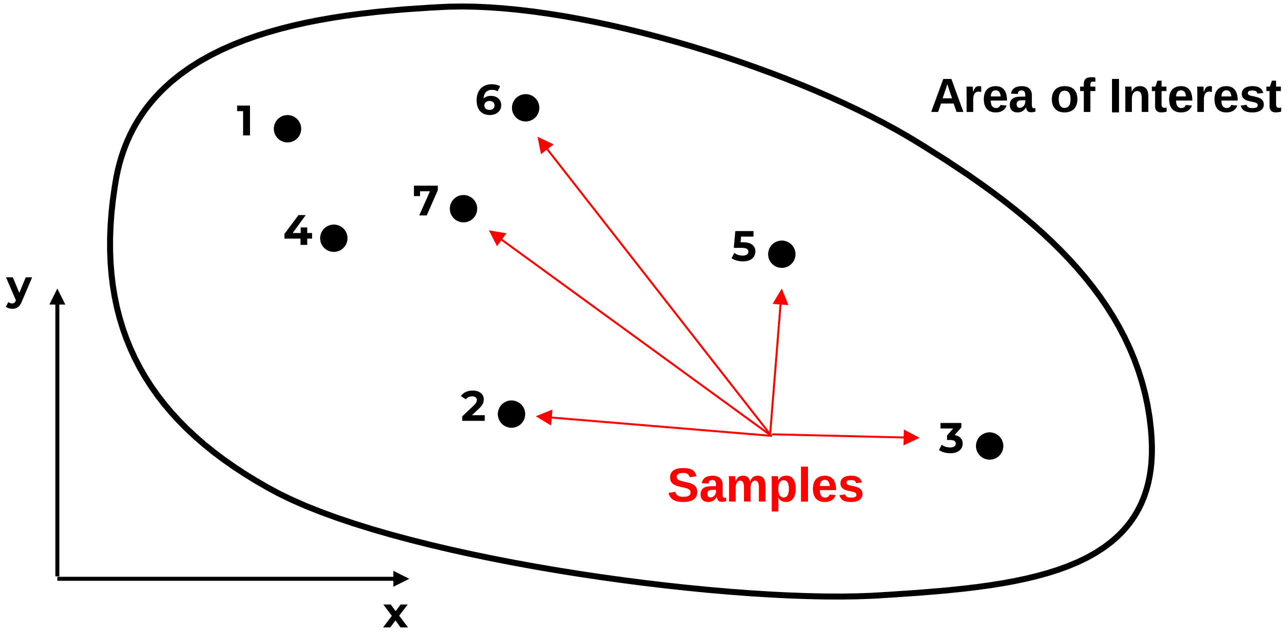
Lecture outline . . .

- Declustering

Michael Pyrcz, The University of Texas at Austin

Clustered Sampling

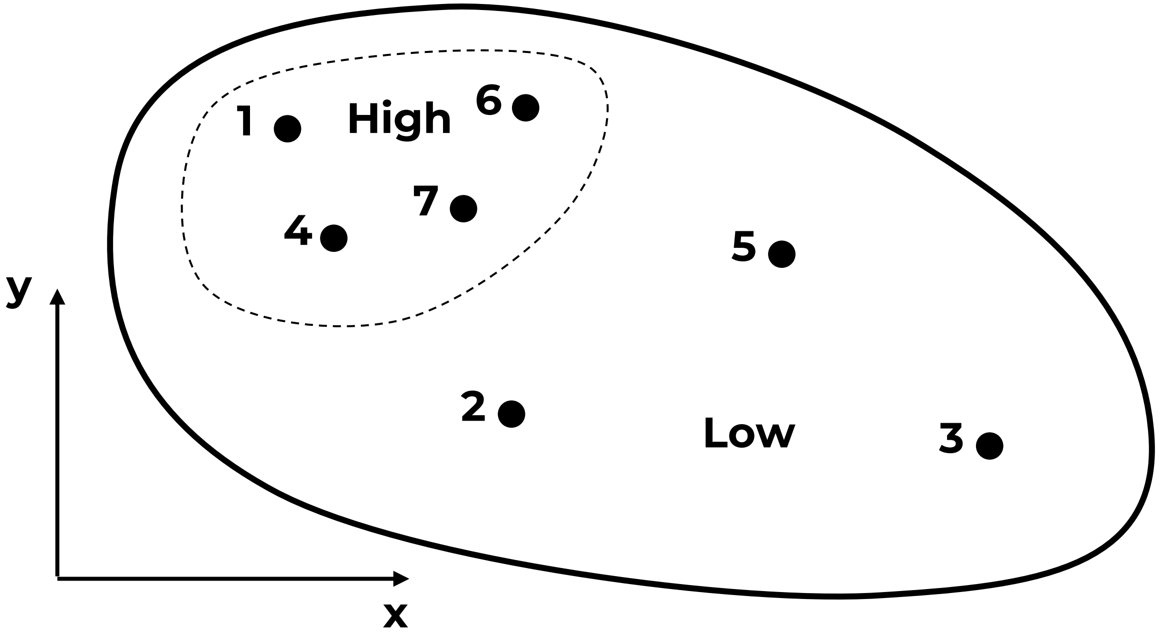
Let's make an estimate for an Area / Volume of Interest:



- e.g. the average porosity to calculate OIP

Clustered Sampling

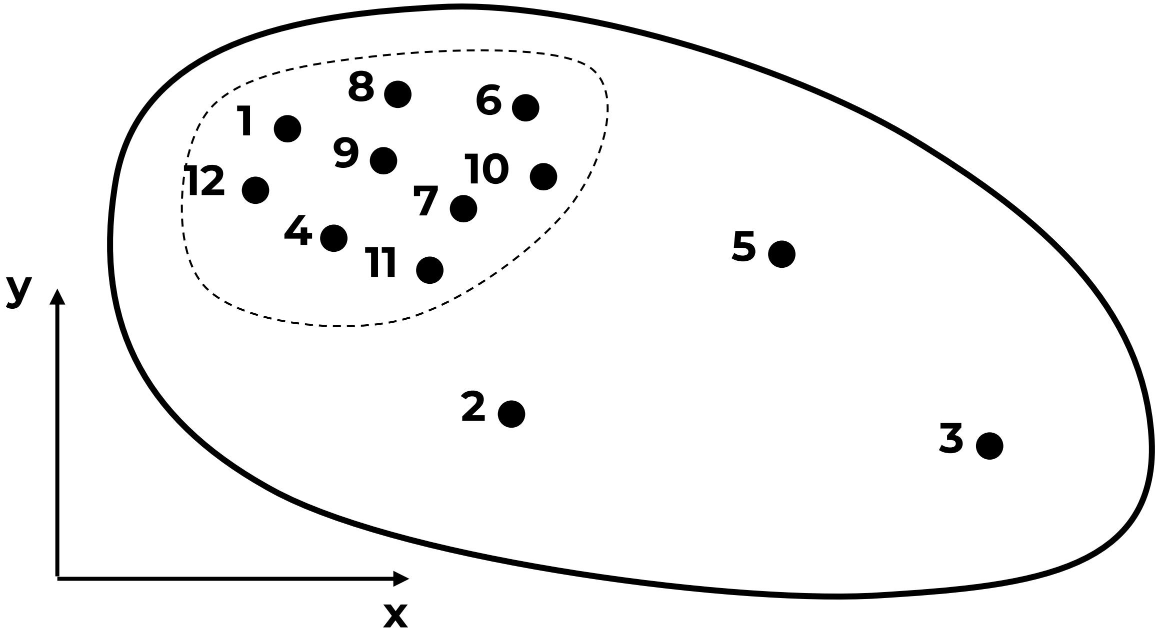
Let's make an estimate for an Area / Volume of Interest:



- What if we knew from seismic that the reservoir quality is better in the top left area?

