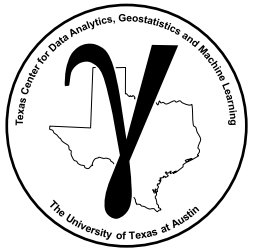
PGE 383 Subsurface Machine Learning

Lecture 4: Data Preparation

Lecture outline:

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

Instructor: Michael Pyrcz, the University of Texas at Austin



Motivation

Any machine learning relies on data!

- Data preparation is still often 80% - 90% of the project effort.
- Garbage in, garbage out
- Our spatial data is quite unique, biased and uncertain

Data Preparation
Other Resources

Recorded Lectures

- Declustering

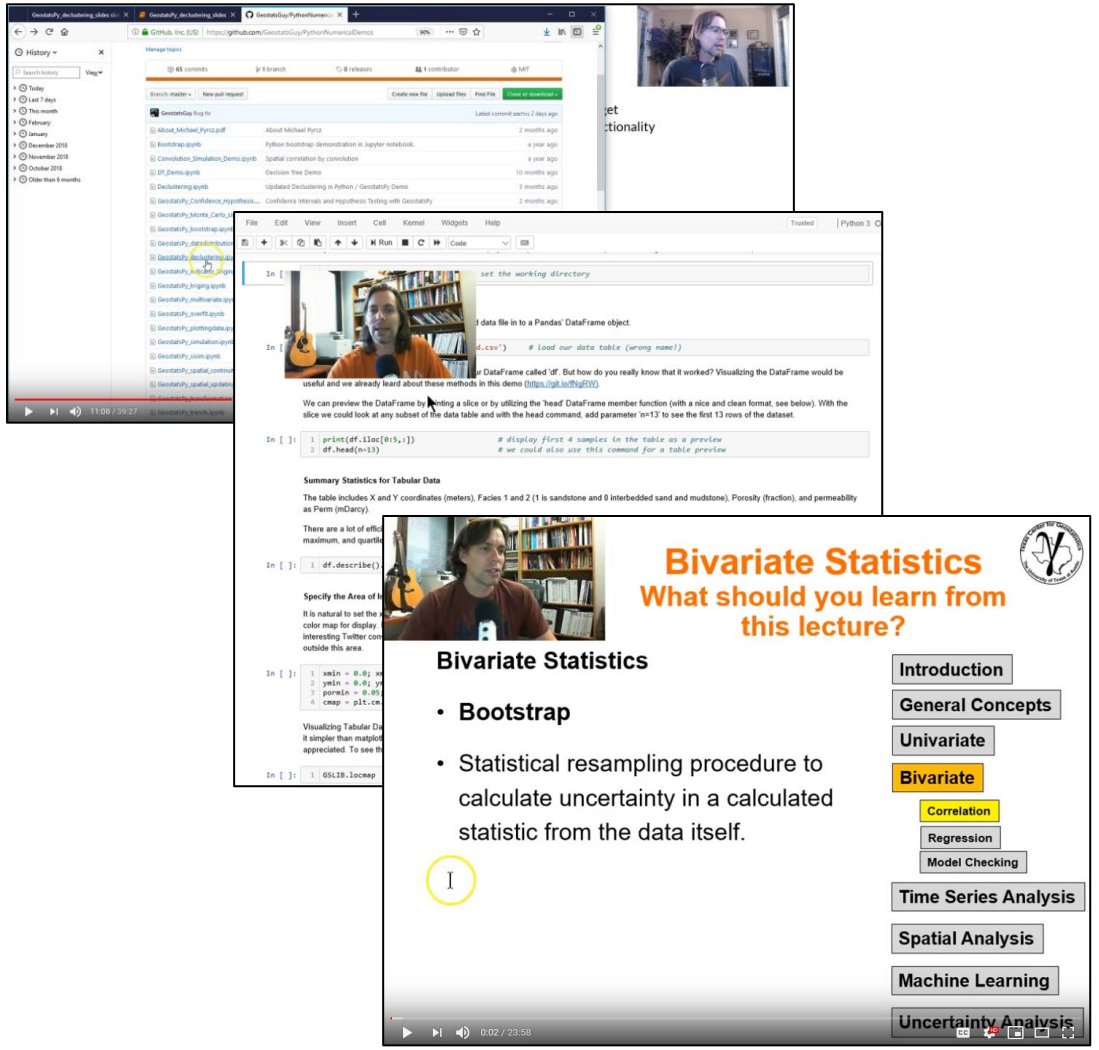
05 Subsurface Modeling Course
https://www.youtube.com/watch?v=k9VbyqafnPk

