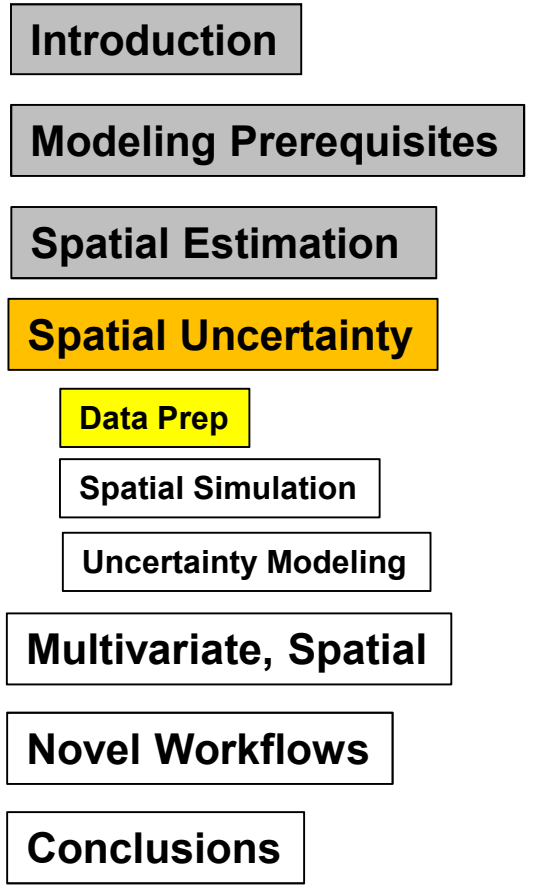
Data Analytics and Geostatistics:
Data Preparation

Lecture outline . . .

- **Sampling Limitations**
- **Declustering**
- **Quantifying Uncertainty**



Instructor: Michael Pyrcz, the University of Texas at Austin

Data Analytics and Geostatistics:
Data Preparation

Lecture outline . . .

- **Sampling Limitations**

Instructor: Michael Pyrcz, the University of Texas at Austin

One Source of Bias
Data Collection

Data is collected to answer questions:

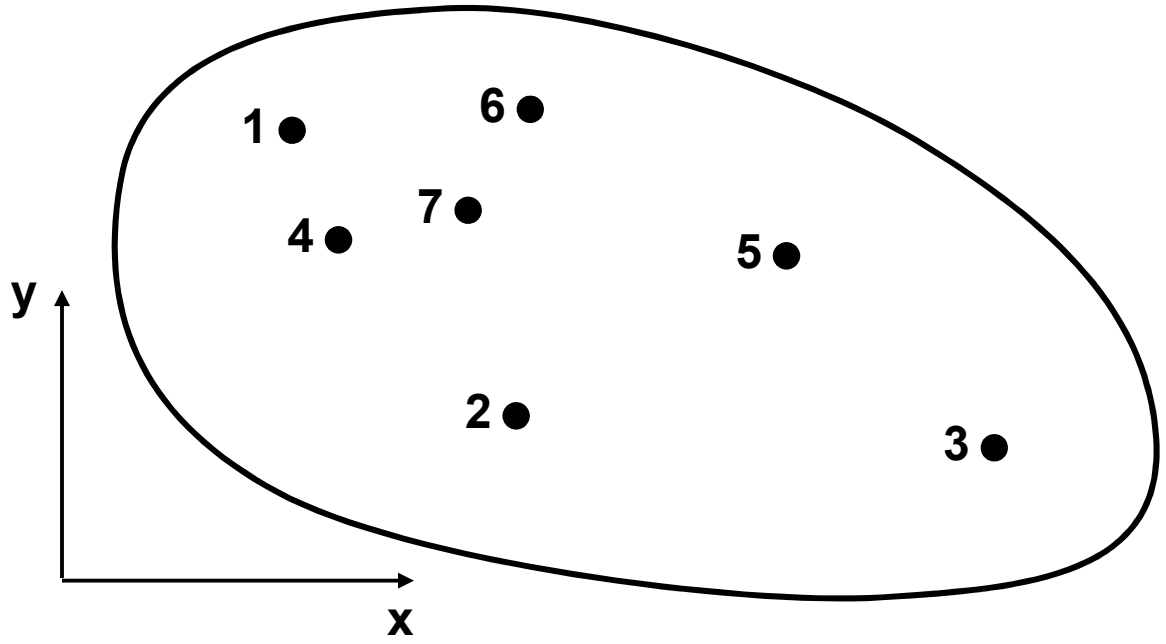
- how far does the contaminant plume extend? – *sample peripheries*
- where is the fault? – *drill based on seismic interpretation*
- what is the highest mineral grade? – *sample the best part*
- who far does the reservoir extend? – *offset drilling*

and to maximize NPV directly:

- maximize production rates

Representativity

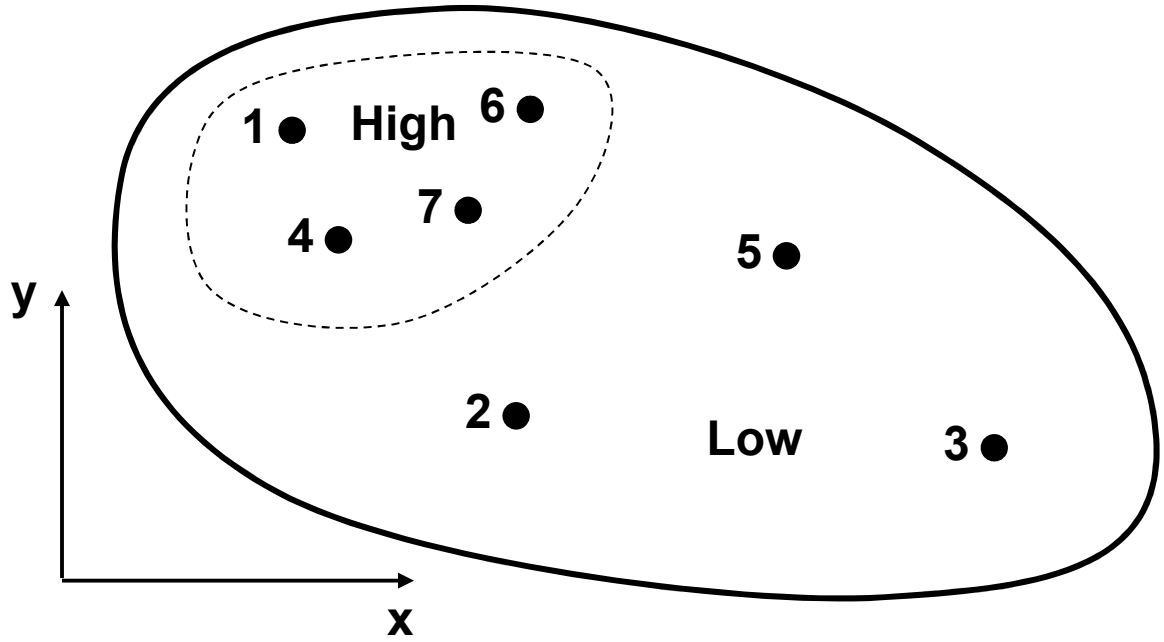
The concern is when we attempt to make an estimate:



e.g. the average porosity to calculate OIP

Representativity

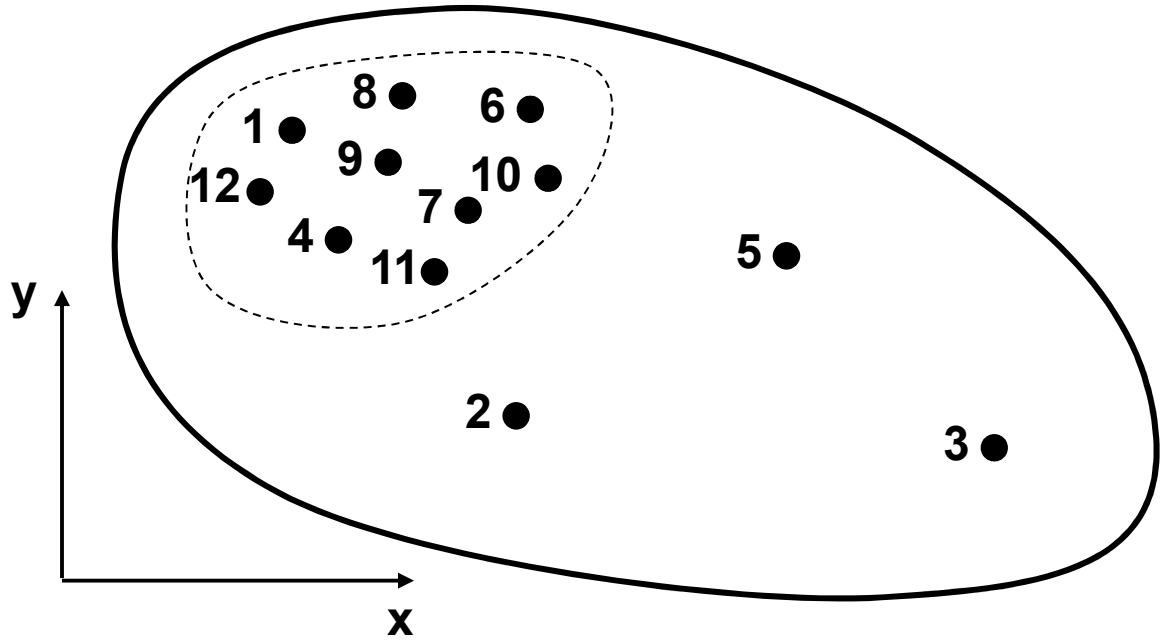
The concern is when we attempt to make an estimate:



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

Representativity

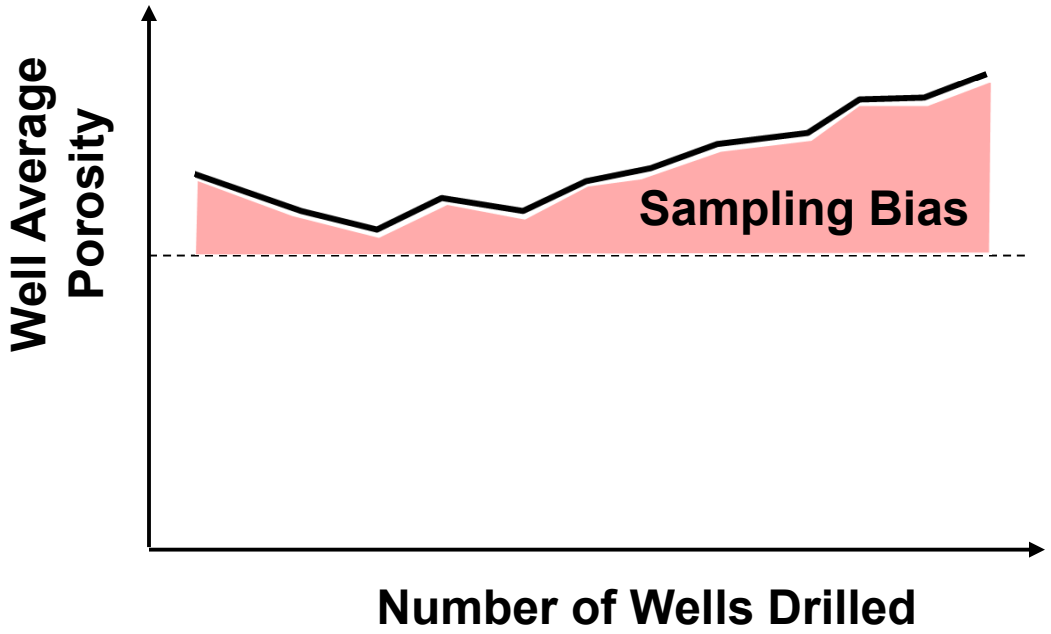
The concern is when we attempt to make an estimate:



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

Representativity

The concern is when we attempt to make an estimate:



We need to mitigate this bias.