Clustered Sampling

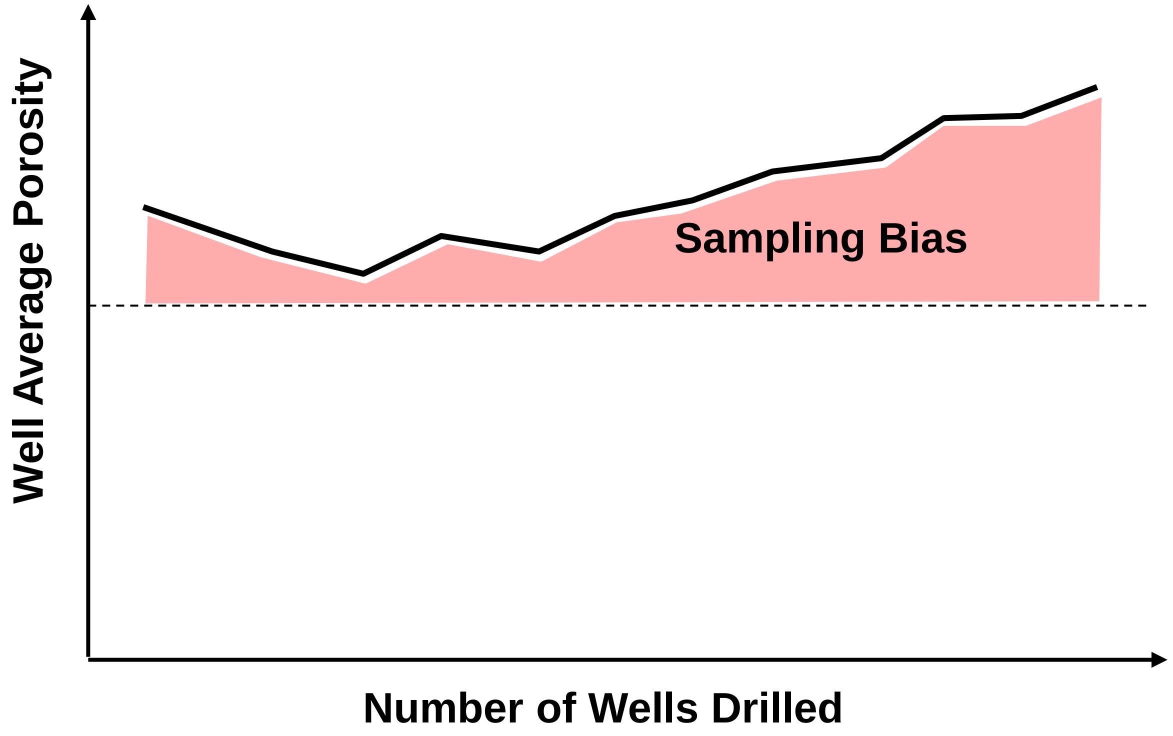
Let's make an estimate for an Area / Volume of Interest:



- What if we kept drilling in the high value region of the area of interest?

Clustered Sampling

How would our estimate of average porosity change as we drilled more wells?



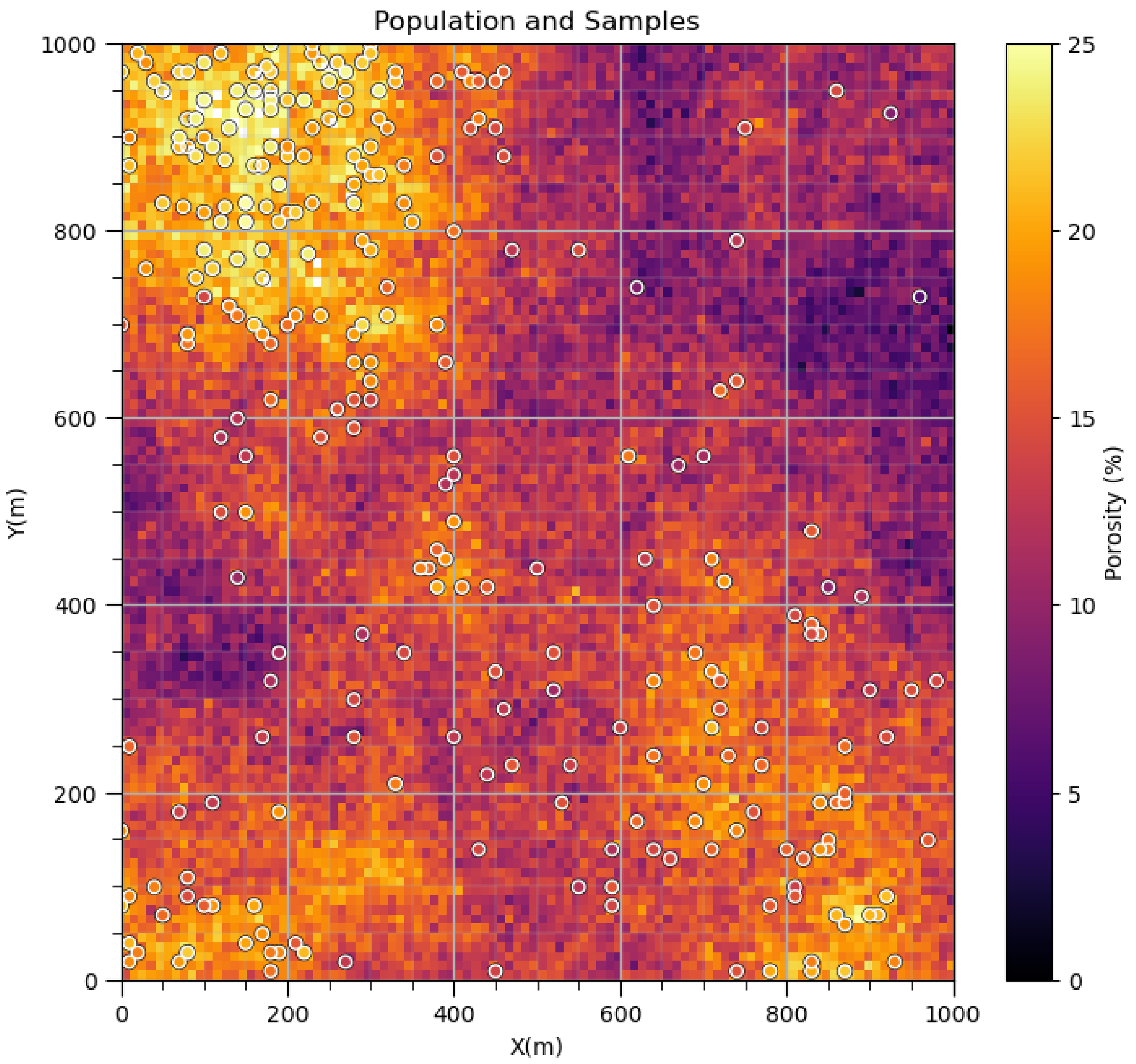
- The naïve sample average becomes more biased!
- We need a method to correct for clustered samples.

Some Clustered Data

Hypothetically, let's say we could also see the population.

- Location map of 270 samples.

- Any issue with the samples vs. the the unknown population.



Samples and population (left), population distribution (upper right) and sample distribution (lower right), from make_nonlinear_MV_spatial_data_v5_sand_only.ipynb.

Solutions to Biased Spatial Data

- There is a need, however, to adjust the histograms and summary statistics to be representative of the entire volume of interest. We use statistics to make decisions!

1. **Declustering techniques assign each datum a weight** based on closeness to surrounding data
   - $w_i, i = 1, \ldots, n$ (weights are greater than 0 and sum to n)
   - Histogram and cumulative histogram use $w_i, i = 1, \ldots, n$ instead of equal weighted, $w_i = 1.0$.