- Declustering Example

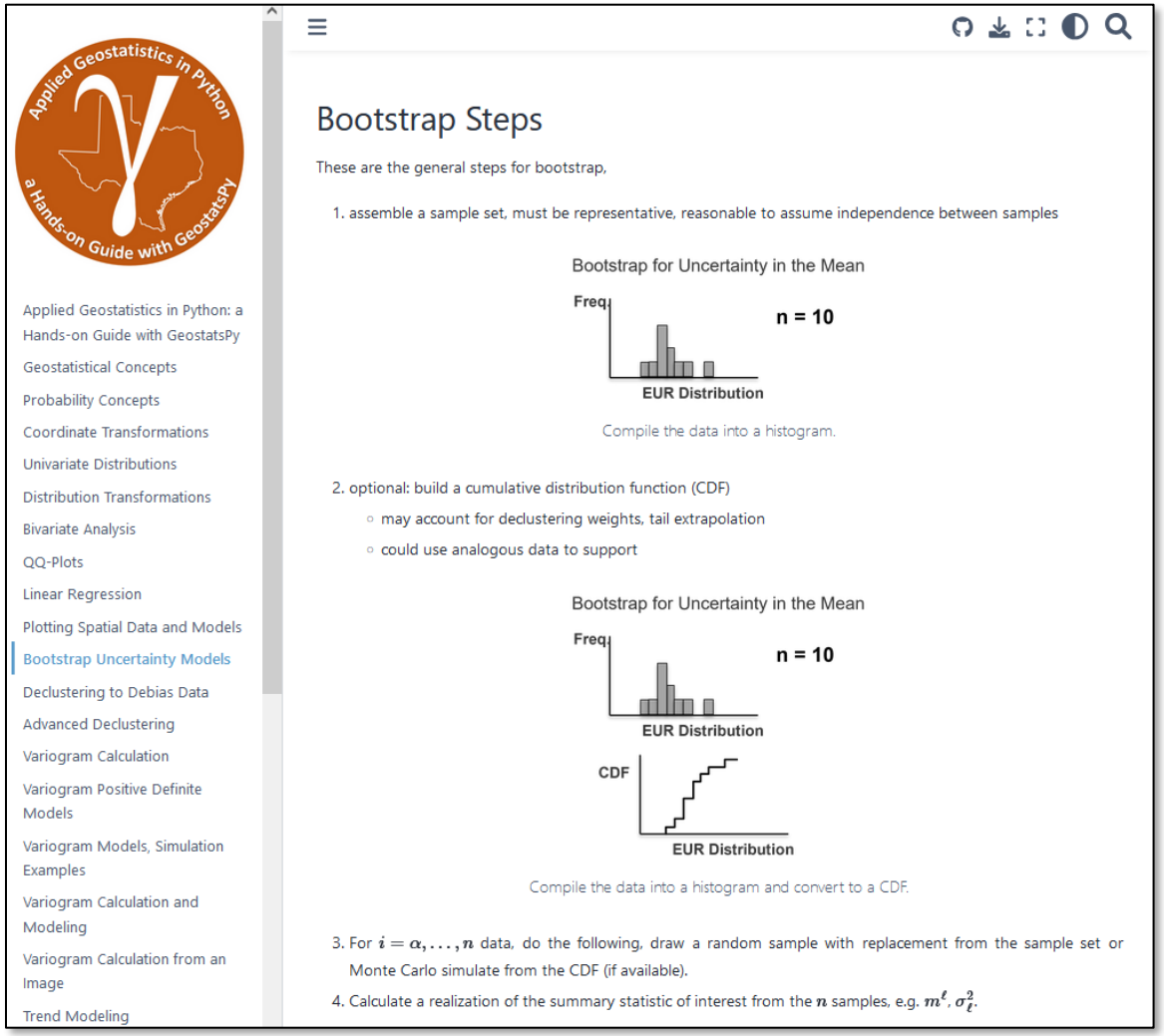
05b Subsurface Modeling Course
https://www.youtube.com/watch?v=rGFO0ulvC5M