Drilling representatively? Why do we drill in a biased manner?

(Geo)statistics Sampling Representatively

How Would We Sample for Representativity?

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?

Regular Sampling: when samples are taken at regular intervals (equally spaced).

- Less reliable than random sampling.
- Warning: May resonate with some unsuspected environmental variable.

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 in the Gulf of Mexico at $150M per well?

We should not change current sampling methods as they result in best economics, we should address sampling bias in the data.

Never use raw spatial data without access sampling bias / correcting.

(Geo)statistics Sampling Bias

Example of Sampling Bias:

1. Well's drilled in part of reservoir identified to have the greatest thickness in seismic.
2. Core extracted from the well bore in the location estimated to have the best reservoir.
3. Core plugs extracted from whole cores for porosity / permeability analysis avoiding shales.

Routine core analysis from https://www.rigzone.com/training/insight.asp?insight_id=325.

(Geo)statistics Sampling Bias

There are also limits to our data collection:

- accessibility to the sample – obstruction, reliable drilling, subsalt imaging
- inability to process the sample – may not be able to recover shale core samples
- can't run permeability evaluation on low permeability rock

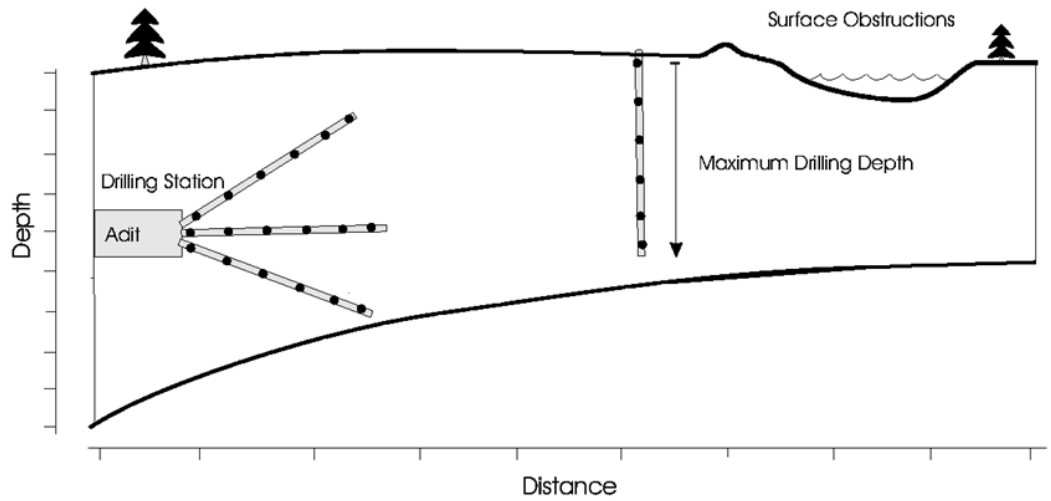


Image from Pyrcz and Deutsch (2003) http://gaa.org.au/pdf/DeclusterDebias-CCG.pdf

Spatially Clustered Data

Data are rarely collected for their statistical representativity:

- Wells are drilled in areas with the greatest probability of high production
- Horizontal wells target stratigraphic zones of interest (high pay)
- Core are taken preferentially from good quality reservoir rock
- These data collection practices should not be changed:
  - best economics
  - most data in the most important locations

Solutions to Biased Sampling

- 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!

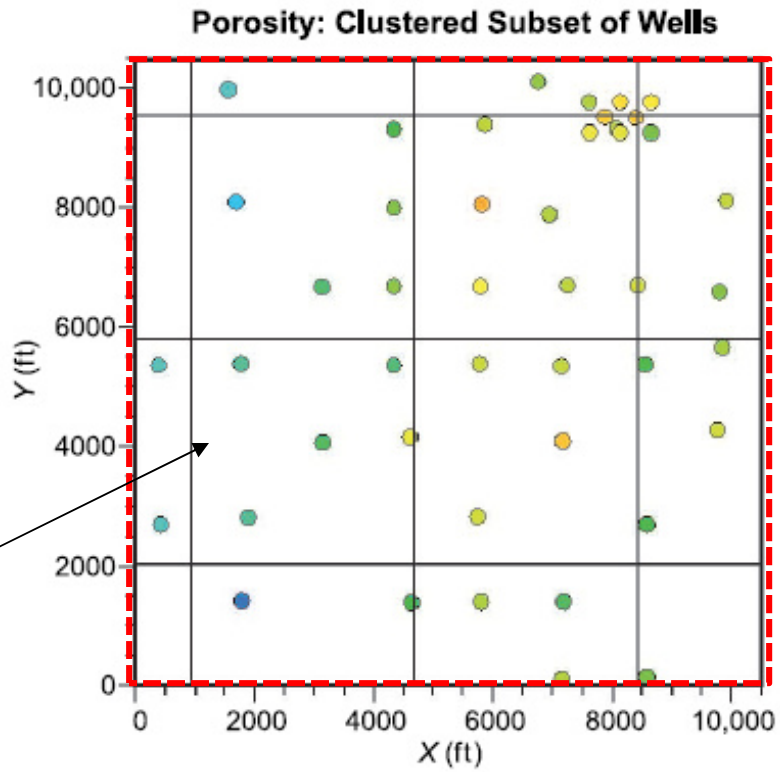
1. **Mapping**: and summarizing average **over map**
2. **Use of Regions**: **to pool and use statistical** over volumes of high / low reservoir quality (e.g. facies)
3. **Declustering techniques** **assign each datum a weight** based on closeness to surrounding data
   - $w_i, i = 1, \ldots, n$ (weights are greater than 0 and sum to n)
   - Histogram and cumulative histogram use $w_i, i = 1, \ldots, n$ instead of equal weighted, $w_i = 1.0$.
4. **Debiasing techniques** derive an entirely new distribution based on a secondary data source such as geophysical measurements or expert interpretation

(Geo)statistics Goal of Sampling and Statistics Example

Addressing Bias:

Would it be fair to calculate the average of these wells and to apply that as an average for this area of interest?

What is the average porosity over this reservoir?



Porosity sample data for an example reservoir (Pyrcz and Deutsch, 2014).

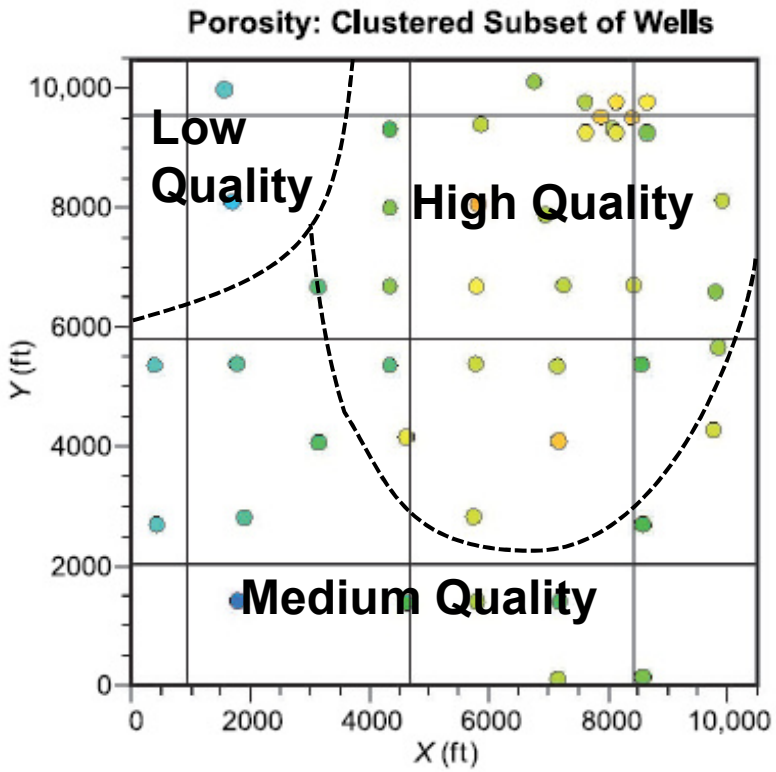
(Geo)statistics Sampling Bias

Addressing Bias with Regions:

Would it be fair to calculate the average of these wells and to apply that as an average for this area of interest?

Break model up into subsets.

- Avoid densely sampled high quality reservoir inflating average over the entire reservoir



Porosity sample data for an example reservoir (Pyrcz and Deutsch, 2014).

(Geo)statistics Sampling Bias

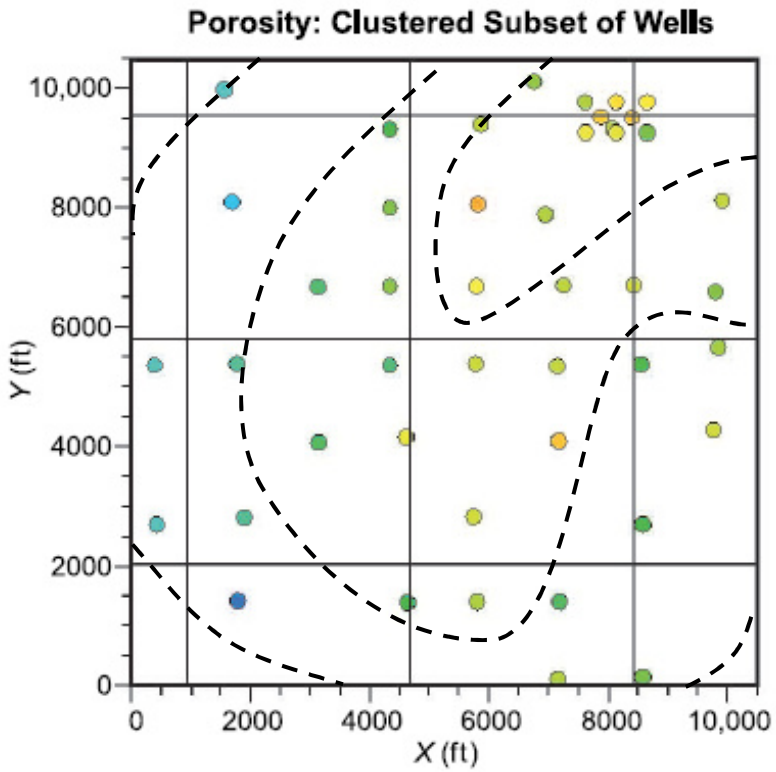
Addressing Bias:

Would it be fair to calculate the average of these wells and to apply that as an average for this area of interest?