2. **Debiasing techniques** derive an entirely new distribution based on a secondary data source such as geophysical measurements or expert interpretation

Cell Declustering

Cell Declustering, a method for calculating declustering weights

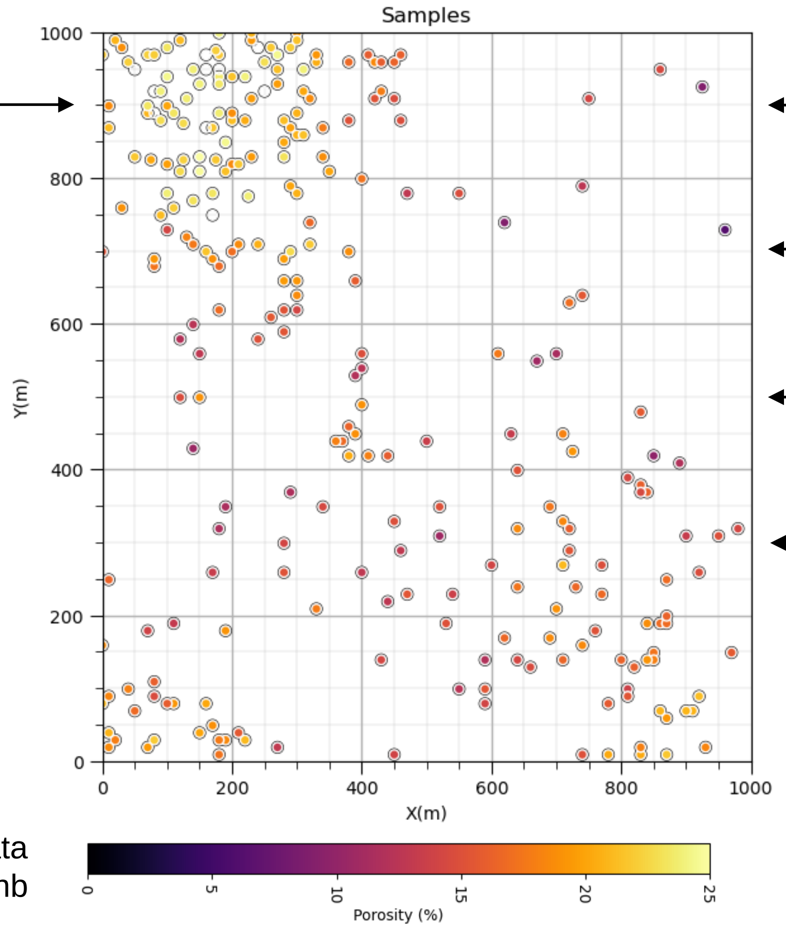
- divide the volume of interest into a grid of cells $l = 1, \dots, L$ count the occupied cells L_o and the number in each cell $n_l, l = 1, \dots, L_o$, weight inversely by number in cell (standardize by L_o)

Cell Declustering Data Weights

$$w(\mathbf{u}_j) = \frac{1}{n_l}\frac{n}{L_0}$$

0.20 weight →

All data in the same cell get the same weight.



Clustered, biased data
GeostatsPy_declustering.ipynb

$$\frac{1}{n_l}\frac{n}{L_0} = \frac{1}{2 \text{ data in cell}} \times \frac{270 \text{ data}}{25 \text{ cells with data}} = \mathbf{5.4 \text{ weight}}$$

$$\frac{1}{n_l}\frac{n}{L_0} = \frac{1}{1 \text{ data in cell}} \times \frac{270 \text{ data}}{25 \text{ cells with data}} = \mathbf{10.8 \text{ weight}}$$

$$\frac{1}{n_l}\frac{n}{L_0} = \frac{1}{3 \text{ data in cell}} \times \frac{270 \text{ data}}{25 \text{ cells with data}} = \mathbf{3.6 \text{ weight}}$$

$$\frac{1}{n_l}\frac{n}{L_0} = \frac{1}{10 \text{ data in cell}} \times \frac{270 \text{ data}}{25 \text{ cells with data}} = \mathbf{1.8 \text{ weight}}$$

$$\frac{1}{n_l}\frac{n}{L_0} = \frac{1}{20 \text{ data in cell}} \times \frac{270 \text{ data}}{25 \text{ cells with data}} = \mathbf{0.9 \text{ weight}}$$

Sum of all weights = n
Nominal / nonclustered weight = 1.0

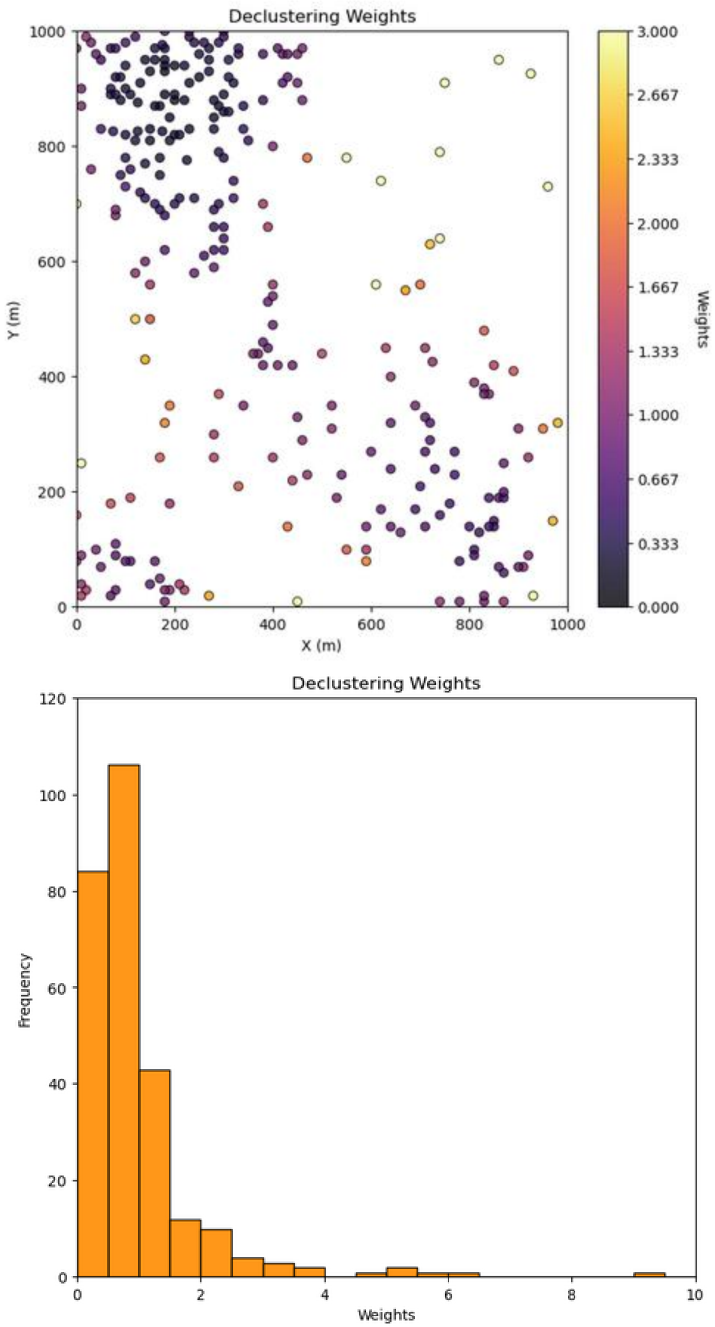
Declustering Weights

- Declustering weights
  1. 1.0 nominal weight
  2. < 1.0 reduced weight
  3. > 1.0 increased weight

- Note: some software programs assume:

$$\sum_{i}^{n} w(\mathbf{u}_i) = 1$$

then ‘nominal weight’ is $\frac{1}{n}$



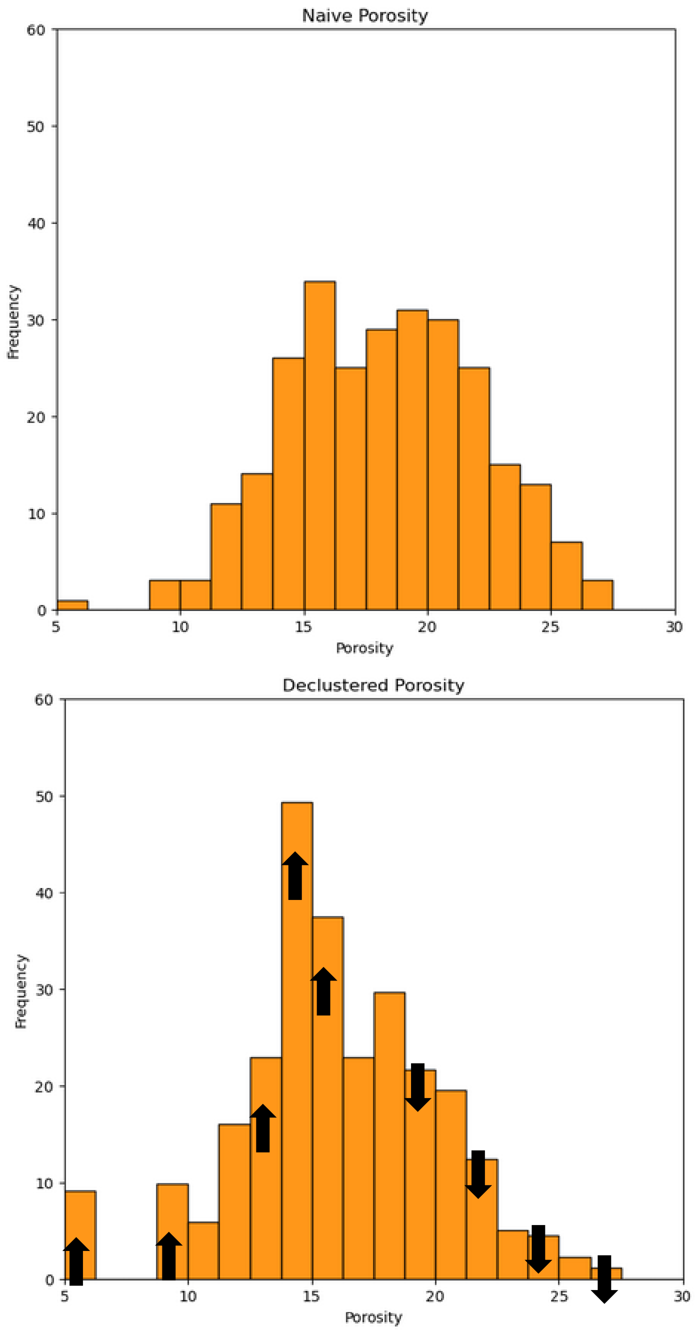
Declustering weights distribution (upper) and location map (lower), from GeostatsPy_declustering.ipynb.

Declustered Distribution

- Updated distribution with declustering weights
- Now data file / table include values and paired weights based on spatial arrangement.
- Possible to calculate any weighted statistic.
  - For example declustered mean:

$$\bar{z} = \frac{\sum_i^n w(\mathbf{u}_i) z(\mathbf{u}_i)}{\sum_i^n w(\mathbf{u}_i) = n}$$

- Python MatPlotLib hist allows for a vector of weights.



Original distribution (upper) and desclustered distribution (lower) from GeostatsPy_declustering.ipynb..

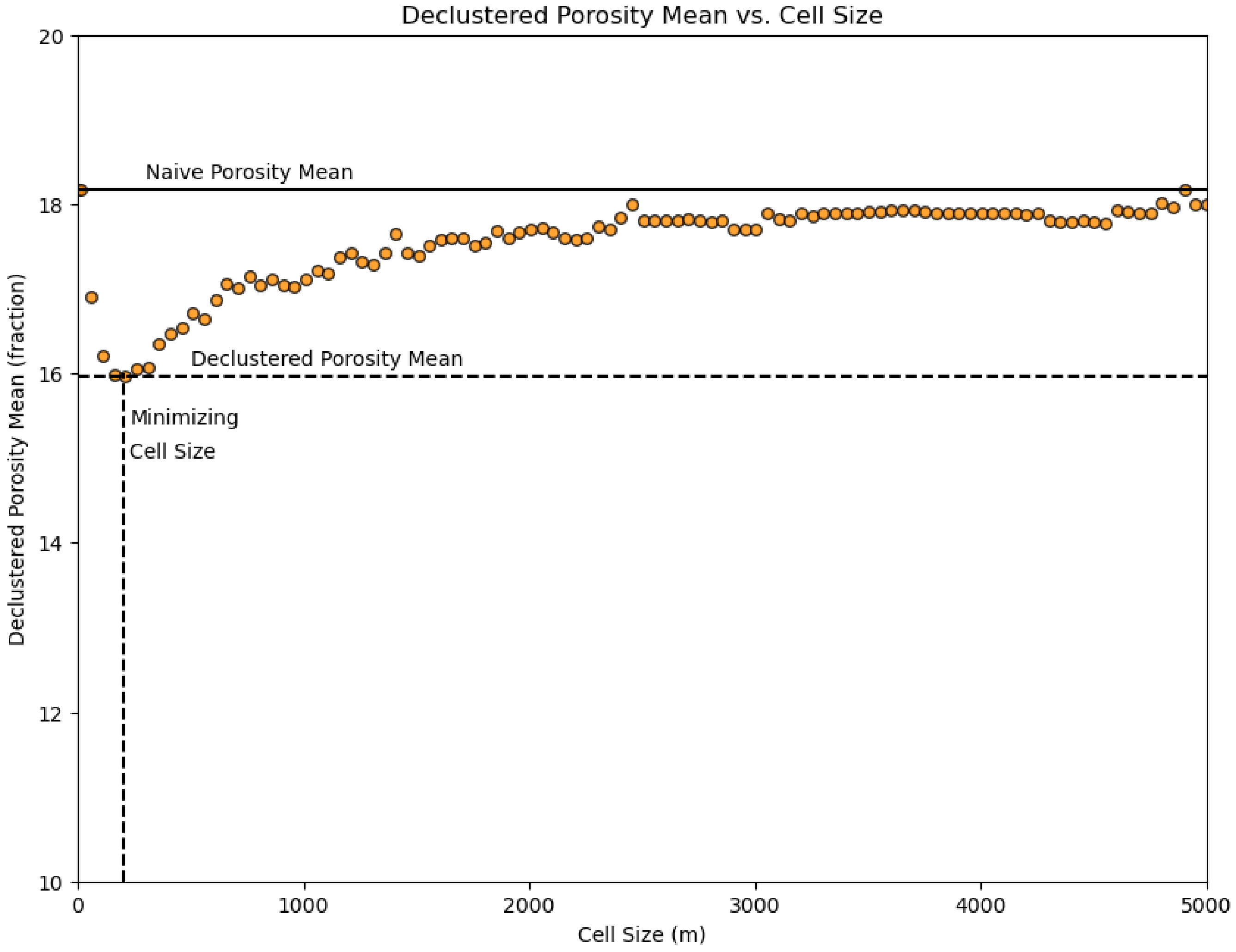
Cell Size Selection

- Plot **declustered mean** versus the **cell size** for a range of cell sizes:

- **There is no theory** that says we are looking for a minimum when the values are clustered in high values or a maximum when clustered in low values – it just seems to make sense

- The result can be very **sensitive to large scale trends** – it is often better to choose the cell size by visual inspection and some sensitivity studies