- Boostrap

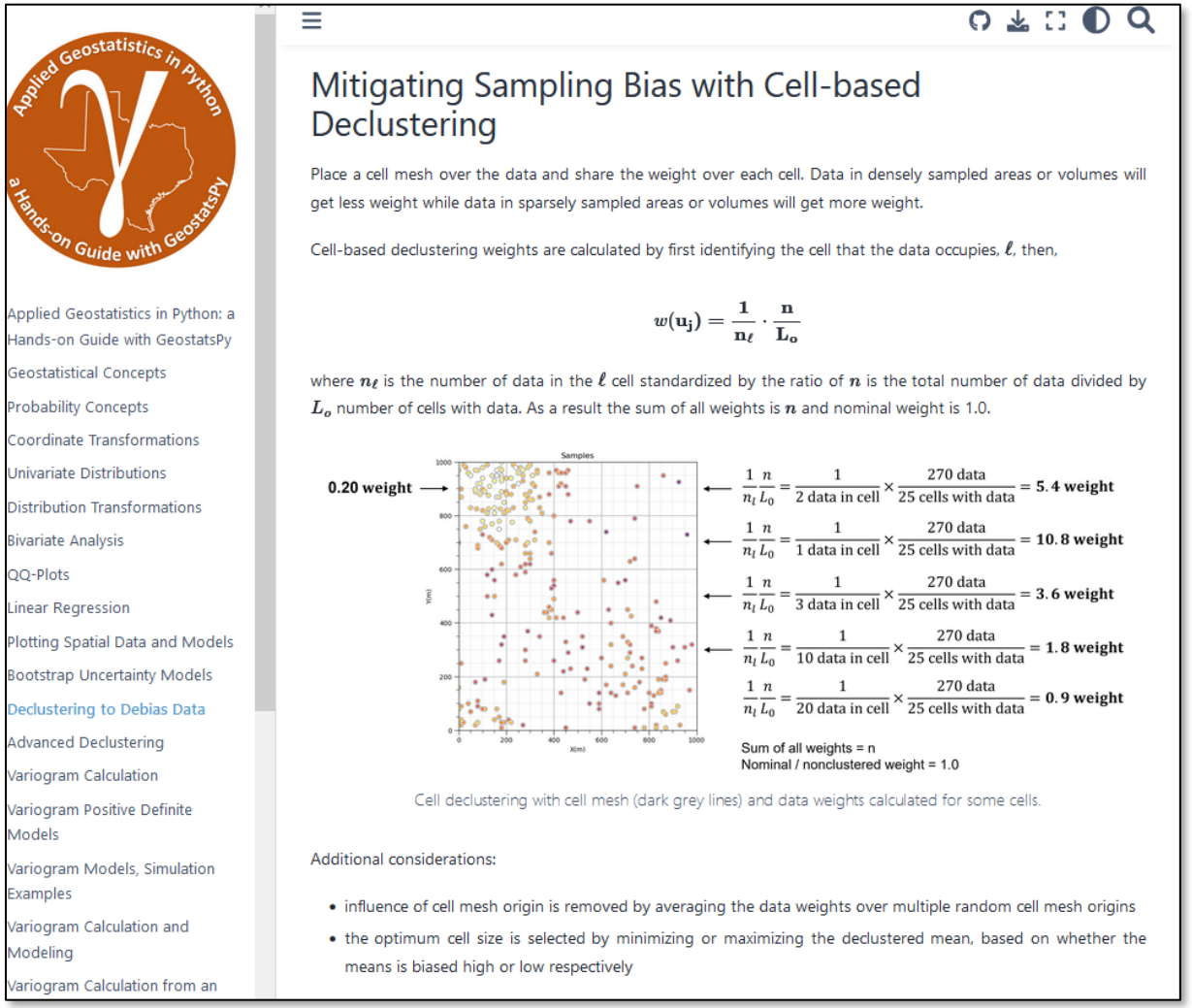
08b Geostatistical Course
https://www.youtube.com/watch?v=wCgdoImlLY0



What is Subsurface Modeling?



Applied Geostatistics in Python, Bootstrap Chapter.



Applied Geostatistics in Python, Declustering Chapter.

PGE 383 Subsurface Machine Learning

Lecture 4: Data Preparation

Lecture outline:

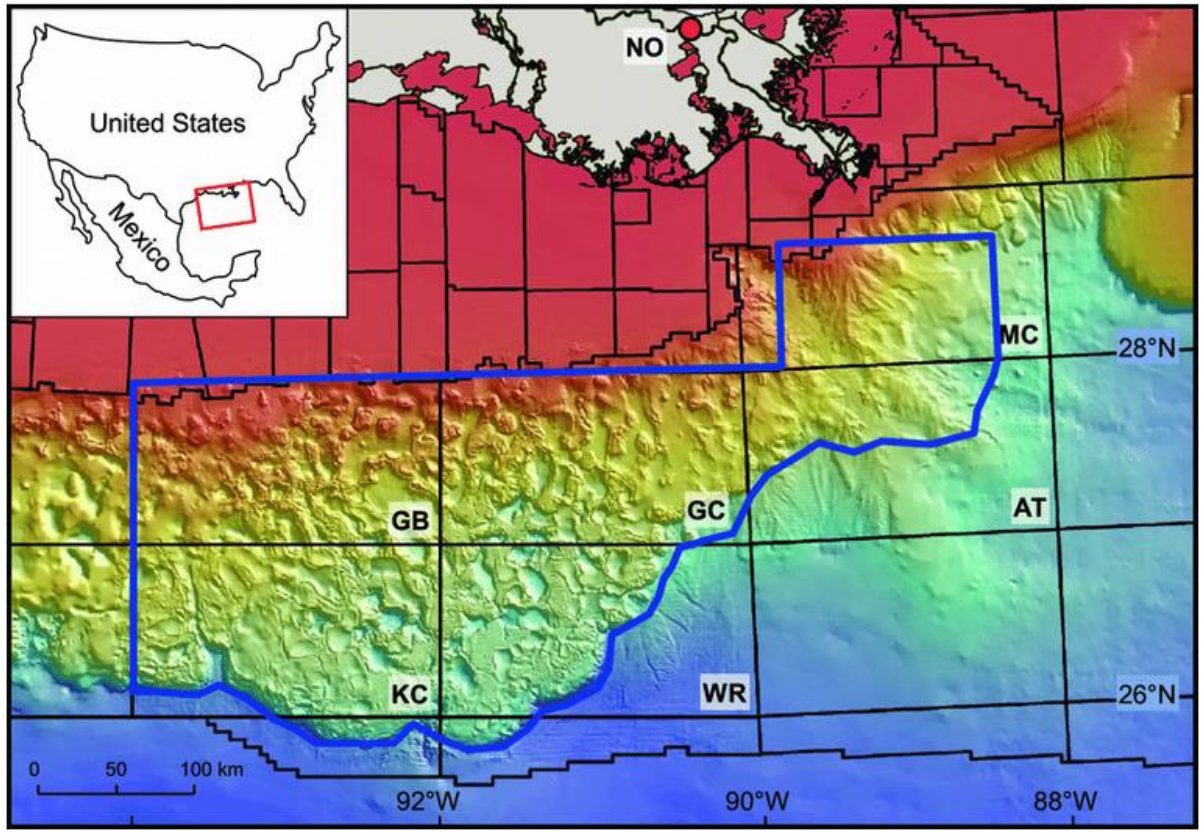
- **Sampling Limitations**

Instructor: Michael Pyrcz, the University of Texas at Austin

Spatial Data Collection

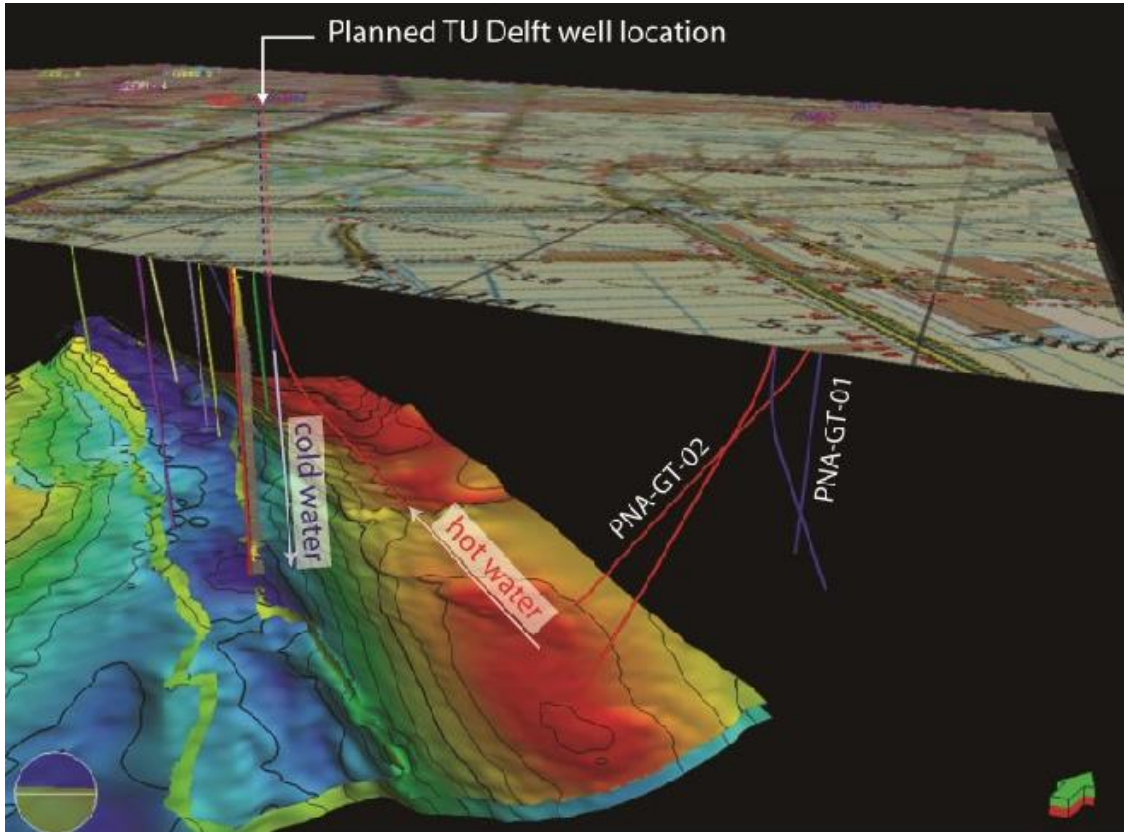
How do we decide where to drill?