Build a map of the property of interest.

Calculate the average of the map

- Avoid densely sampled high

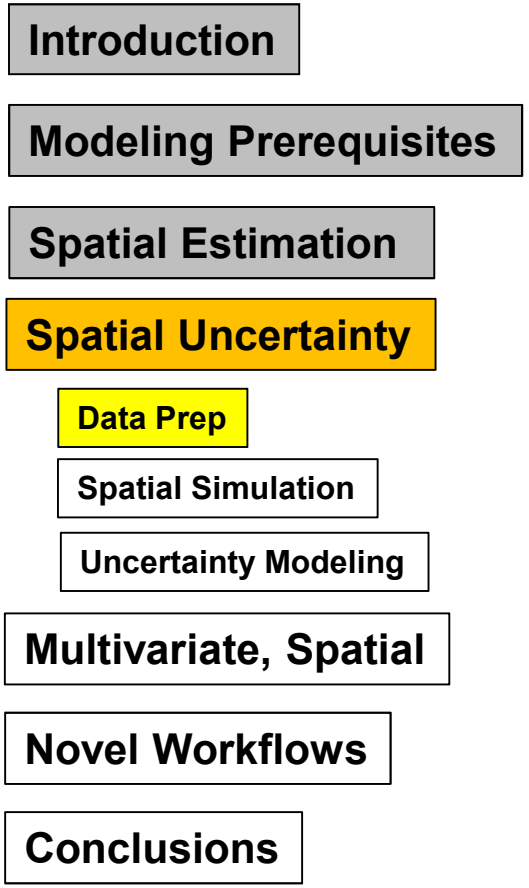


Porosity sample data for an example reservoir (Pyrcz and Deutsch, 2014).

Data Analytics and Geostatistics:
Data Preparation

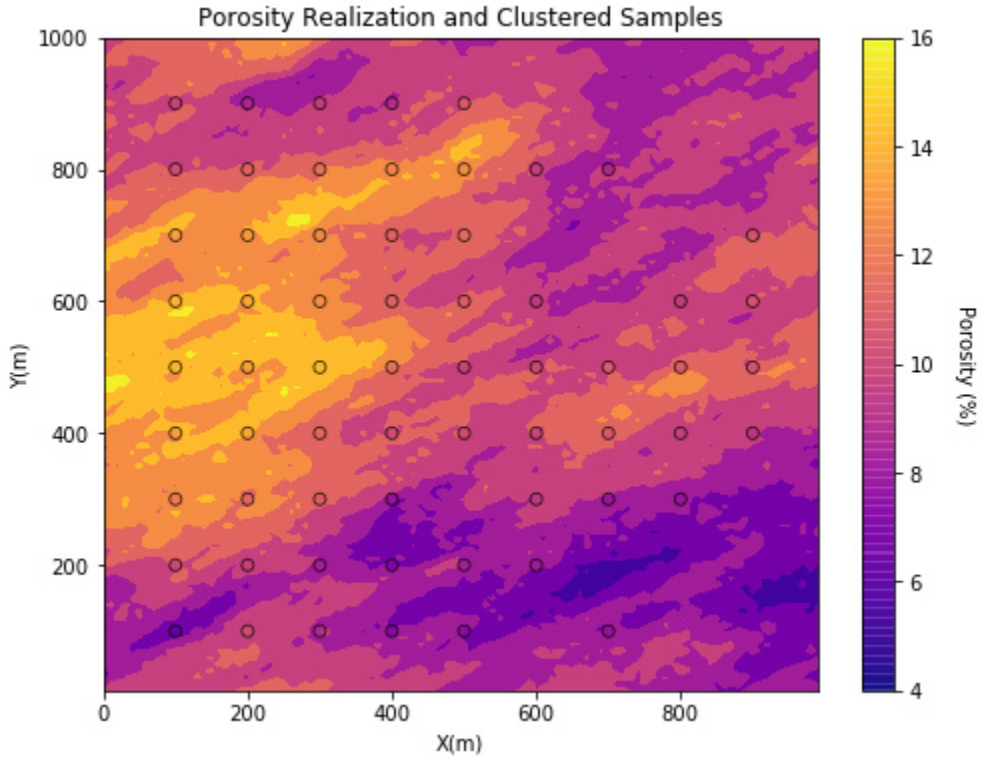
Lecture outline . . .

- **Declustering**



Instructor: Michael Pyrcz, the University of Texas at Austin

Spatially Clustered Data Example



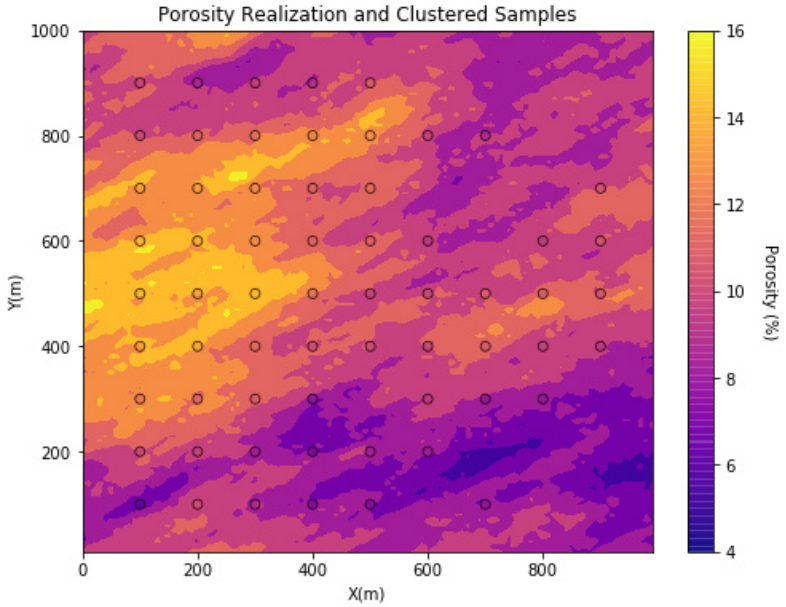
What is wrong with this sample set?

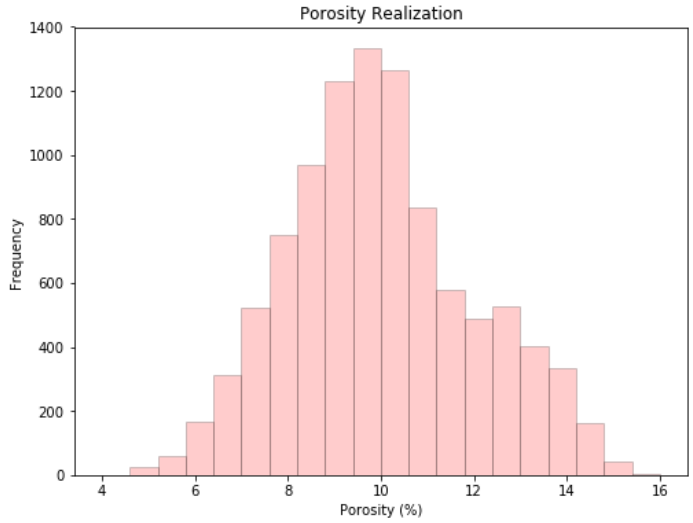
Spatially Clustered Data

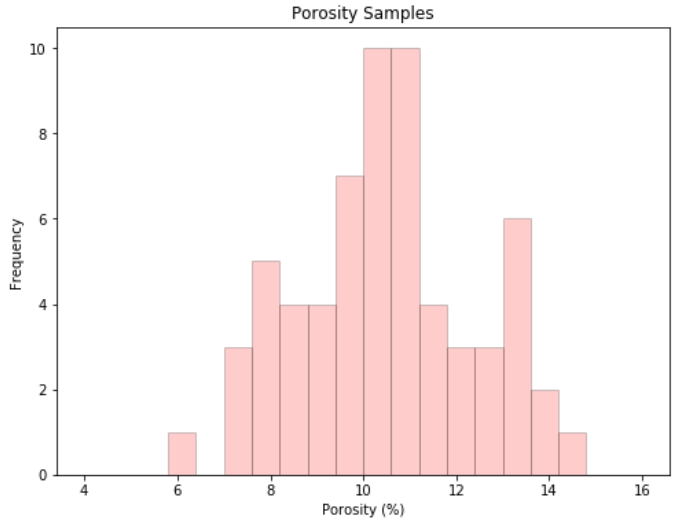
- There is a need, however, to adjust the histograms and summary statistics to be representative of the entire volume of interest.
- Declustering techniques assign each datum a weight based on closeness to surrounding data
  - $w_i, i = 1, \ldots, n$ (weights are greater than 0 and sum to n)
  - Histogram and cumulative histogram use $w_i, i = 1, \ldots, n$ instead of equal weighted, $w_i = 1.0$.
- Debiasing techniques derive an entirely new distribution based on a secondary data source such as geophysical measurements or expert interpretation

Spatially Clustered Data

- Location map of 64 wells. with truth model.
- See the error between the samples and the underlying truth model.