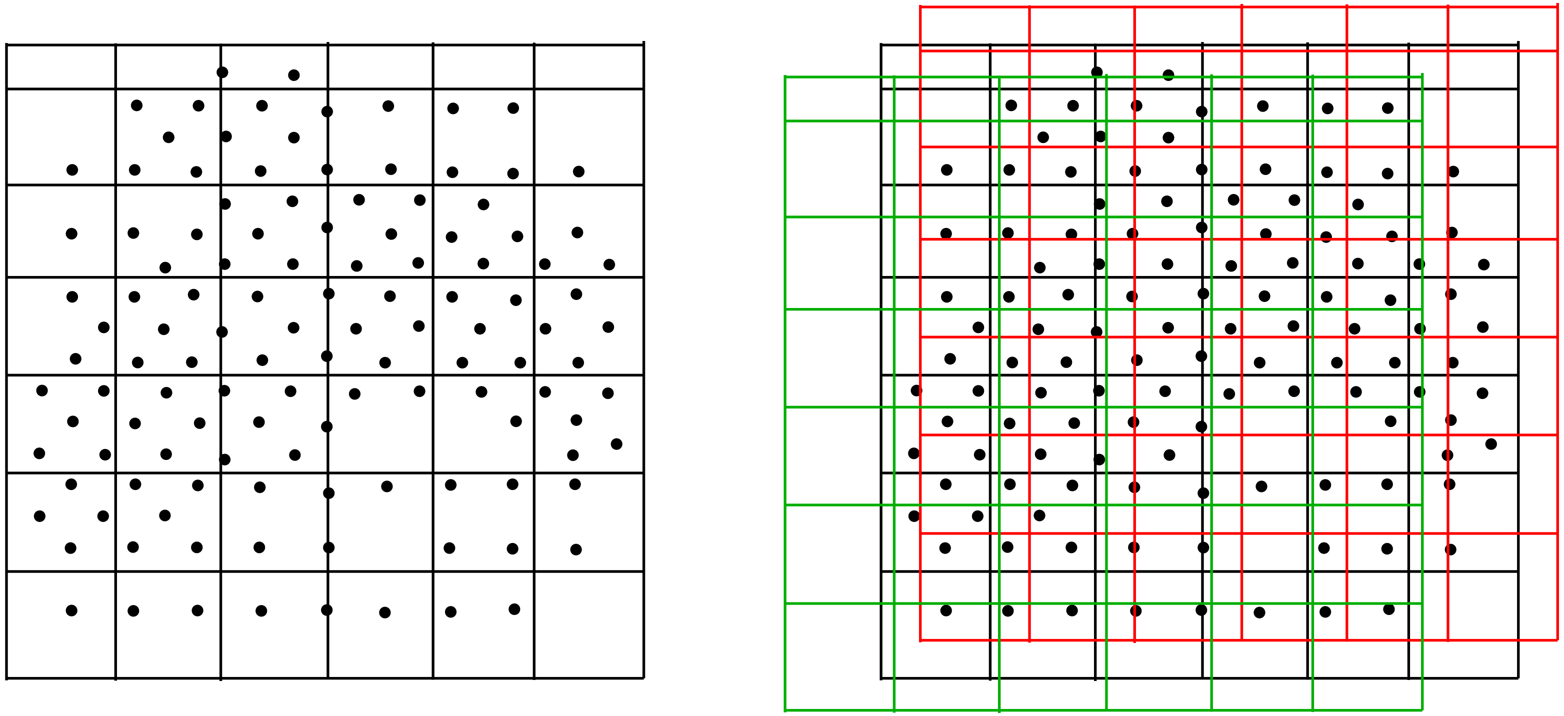
- Could choose the cell size so that there is **approximately one datum per cell in the sparsely sampled areas**, the nominal spacing



Declustered mean vs. cell size from rom GeostatsPy_declustering.ipynb.

Cell-based Declustering Offsets

- The result is sensitive to exact location of the cell mesh



Sample data and a single cell mesh (left), sample data and multiple cell meshes (right).

- This sensitivity is removed by iterativing the mesh position, calculating the weights for each and averaging the result.

Summary on Cell-based Declustering

- Sensitive to cell size choice, minimizing / maximizing declustered mean or select based on data configuration.
  - We'll use minimizing or maximizing approach in this class, by calculating the declustered mean over a wide range of cell sizes

- Removed sensitivity to exact cell mesh location by averaging over multiple cell meshes.

- Low / Little Sensitivity to Data Boundary
  - We have another method available - Polygonal Declustering

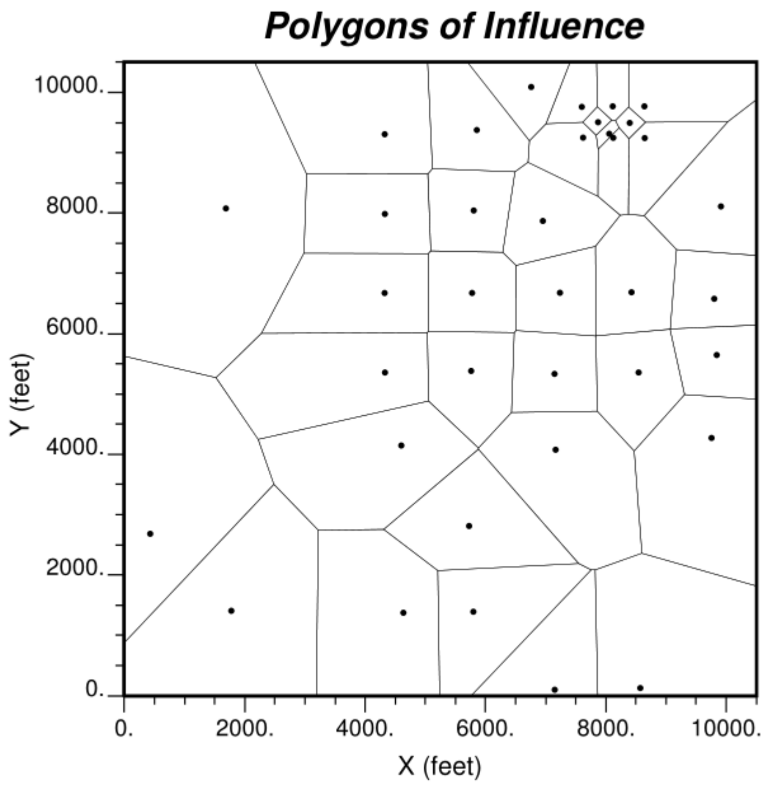
Polygonal Declustering

- Split up the area of interest with Voronoi partition.
  - Intersected perpendicular bisectors between adjacent data points
  - Segments are by nearest data point

$$w(\mathbf{u}_j) = \frac{A_j}{\sum_{j=1}^{n} A_j} \text{ for } \sum_{j=1}^{n} w(\mathbf{u}_j) = 1$$

$$w(\mathbf{u}_j) = n\frac{A_j}{\sum_{j=1}^{n} A_j} \text{ for } \sum_{j=1}^{n} w(\mathbf{u}_j) = n$$

  - This method is sensitive to boundary
  - Commonly applied in a variety of scientific fields for weighted averages of spatial phemenon with irregular sampling.



Sample data and polygons of influence.

Declustered Statistics

We apply the declustering weights to calculate all required statistics.

- The sample mean:

$$\hat{m} = \frac{\sum_{j=1}^{N} w(\mathbf{u}_j) z(\mathbf{u}_j)}{\sum_{j=1}^{N} w(\mathbf{u}_j)}$$

- The sample variance:

$$s^2 = \frac{1}{\sum_{j=1}^{n} w(\mathbf{u}_j) - 1} \sum_{j=1}^{n} w(\mathbf{u}_j)\left(z(\mathbf{u}_j) - \hat{m}\right)^2, \text{ where } \sum_{j=1}^{n} w(\mathbf{u}_j) = n$$

- The covariance:

$$\boldsymbol{C}_{\boldsymbol{x},\boldsymbol{y}} = \frac{1}{\sum_{j=1}^{n} w(\mathbf{u}_j)} \sum_{j=1}^{n} w(\mathbf{u}_j)\left(x(\mathbf{u}_j) - \bar{\boldsymbol{x}}\right)\left(y(\mathbf{u}_j) - \bar{y}\right)$$

- The entire CDF:
  - If $\sum_{j=1}^{n} w(\mathbf{u}_j) = 1$, then $F_Z(z) \approx \sum_{j=1}^{n(Z<z)} w(\mathbf{u}_j)$

    the sum of the weights of all data $z(\mathbf{u}_j) < z$

- Statistics from raw spatial data with no effort to correct for bias are called **naïve statistics**, e.g., naïve mean, naïve standard deviation etc.

Excel Declustering
Hands On

Well-documented Excel example of declustering (and debiasing).

Cell-based Declustering By-Hand in Excel, Michael Pyrcz, University of Texas at Austin, @GeostatsGuy on Twitter

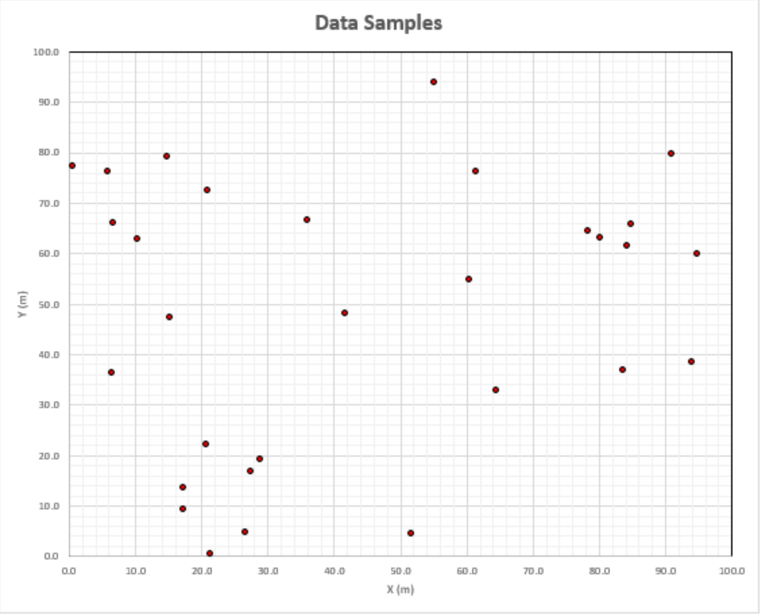
About: This demonstration includes cell-based declustering applied on a random sample set from a truth model.
Dataset: The truth model is a simple 2D geometric function with the high in the center of the area of interest (at 50m,50m).
Objective: Provide an opportunity to experiment with declustering for a variety of data configurations.