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 / minimize uncertainty

- 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*

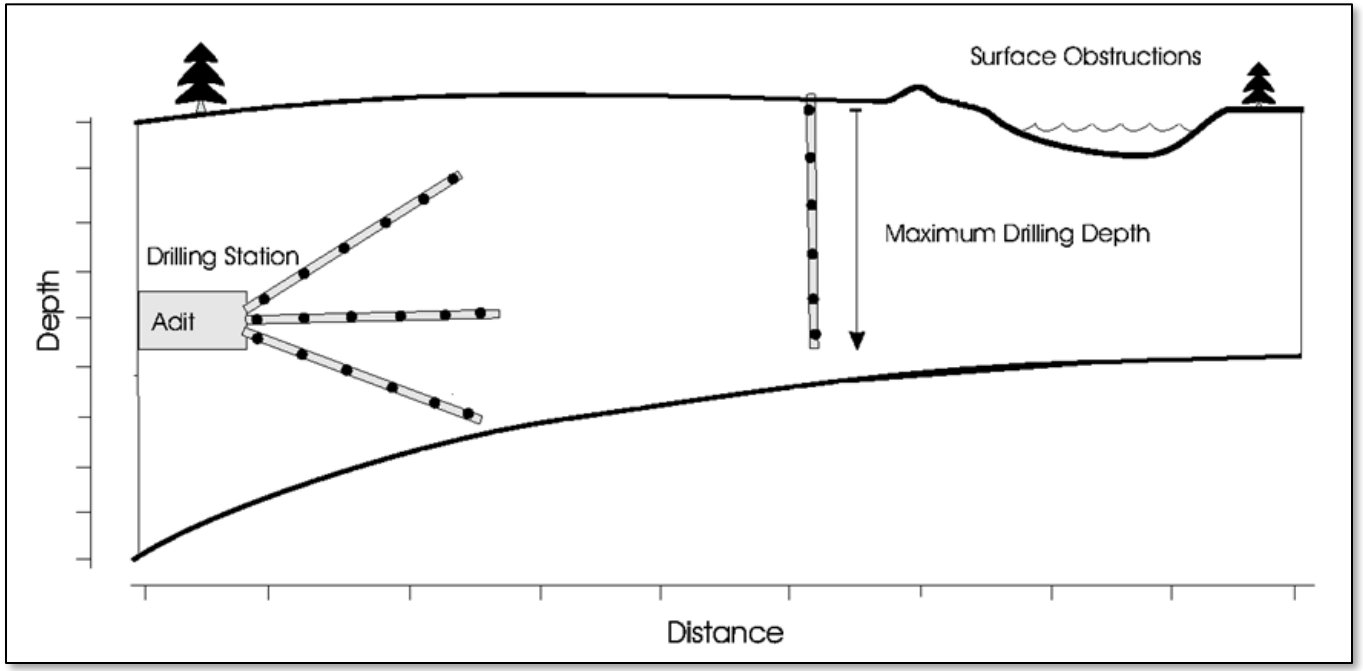
2. maximize NPV directly

- 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

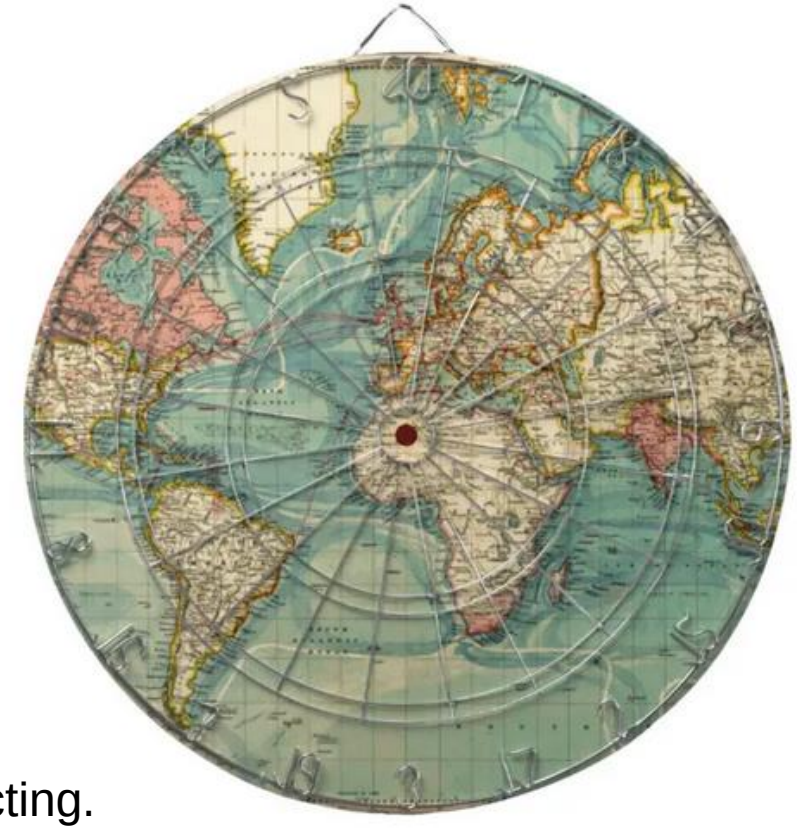
If we were sampling for representativity of the sample set and resulting sample statistics, by theory we have 2 options:

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 (https://www.zazzle.com/vintage_world_map_dartboard_with_darts).

Representative Sampling Definition

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

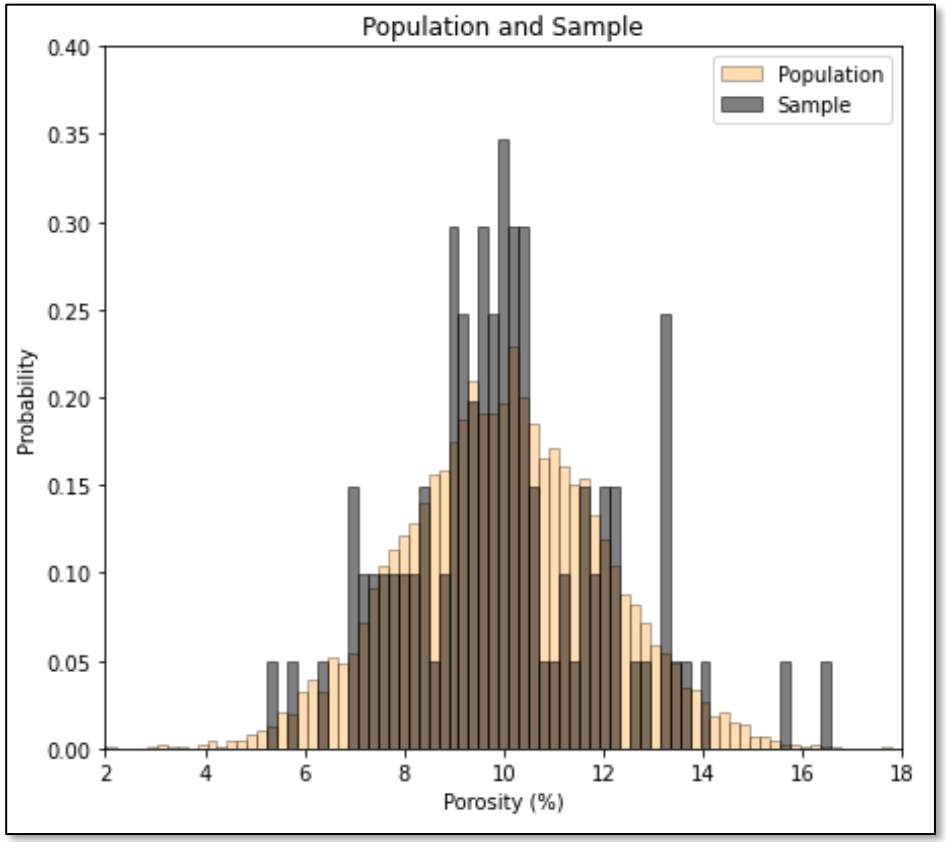
Sampling that results in statistics that match the population parameters in expectation.