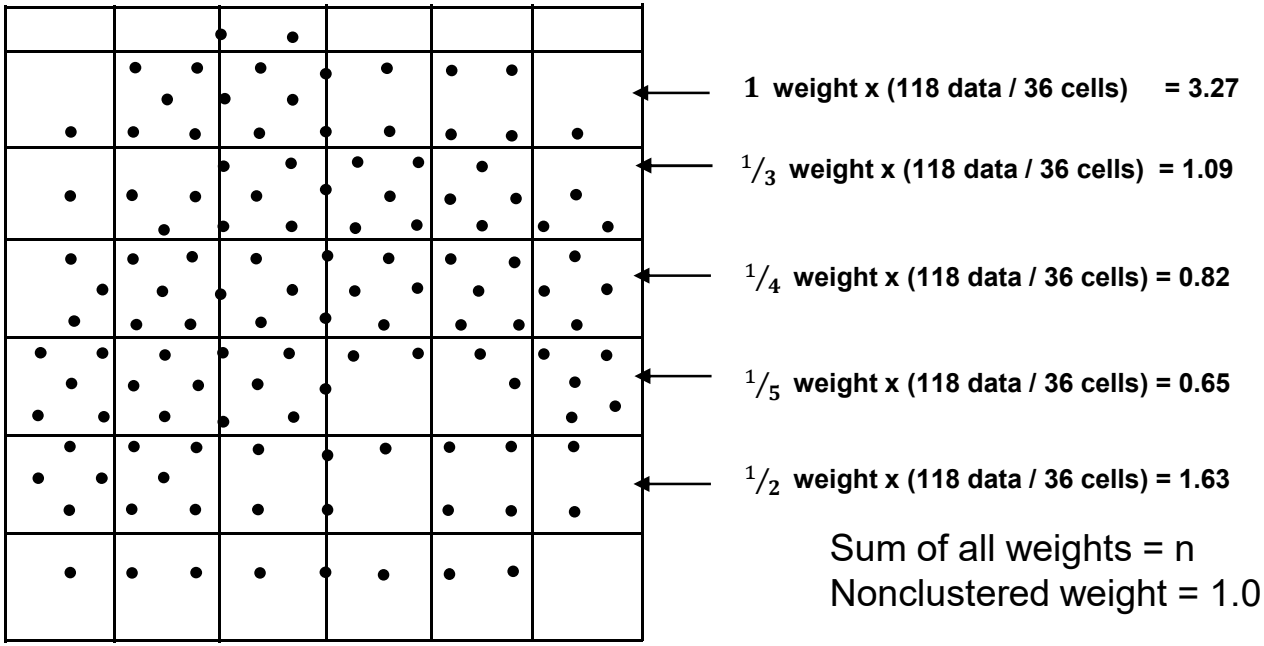




Truth Mean = 10.0 , Clustered Sample Mean = 10.48 , Error = 4.8 %

Cell Declustering

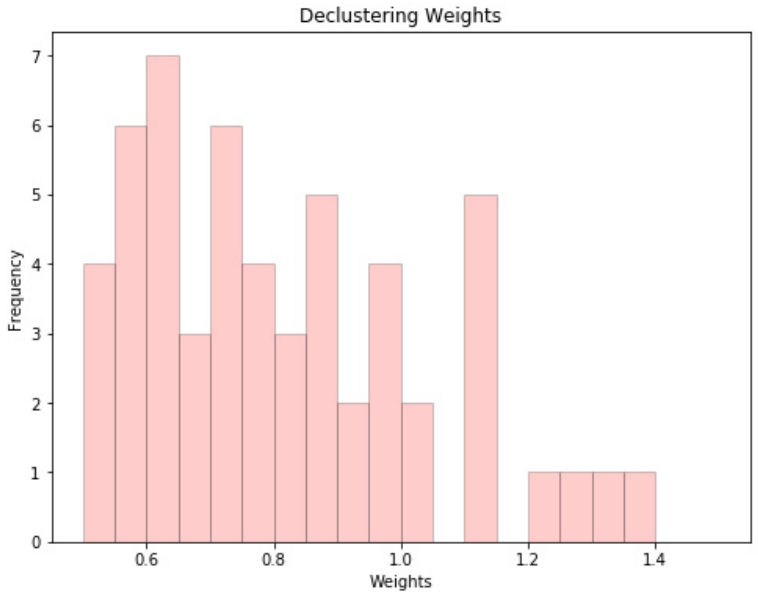
- *Cell Declustering*, is robust in 3-D and when the limits are poorly defined:
  - divide the volume of interest into a grid of cells *l=1,…,L*
  - count the occupied cells L_o and the number in each cell n_l, *l=1,…,* L_o
  - weight inversely by number in cell (standardize by L_o)

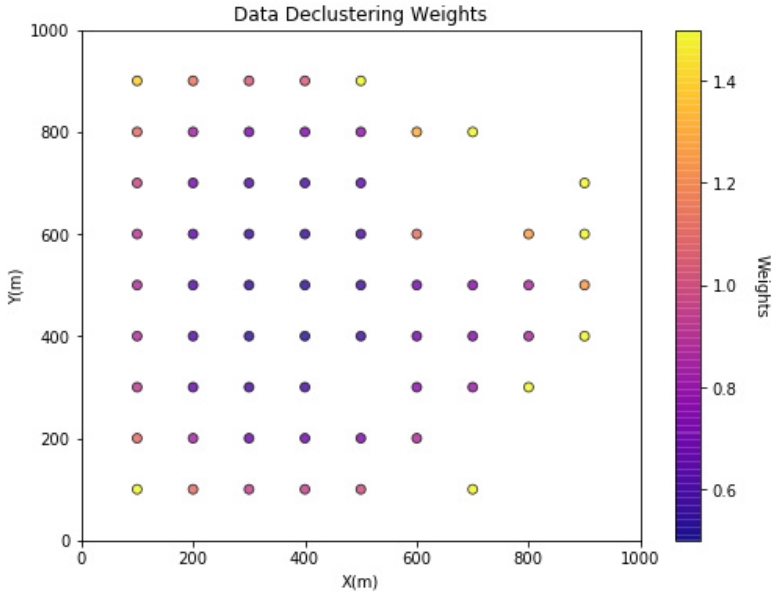


- The issue, of course, is how to choose the cell size…

Declustering Weights

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

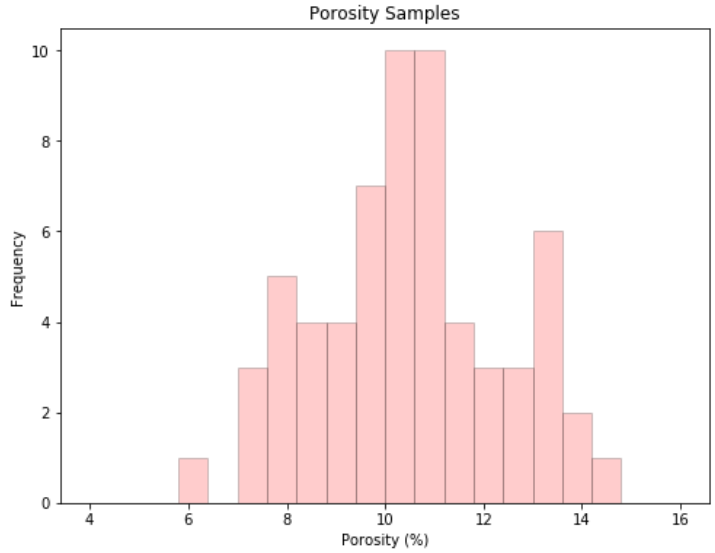




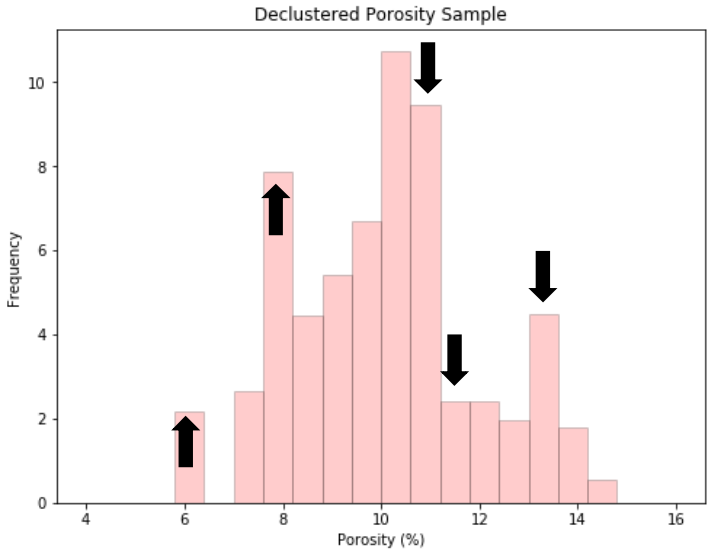
Declustered Distribution

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

$$\bar{z} = \frac{\sum_i^n w_i z_i}{\sum_i^n w_i = n}$$



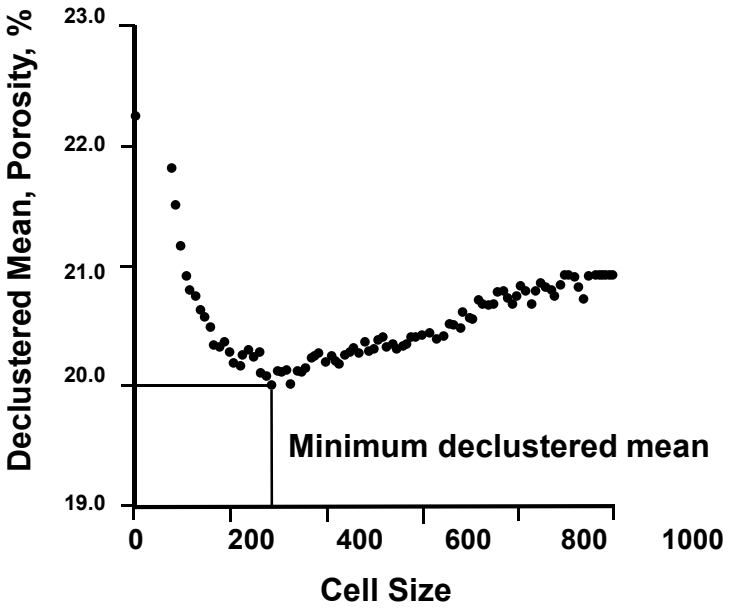
Truth Mean = 10.0 , Clustered Sample Mean = 10.48 , Error = 4.8 %

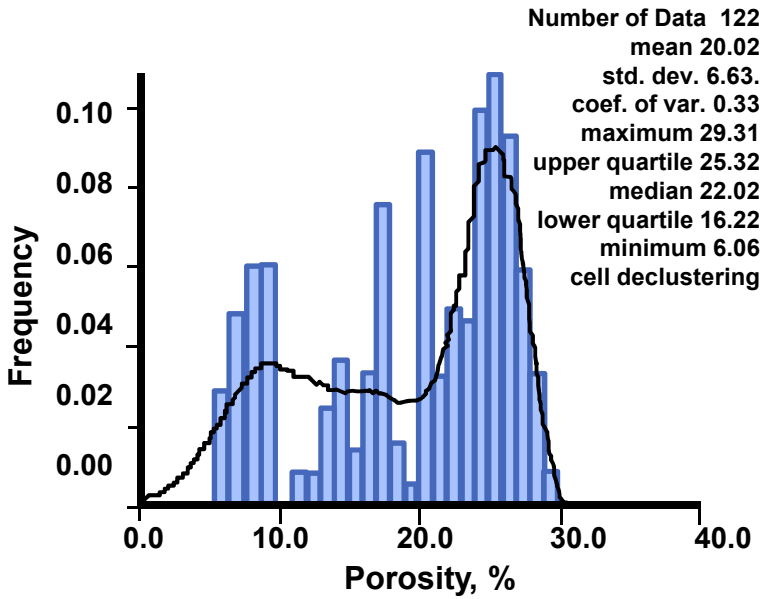


Truth Mean = 10.0 , Clustered Sample Mean = 10.48 , Error = 5.0 %
Declustered Mean = 10.07 , Error = 1.0 %