cell size	20
Lo	15

sample mean	60.5
dec. mean	63.9
true mean	61.7

	x	y	ix	iy	z	wtx	wty	wt	z x wt
1	28.7	19.4	29	20	63.4	2	1	0.50	31.7
2	6.6	66.1	7	67	53.8	1	4	0.40	21.5
3	17.0	9.5	18	10	48.8	1	1	1.00	48.8
4	83.5	37.2	84	38	63.9	5	2	1.00	63.9
5	54.9	94.1	55	95	54.7	3	5	2.00	109.4
6	84.6	66.0	85	67	61.1	5	4	0.50	30.5
7	17.0	13.8	18	14	51.8	1	1	1.00	51.8
8	79.9	63.2	80	64	66.9	4	4	0.67	44.6
9	0.5	77.4	1	78	43.6	1	4	0.40	17.4
10	21.3	0.6	22	1	43.6	2	1	0.50	21.8
11	93.9	38.6	94	39	54.6	5	2	1.00	54.6
12	84.0	61.6	85	62	63.0	5	4	0.50	31.5
13	14.7	79.3	15	80	53.9	1	4	0.40	21.6
14	26.5	5.0	27	6	50.4	2	1	0.50	25.2
15	6.4	36.4	7	37	55.1	1	2	2.00	110.2
16	15.1	47.4	16	48	65.9	1	3	2.00	131.9
17	64.3	32.9	65	33	77.3	4	2	2.00	154.7
18	27.2	16.9	28	17	60.3	2	1	0.50	30.2
19	94.7	60.1	95	61	53.7	5	4	0.50	26.8
20	78.2	64.6	79	65	67.4	4	4	0.67	44.9
21	20.7	22.3	21	23	60.4	2	2	2.00	120.8
22	20.8	72.7	21	73	63.0	2	4	1.00	63.0
23	51.6	4.6	52	5	55.0	3	1	2.00	109.9
24	35.8	66.8	36	67	78.0	2	4	1.00	78.0
25	61.3	76.5	62	77	70.5	4	4	0.67	47.0
26	90.8	79.9	91	80	49.2	5	4	0.50	24.6
27	41.5	48.3	42	49	91.9	3	3	2.00	183.9
28	10.2	62.9	11	63	58.9	1	4	0.40	23.6
29	60.3	55.1	61	56	87.5	4	3	2.00	174.9
30	5.8	76.3	6	77	48.4	1	4	0.40	19.4

Sum Wt. 30.00



https://github.com/GeostatsGuy/ExcelNumericalDemos/blob/master/Declustering_Debias_Demo.xlsx **or** https://git.io/fxA3O

Python Interactive Demonstration

Here’s interactive cell-based declustering in Python.

Interactive Cell-based Declustering Demostration

- select the cell size and number of cell mesh offsets and visualize the declustering method

Michael Pyrcz, Associate Professor, University of Texas at Austin

Twitter | *GitHub* | *Website* | *GoogleScholar* | *Book* | *YouTube* | *LinkedIn* | *GeostatsPy*

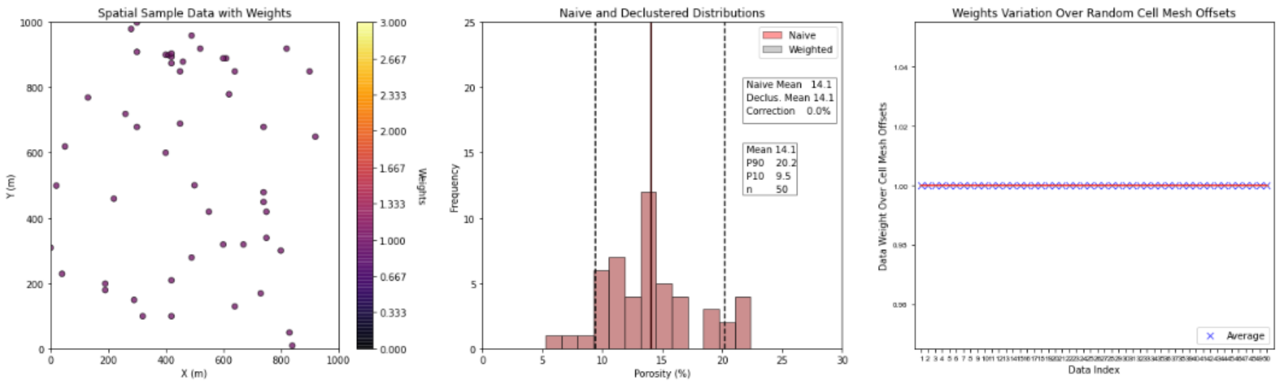
The Inputs

- **Cell Size**: the size the of the cells in the mesh, **Number of Offsets**: number of cell mesh offsets to average to calculate the data weights

```
display(ui, interactive_plot)                          # display the interactive plot
```

Cell-based Declustering, Michael Pyrcz, Associate Professor, The University of Texas at Austin

Cell Size 5.00 Number of Offsets 5 Show Grid



Interactive desclutering in Python, file is interactive_declustering.ipynb

Python GeostatsPy Declustering Demo

Here's a demonstration of cell-based declustering in Python.

GeostatsPy: Cell-based Declustering with Basic Univariate Statistics and Distribution Representativity for Subsurface Data Analytics in Python

Michael Pyrcz, Associate Professor, University of Texas at Austin

Twitter | *GitHub* | *Website* | *GoogleScholar* | *Book* | *YouTube* | *LinkedIn* | *GeostatsPy*

PGE 383 Exercise: Basic Univariate Summary Statistics and Data Distribution Representativity Plotting in Python with GeostatsPy

Here's a simple workflow with some basic univariate statistics and distribution representativity. This should help you get started data declustering to address spatial sampling bias.

Geostatistical Sampling Representativity

In general, we should assume that all spatial data that we work with is biased.

Source of Spatial Sampling Bias

Data is collected to answer questions:

- how far does the contaminant plume extend? – sample peripheries
- where is the fault? – drill based on seismic interpretation
- what is the highest mineral grade? – sample the best part
- who far does the reservoir extend? – offset drilling and to maximize NPV directly:
- maximize production rates

Random Sampling: when every item in the population has a equal chance of being chosen. Selection of every item is independent of every other selection. Is random sampling sufficient for subsurface? Is it available?

- it is not usually available, would not be economic
- data is collected answer questions
  - how large is the reservoir, what is the thickest part of the reservoir
- and wells are located to maximize future production

Declustering demonstration in Python, file is GeostatsPy_Declustering.ipynb.