The Cell Size

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





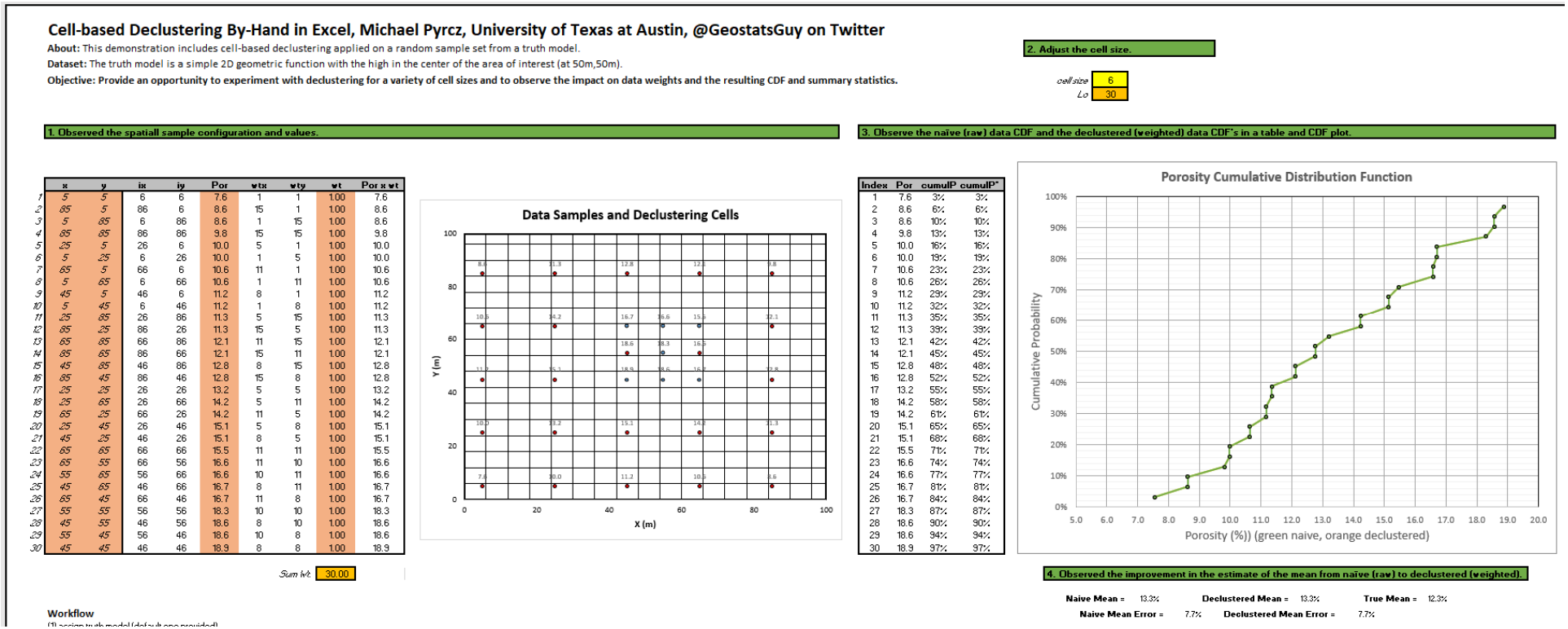
- There is no theory that says we are looking for a minimum when the values are clustered in high values or a maximum when clustered in low values – it just seems to make sense
- The result can be very sensitive to large scale trends – it is often better to choose the cell size be visual inspection and some sensitivity studies

Declustering Hands-on

Here's an opportunity for experiential learning with Cell-based Declustering.

- **Things to try:**

1. Set the cell size very small (1). What's the data weights and CDF / mean? Very large?
2. What do you think is the best cell size for this data configuration?



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 cell sizes and to observe the impact on data weights and the resulting CDF and summary statistics.

2. Adjust the cell size.

cell size: 6
Lo: 30

1. Observed the spatiall sample configuration and values.

	x	y	ix	iy	Por	wtx	wty	wt	Por x wt
1	5	5	6	6	7.6	1	1	1.00	7.6
2	85	5	86	6	8.6	15	1	1.00	8.6
3	5	85	6	86	8.6	1	15	1.00	8.6
4	85	85	86	86	9.8	15	15	1.00	9.8
5	25	5	26	6	10.0	5	1	1.00	10.0
6	5	25	6	26	10.0	1	5	1.00	10.0
7	65	5	66	6	10.6	11	1	1.00	10.6
8	5	65	6	66	10.6	1	11	1.00	10.6
9	45	5	46	6	11.2	8	1	1.00	11.2
10	5	45	6	46	11.2	1	8	1.00	11.2
11	25	85	26	86	11.3	5	15	1.00	11.3
12	85	25	86	26	11.3	15	5	1.00	11.3
13	65	85	66	86	12.1	11	15	1.00	12.1
14	85	65	86	66	12.1	15	11	1.00	12.1
15	45	85	46	86	12.8	8	15	1.00	12.8
16	85	45	86	46	12.8	15	8	1.00	12.8
17	25	25	26	26	13.2	5	5	1.00	13.2
18	25	65	26	66	14.2	5	11	1.00	14.2
19	65	25	66	26	14.2	11	5	1.00	14.2
20	25	45	26	46	15.1	5	8	1.00	15.1
21	45	25	46	26	15.1	8	5	1.00	15.1
22	65	65	66	66	15.5	11	11	1.00	15.5
23	65	55	66	56	16.6	11	10	1.00	16.6
24	55	65	56	66	16.6	10	11	1.00	16.6
25	45	65	46	66	16.7	8	11	1.00	16.7
26	65	45	66	46	16.7	11	8	1.00	16.7
27	55	55	56	56	18.3	10	10	1.00	18.3
28	45	55	46	56	18.6	8	10	1.00	18.6
29	55	45	56	46	18.6	10	8	1.00	18.6
30	45	45	46	46	18.9	8	8	1.00	18.9

Sum wt: 30.00

3. Observe the naïve (raw) data CDF and the declustered (weighted) data CDF's in a table and CDF plot.

Index	Por	cumulP	cumulP*
1	7.6	3%	3%
2	8.6	6%	6%
3	8.6	10%	10%
4	9.8	13%	13%
5	10.0	16%	16%
6	10.0	19%	19%
7	10.6	23%	23%
8	10.6	26%	26%
9	11.2	29%	29%
10	11.2	32%	32%
11	11.3	35%	35%
12	11.3	39%	39%
13	12.1	42%	42%
14	12.1	45%	45%
15	12.8	48%	48%
16	12.8	52%	52%
17	13.2	55%	55%
18	14.2	58%	58%
19	14.2	61%	61%
20	15.1	65%	65%
21	15.1	68%	68%
22	15.5	71%	71%
23	16.6	74%	74%
24	16.6	77%	77%
25	16.7	81%	81%
26	16.7	84%	84%
27	18.3	87%	87%
28	18.6	90%	90%
29	18.6	94%	94%
30	18.9	97%	97%

4. Observed the improvement in the estimate of the mean from naïve (raw) to declustered (weighted).

Naive Mean = 13.3% Declustered Mean = 13.3% True Mean = 12.3%
Naive Mean Error = 7.7% Declustered Mean Error = 7.7%

Workflow

File Name: Declustering_Debiasing_Demo.xlsx File is at: https://git.io/fxA3O

Python / GSLIB Declustering Demo

Here's a workflow that used GeostatsPy.

Things to demonstrate:

1. Load data, visualize
2. Cell-base declustering
3. Visualize the weights
4. Calculated weighted statistics
5. Check multiple cell sizes

GeostatsPy: 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

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
  - dual purpose appraisal and injection / production wells!

Regular Sampling: when samples are taken at regular intervals (equally spaced).

- less reliable than random sampling.
- Warning: may resonate with some unsuspected environmental variable.

What do we have?

- we usually have biased, opportunity sampling
- we must account for bias (debiasing will be discussed later)

So if we were designing sampling for representativity of the sample set and resulting sample statistics, by theory we have 2 options, random sampling and regular sampling.

Data File is at: https://git.io/fh0CW and Jupyter Notebook Workflow is at: https://git.io/fhgJI

Probability and Statistics New Tools

Topic	Application to Subsurface Modeling
Awareness	Every subsurface dataset is sampled to answer questions and add value, not for statistical representativity. *Assume all data sets are biased, test for bias.*
Cell Declustering	Given the spatial location of the sample data, calculate declustering weights. *Build representative sample statistics that correct for sampling bias.*

Data Analytics and Geostatistics:
Data Preparation

Lecture outline . . .

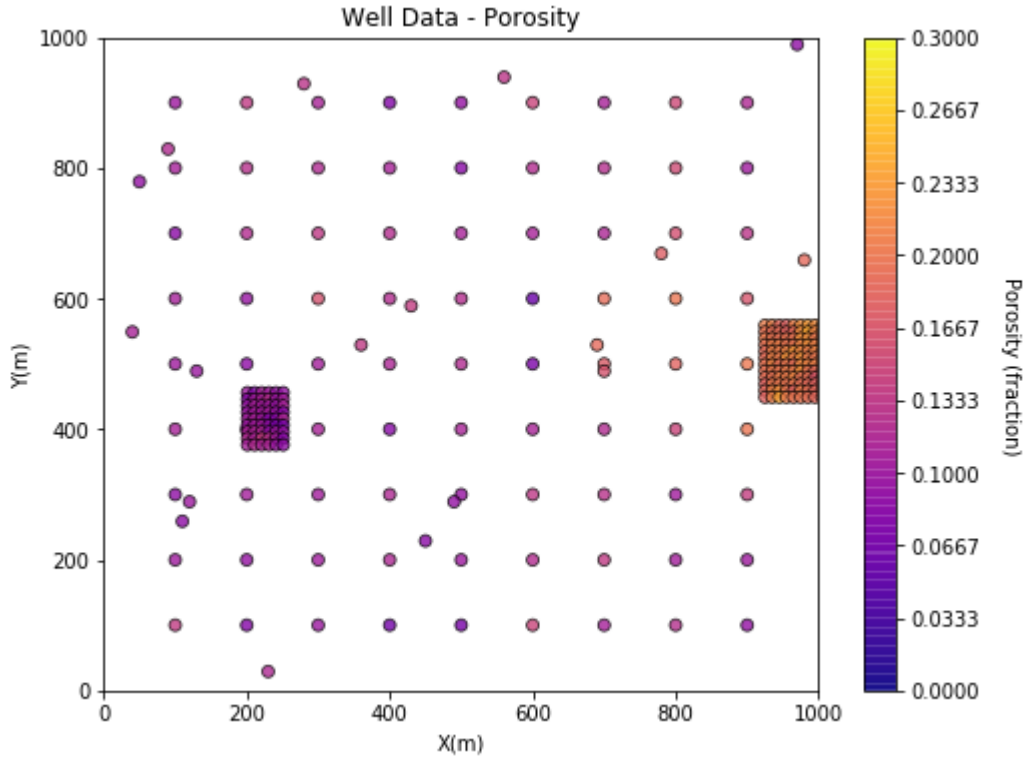
- **Quantifying Uncertainty**

Instructor: Michael Pyrcz, the University of Texas at Austin

Bootstrap Motivation

Uncertainty in the Sample Statistics

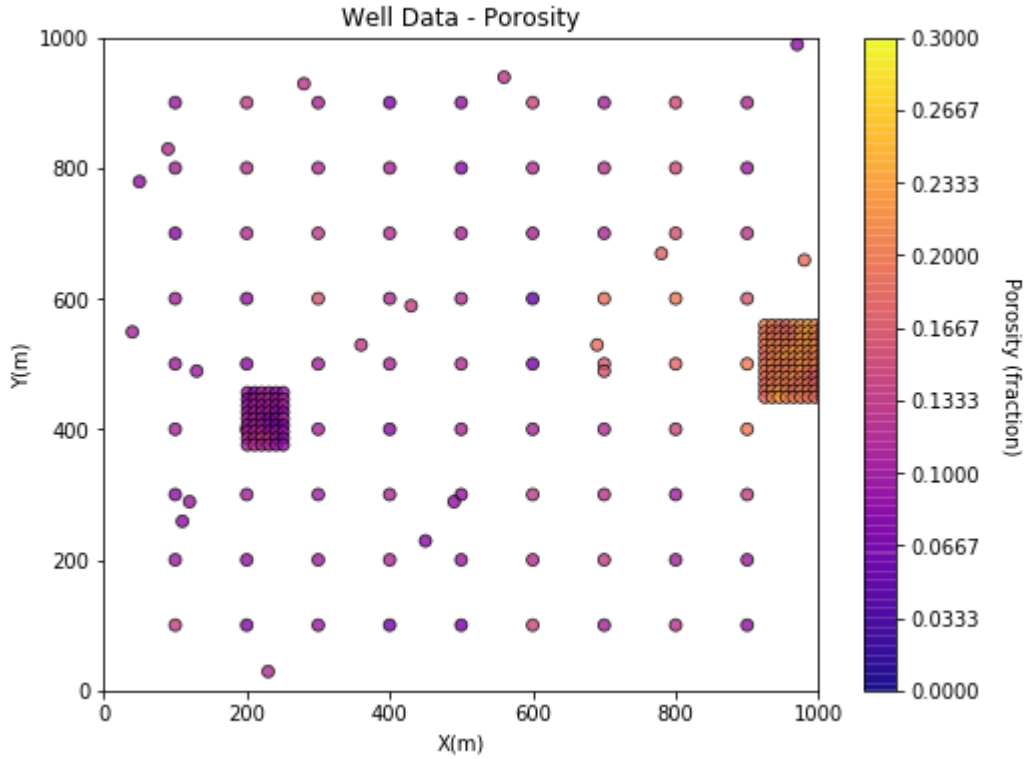
- One source of uncertainty is the paucity of data.
- Do these 200 or so wells provide a precise (and accurate estimate) of the mean? standard deviation? skew? P13?



Bootstrap Motivation

Would it be Useful to Know the Uncertainty in these Statistics Due to Limited Sampling?

- What is the impact of uncertainty in the mean porosity e.g. 20%+/-2%?



Bootstrap Definition

Bootstrap

- method to assess the uncertainty in a sample statistic by repeated random sampling with replacement

Assumptions

- sufficient, representative sampling

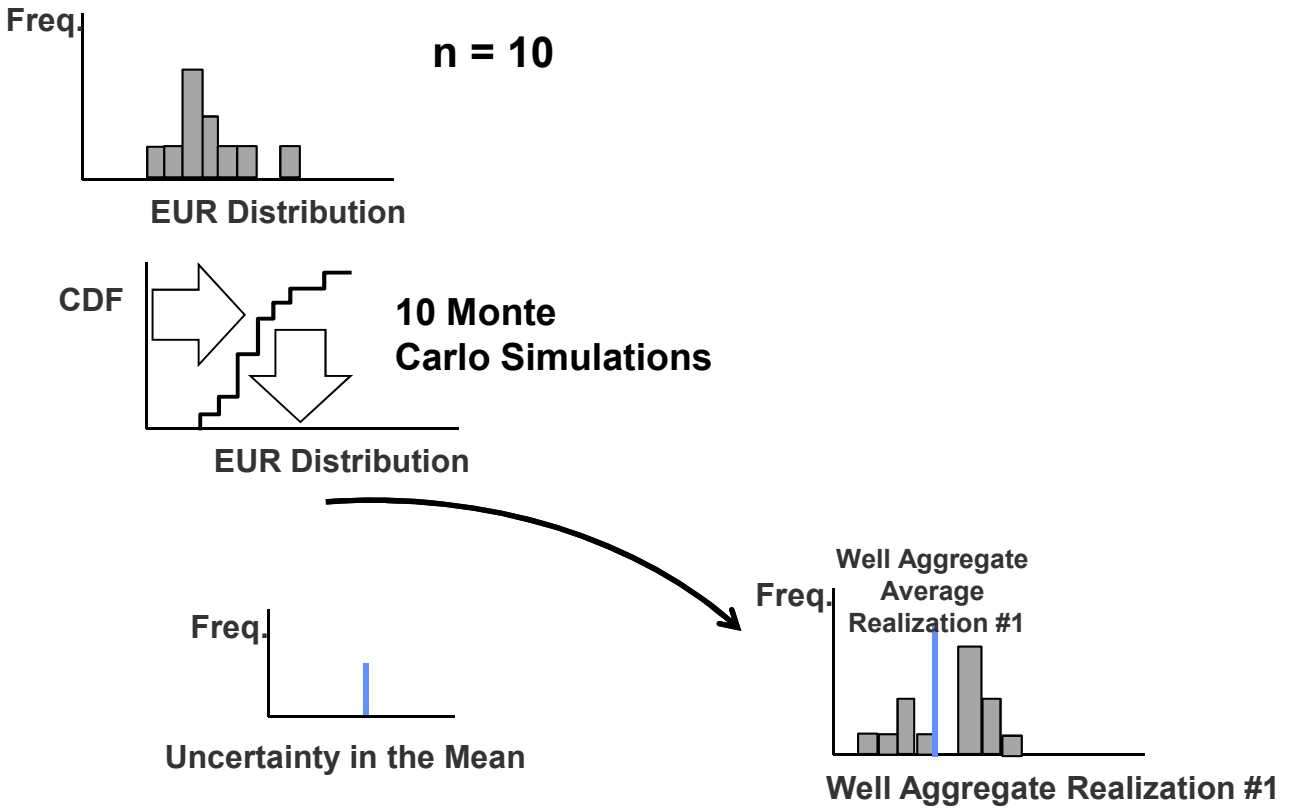
Limitations

- assumes the samples are representative
- assumes stationarity
- only accounts for uncertainty due to too few samples, e.g. no uncertainty due to changes away from data
- does not account for area of interest
- assumes the samples are independent
- does not account for other local information sources