Comments on Cell Declustering

- Perform an areal 2-D declustering when the wells are vertical or near vertical
  - The problem simplifies to 2D only

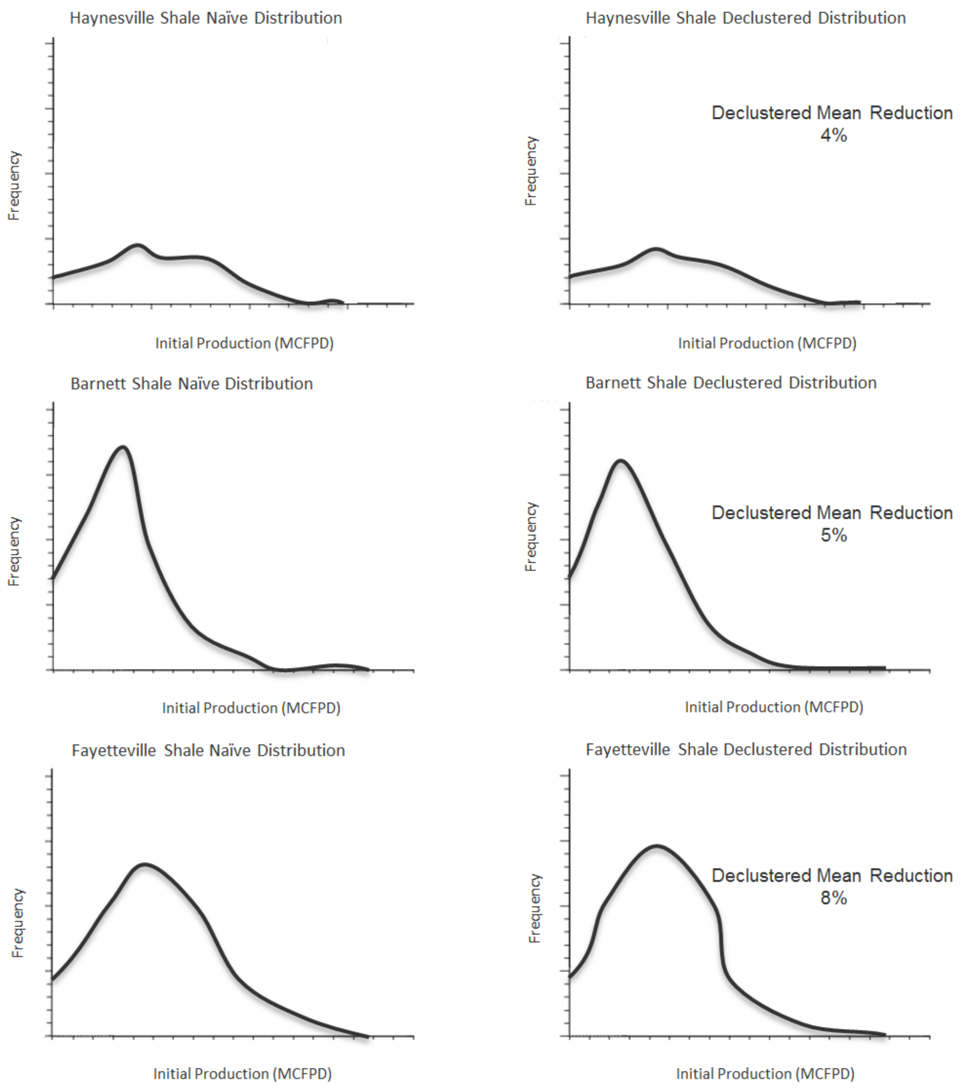
- Consider 3-D declustering when there are horizontal or highly deviated wells

- The shape of the cells depends on the geometric configuration of the data
  - adjust the shape of the cells to conform to major directions of preferential sampling

Declustering in Unconventionals

Representative Statistics

- Compiled IP datasets for domestic shale plays
  - Filtered datasets to reduce influence of completions
- Representativity an issue even with large datasets and relatively good coverage
  - Observed changes in naïve to declustered means of 4 – 8%



Naïve and declustered distributions from cell-based declustering (modified from Pyrcz et al, 2017)

PGE 338 Data Analytics and Geostatistics

Lecture 9b: Spatial Bias

Lecture outline . . .

- Debiasing with Secondary Data

This entire Section is Reference Only. Not on Exam.

Michael Pyrcz, The University of Texas at Austin

Introduction

General Concepts

Univariate

Bivariate

- Correlation
- Regression
- Model Checking

Time Series Analysis

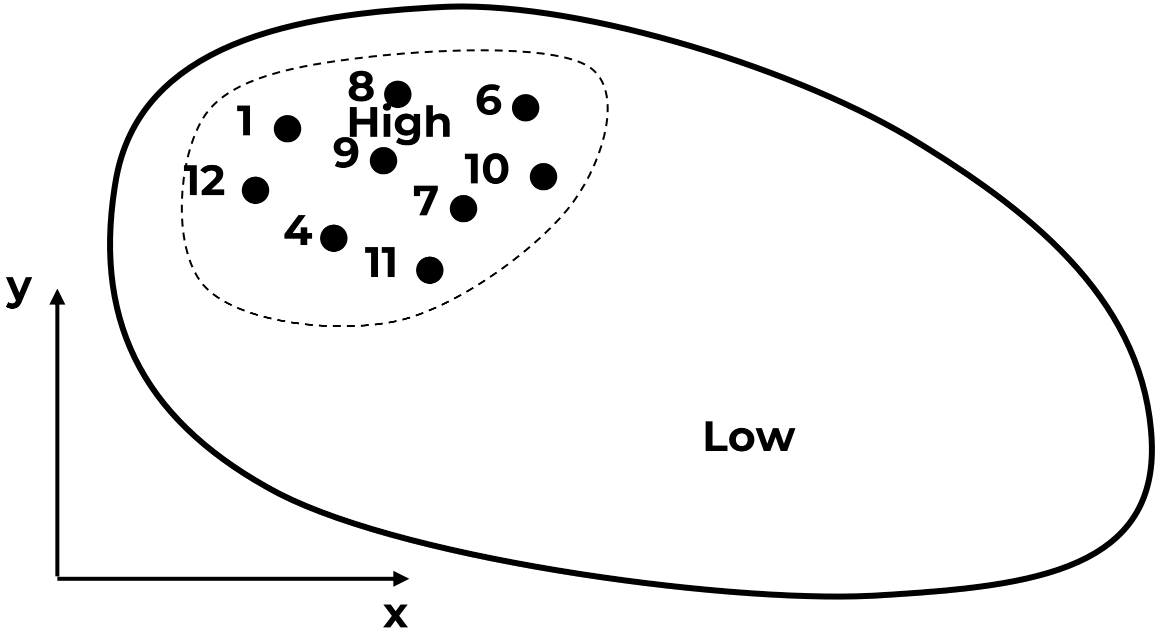
Spatial Analysis

Machine Learning

Uncertainty Analysis

Missing Samples

Let's once again make an estimate for an Area / Volume of Interest:

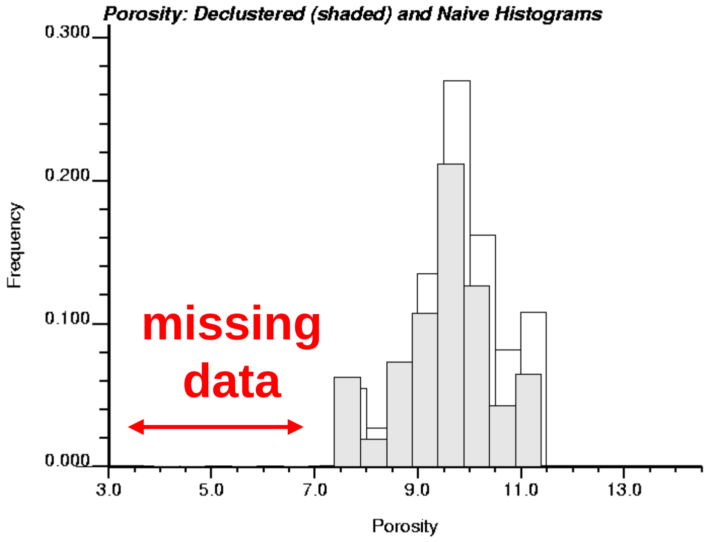


- What if we never sampled the lowest part of the area of interest?

Reference Only. Not on Exam.

Comments on Spatial Debiasing

What do we do when there are too few data, or the data are not representative?