No spatial Context

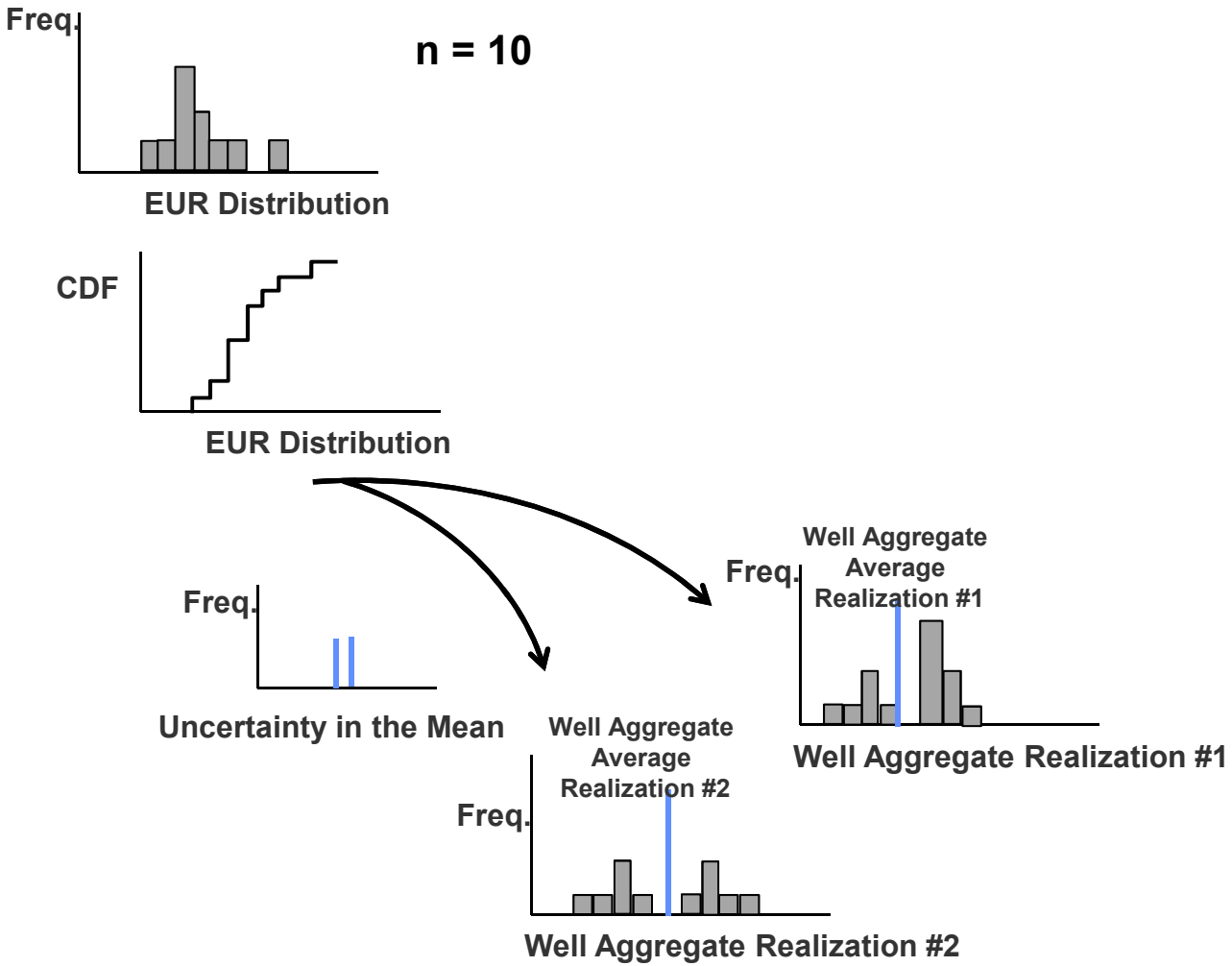
Univariate Statistics
Bootstrap

Bootstrap for Uncertainty in the Mean



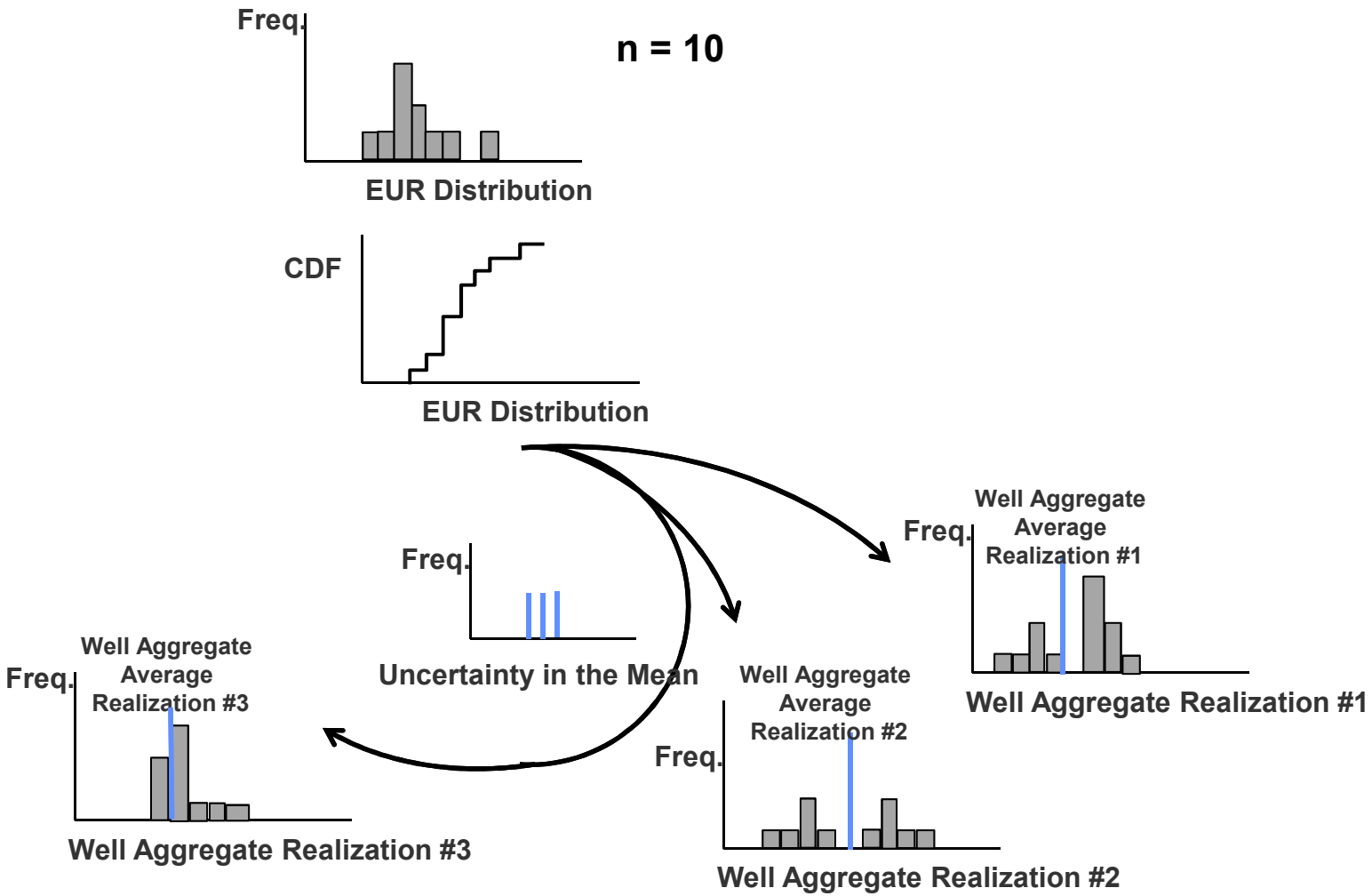
Univariate Statistics
Bootstrap

Bootstrap for Uncertainty in the Mean



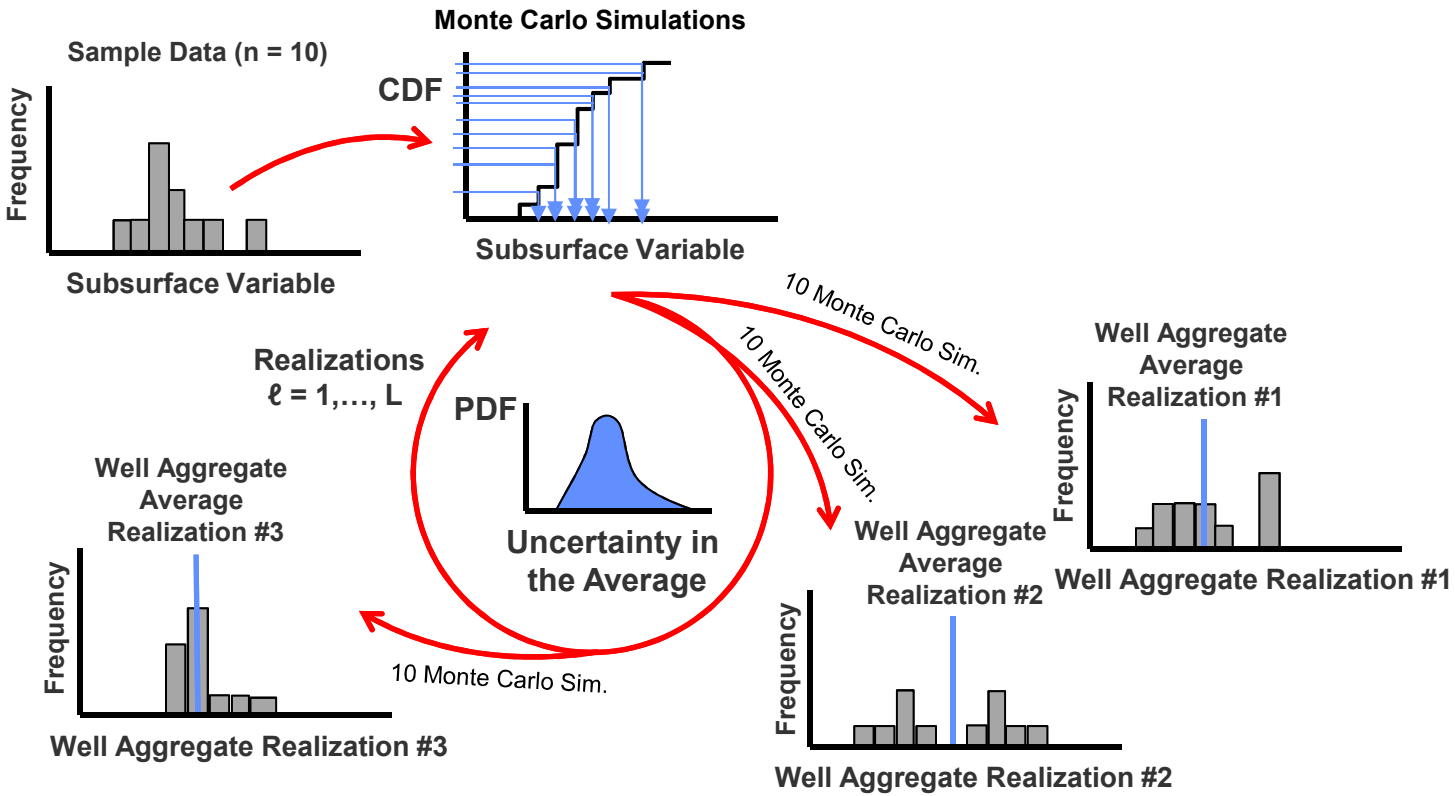
Univariate Statistics
Bootstrap

Bootstrap for Uncertainty in the Mean



Univariate Statistics
Bootstrap

Bootstrap for Uncertainty in the Mean

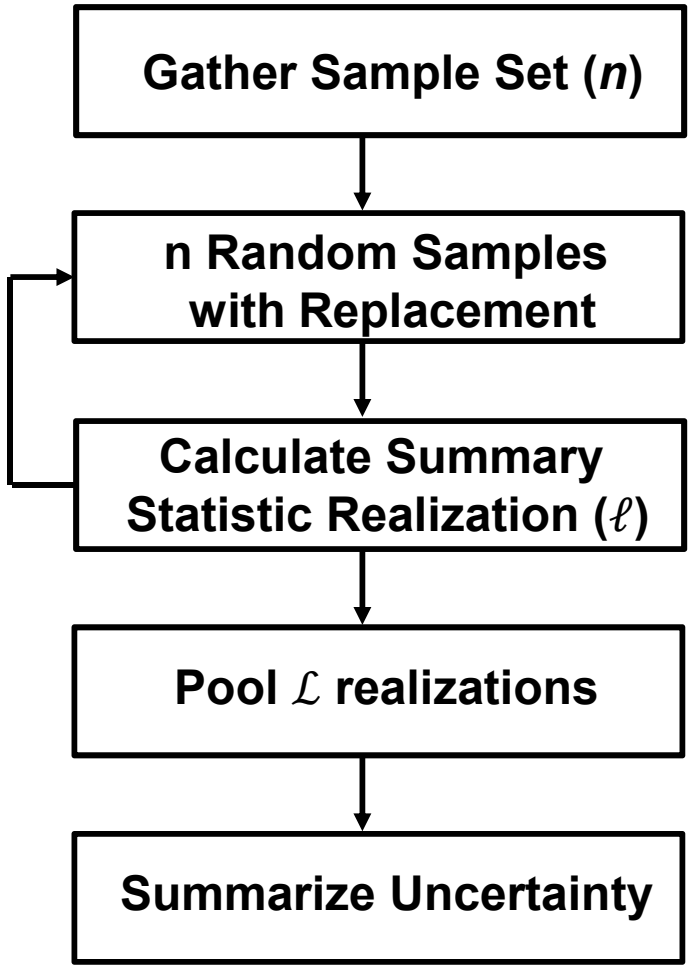


Univariate Statistics
Bootstrap

- Bootstrap Approach (Efron, 1982)
- Statistical resampling procedure to calculate uncertainty in a calculated statistic from the data itself.
- For uncertainty in the mean solution is standard error:

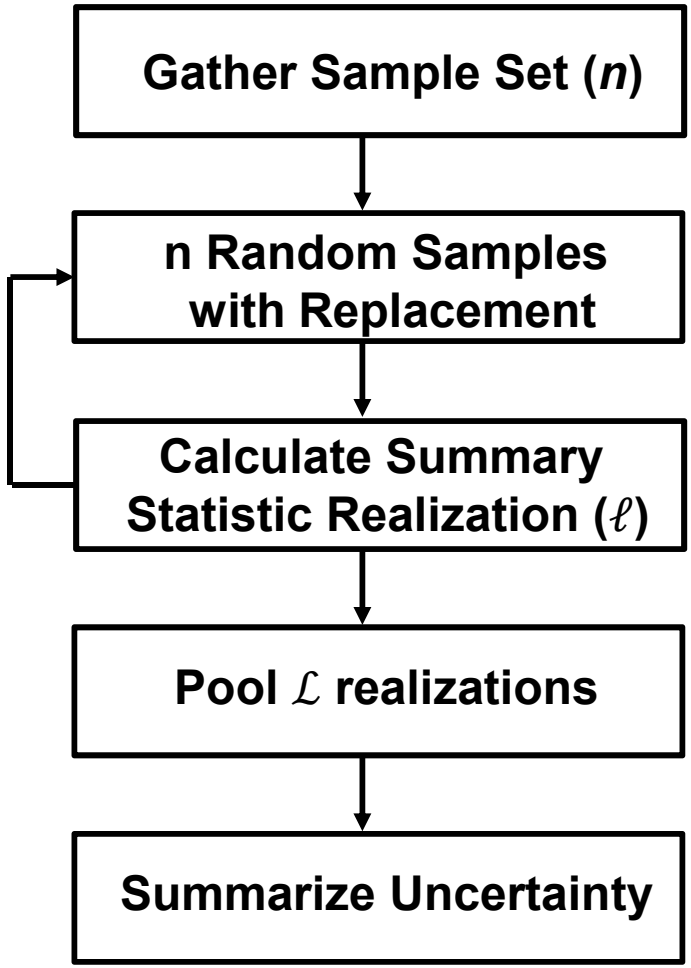
$$\sigma_{\bar{x}}^2 = \frac{\sigma_s^2}{n}$$

- Extremely powerful. Could get uncertainty in any statistic! e.g. P13, skew etc.
- Would not be possible without bootstrap.
- Advanced forms account for spatial information and strategy (game theory).



Univariate Statistics
Bootstrap

- You now know about one of the most powerful tools ever!
- Caveats:
  - assumes the sample set is representative
  - unbiased and covers the full range
  - assumes all samples are independent if not consider Journel's spatial bootstrap (1993).
- You can do bootstrap in Excel.



Bootstrap Hands-on

Here's an opportunity for experiential learning with Bootstrap Uncertainty in Excel.

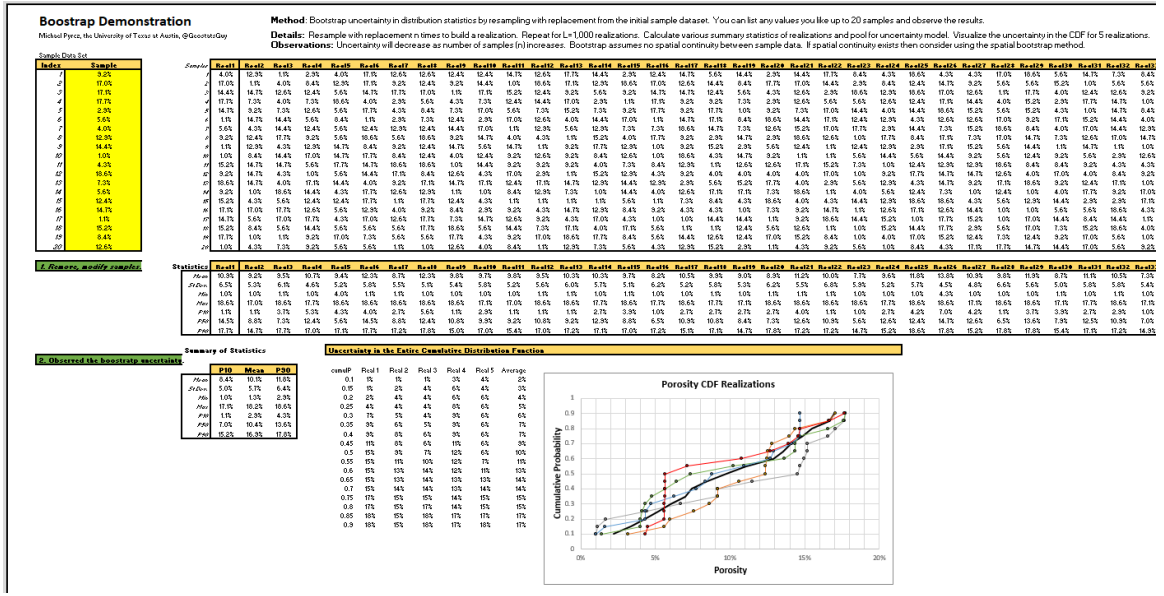
Things to try:

1. Erase some data. Observe the uncertainty range. Does it increase or decrease?

2. Add an extreme value.
   Any impact on uncertainty?

3. Reduce the variance in the samples. Uncertainty now?

4. Compare result to standard error: $SE = \frac{s}{\sqrt{n}}$
   standard deviation divided by number of samples



File Name: Bootstrap_Demo_SImple.xlsx File is at: https://git.io/fjvvP

GeostatsPy Package

Univariate Statistics
Bootstrap Demo

Things to demonstrate:

1. Load data, visualize
2. Summary statistics
3. Bootstrap Realizations
4. Summarization over Bootstrap Realizations
5. Uncertainty in Average and Standard Deviation

GeostatsPy: Bootstrap for Subsurface Data Analytics in Python

Michael Pyrcz, Associate Professor, University of Texas at Austin

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

PGE 383 Exercise: Bootstrap for Subsurface Data Analytics in Python

Here's a simple workflow, demonstration of bootstrap for subsurface modeling workflows. This should help you get started with building subsurface models that integrate uncertainty in the sample statistics.

Bootstrap

Uncertainty in the sample statistics

- one source of uncertainty is the paucity of data.
- do 200 or even less wells provide a precise (and accurate estimate) of the mean? standard deviation? skew? P13?

Would it be useful to know the uncertainty in these statistics due to limited sampling?

- what is the impact of uncertainty in the mean porosity e.g. 20%+/-2%?

Bootstrap is a method to assess the uncertainty in a sample statistic by repeated random sampling with replacement.

Assumptions

- sufficient, representative sampling, identical, idependent samples

Limitations

1. assumes the samples are representative
2. assumes stationarity
3. only accounts for uncertainty due to too few samples, e.g. no uncertainty due to changes away from data
4. does not account for boundary of area of interest
5. assumes the samples are independent
6. does not account for other local information sources

The Bootstrap Approach (Efron, 1982)

Statistical resampling procedure to calculate uncertainty in a calculated statistic from the data itself.

- Does this work? Prove it to yourself, for uncertainty in the mean solution is standard error:

$$\sigma^2_{\bar{x}} = \frac{\sigma^2_x}{n}$$

Extremely powerful - could calculate uncertainty in any statistic! e.g. P13, skew etc.

- Would not be possible access general uncertainty in any statistic without bootstrap.
- Advanced forms account for spatial information and sampling strategy (game theory and Journel's spatial bootstrap (1993).

Steps:

1. assemble a sample set, must be representative, reasonable to assume independence between samples
2. optional: build a cumulative distribution function (CDF)

Data File is at: https://git.io/fh0CW and Jupyter Notebook Workflow is at: https://git.io/fhgUW

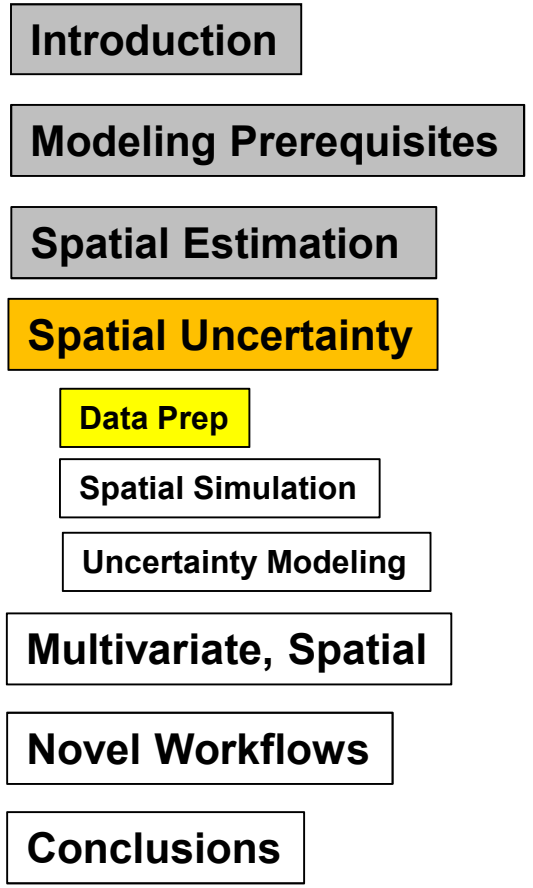
Univariate Statistics
New Tools

Topic	Application to Subsurface Modeling
Awareness of Uncertainty Due to Sparse Sampling	Sample statistics are uncertain due to limited sampling *Quantify and apply this uncertainty model in subsurface modeling workflows.*
Bootstrap	Resamling with replacement to calculate realizations of statistics *While aware of the limitations, use them method to calculate uncertainty in e.g. mean porosity and carry through workflow as scenarios*

Data Analytics and Geostatistics:
Data Preparation

Lecture outline . . .

- **Sampling Limitations**
- **Declustering**
- **Quantifying Uncertainty**



Instructor: Michael Pyrcz, the University of Texas at Austin