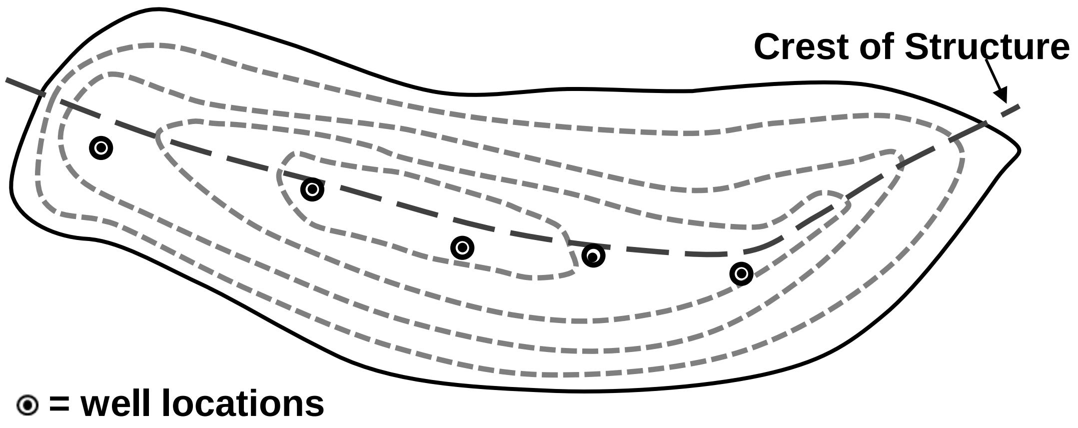
Delcustering is not possible when part of the distribution is missing.

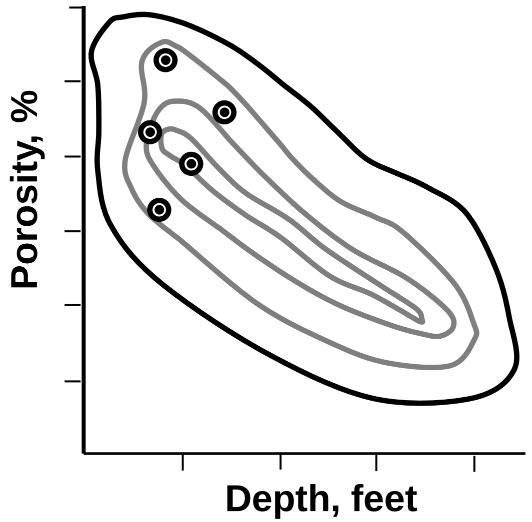
- Declustering can only change the data weights / change the heights of the histogram bars, it can't fill in missing data / add new bars!
  - e.g. porosity values < 7.5% are not sampled!
- We need another method, spatial data debiasing.

Reference Only. Not on Exam.

Comments on Spatial Debiasing

- What do we do when there are too few data, or the data are not representative?
- Nothing, unless there is some secondary information





- How can we use this soft data to correct histogram?
  - Extrapolate porosity data using the full depth distribution

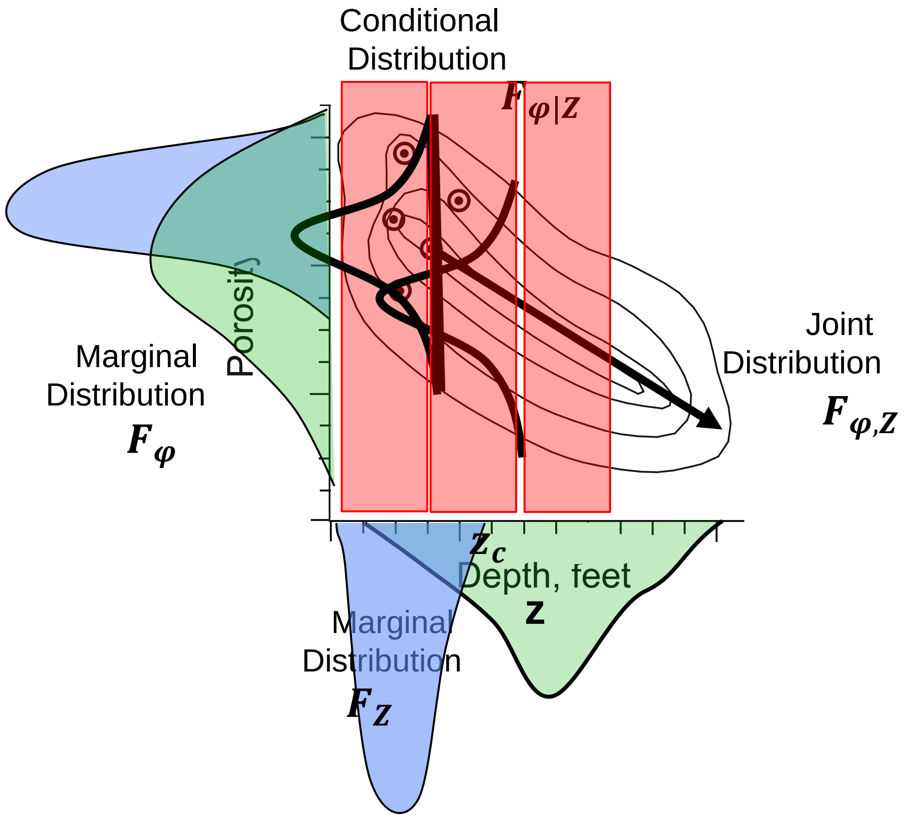
Reference Only. Not on Exam.

Image modified from Pyrcz and Deutsch (2014)

Calibration Approach with Conditional Distributions

- We model the bivariate relationship between porosity and depth.

- If we assume a linear compaction curve then this is a simple linear extrapolation of the condition distribution, $F_{\varphi|Z}$



Reference Only. Not on Exam.

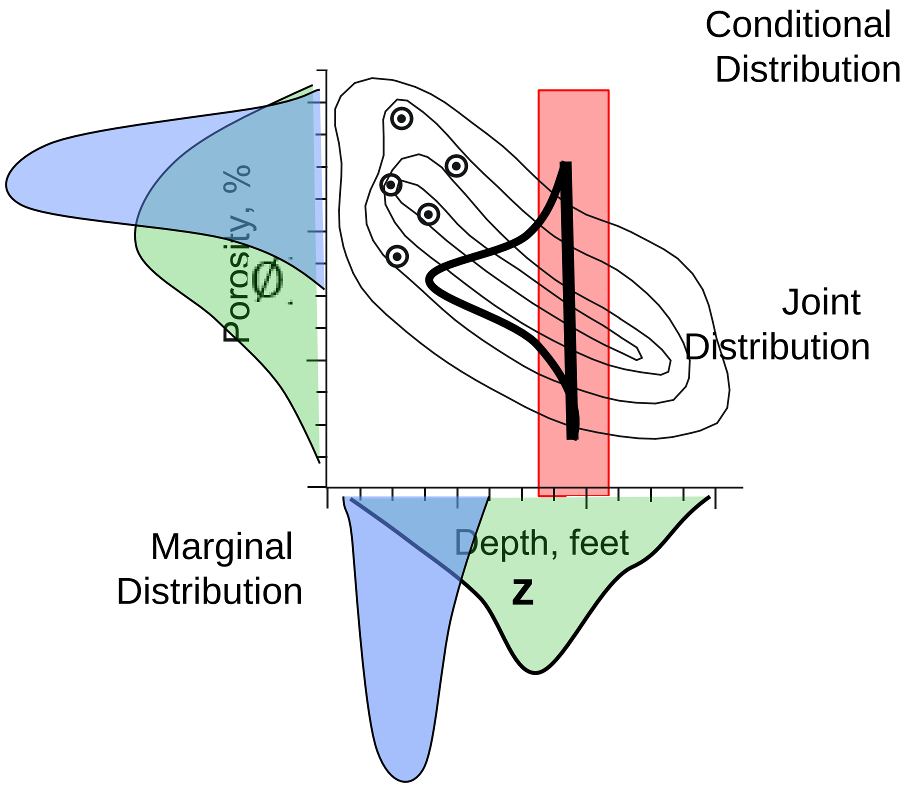
Calibration Approach with Conditional Distributions

- Empirical spatial debiasing approach for porosity, ∅ | depth, *Z* :
  - map a secondary variable *Z* at all locations, we have full Z distribution
  - develop a bivariate relationship between *Z* and ∅ variables
  - generate a distribution of ∅ by combining conditional distributions

- The marginal PDF of porosity may be found by:

  $f_\emptyset(\emptyset) = \int_z f_{\emptyset\,|\,z} \cdot f_z \, dz$, integrate over the conditional distributions.

- The calibration (modeling of the permeability and porosity distribution is critical



Reference Only. Not on Exam.

PGE 338 Data Analytics and Geostatistics

Lecture 9b: Spatial Bias

Lecture outline . . .

- Sampling Bias
- Declustering
- Debiasing with Secondary Data

Michael Pyrcz, The University of Texas at Austin