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

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

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

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

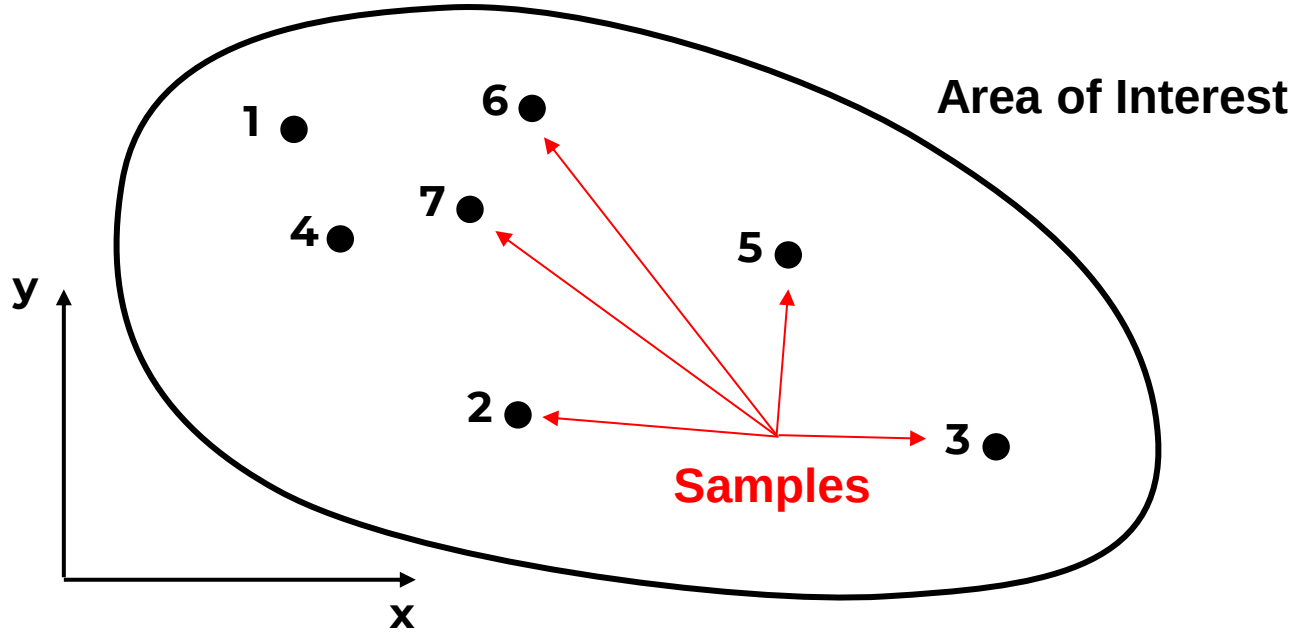
There is a need, however, to adjust the histograms and summary statistics to be representative of the entire volume of interest. We use statistics to make decisions!

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

Clustered Sampling

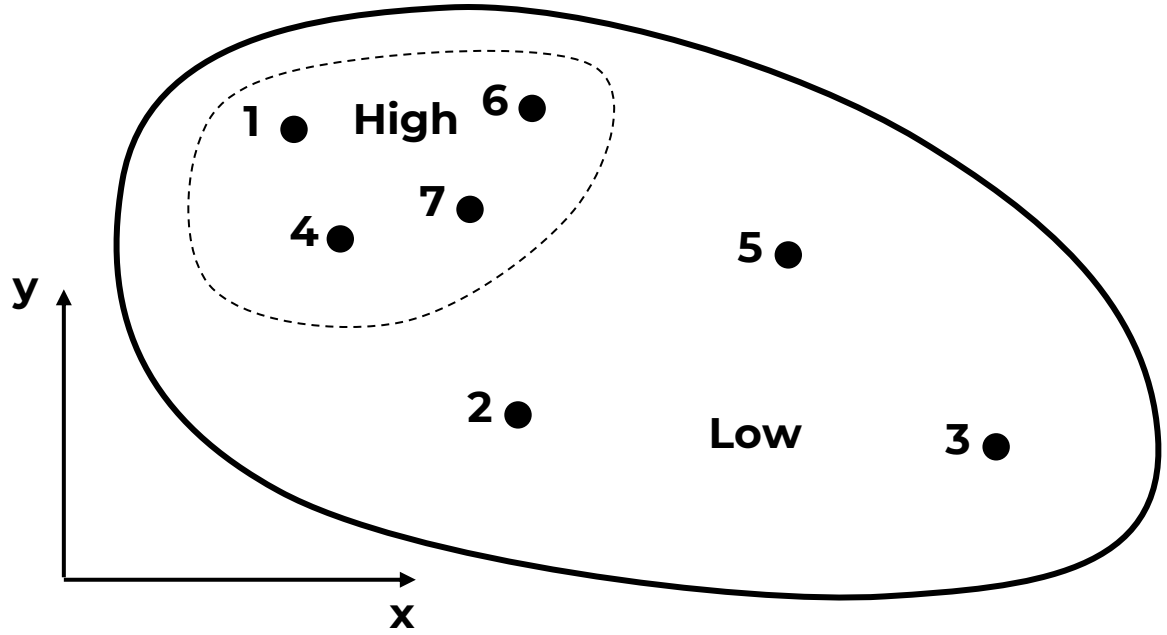
Let's make an estimate for an Area / Volume of Interest:



- e.g. the average porosity to calculate OIP

Clustered Sampling

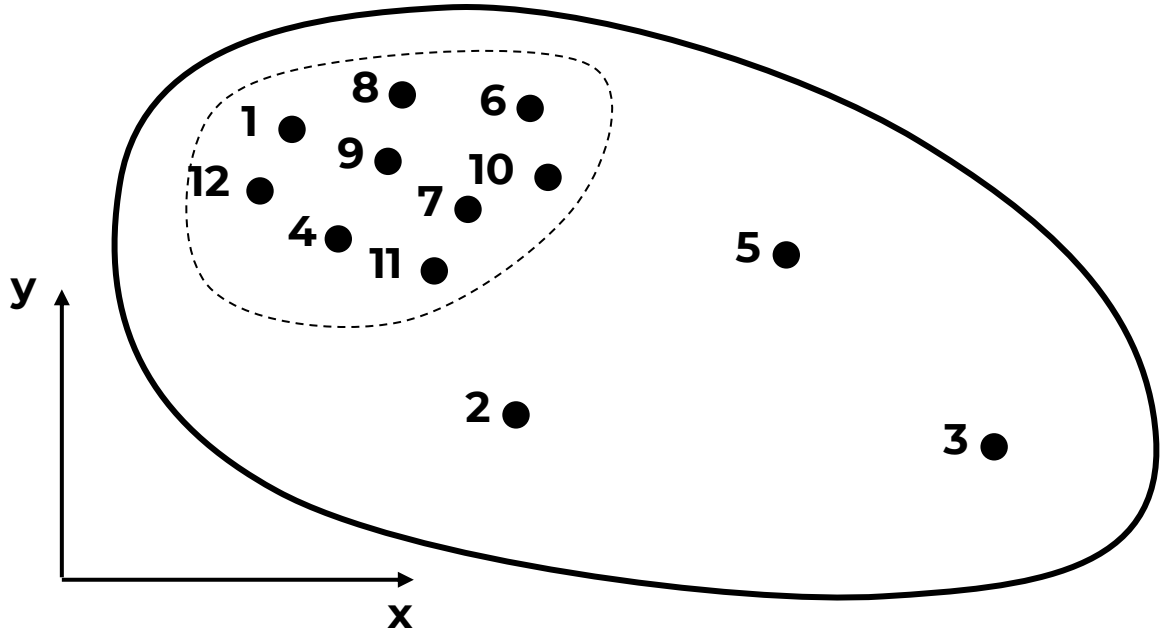
Let's make an estimate for an Area / Volume of Interest:



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

Clustered Sampling

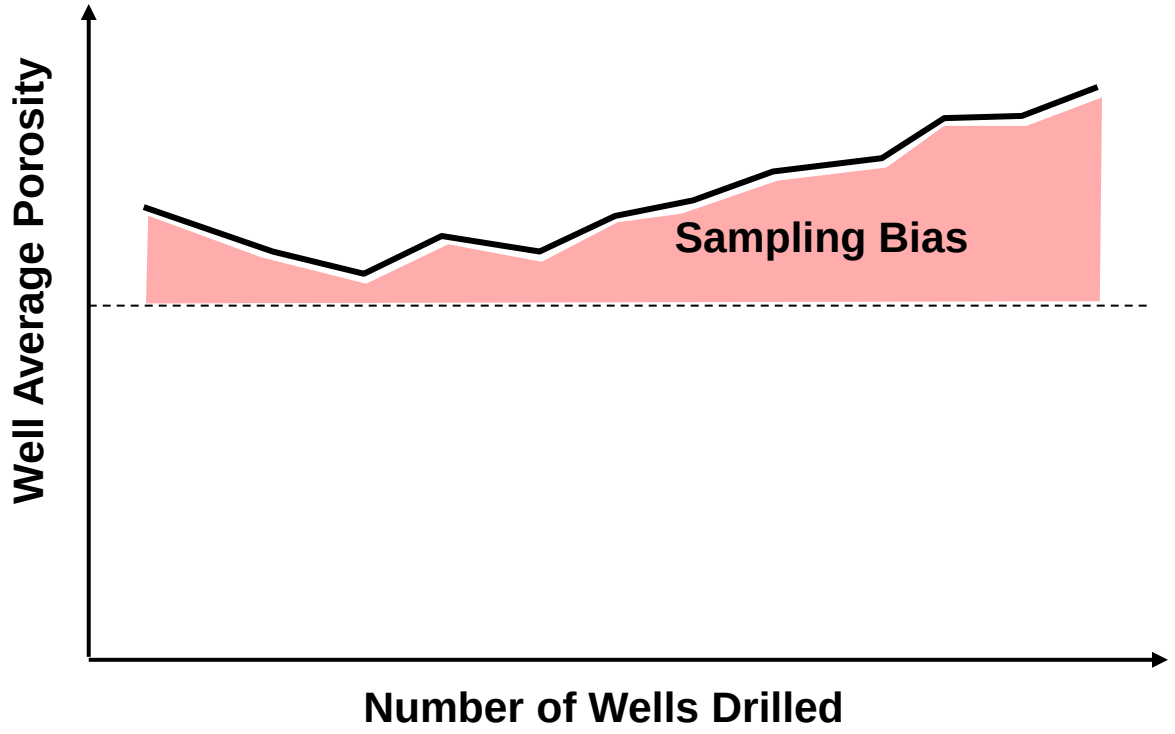
Let’s make an estimate for an Area / Volume of Interest:



- 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?

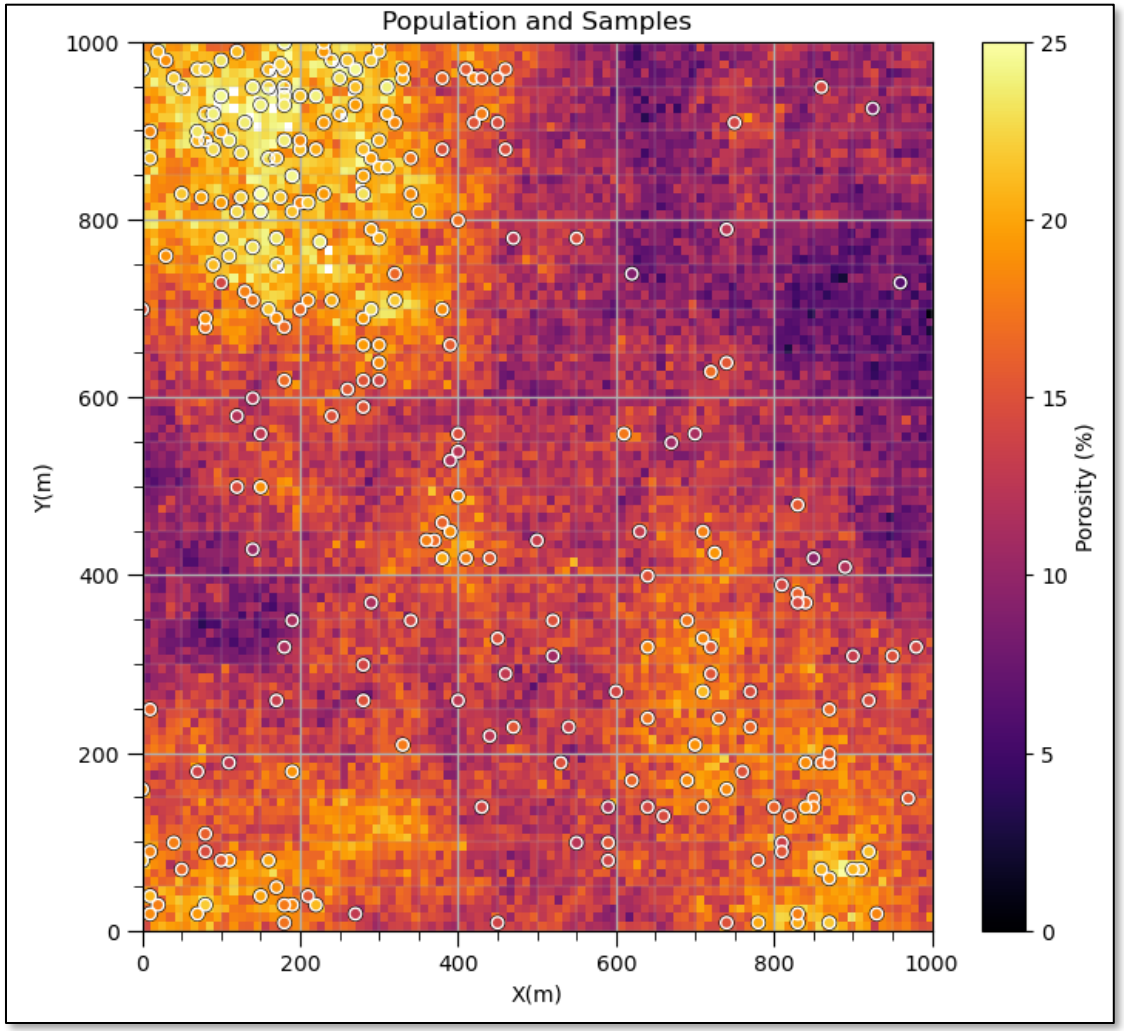


- The naïve sample average becomes more biased!
- We need a method to correct for clustered samples.

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 Declustering chapter of Applied Geostatistics in Python e-book.

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

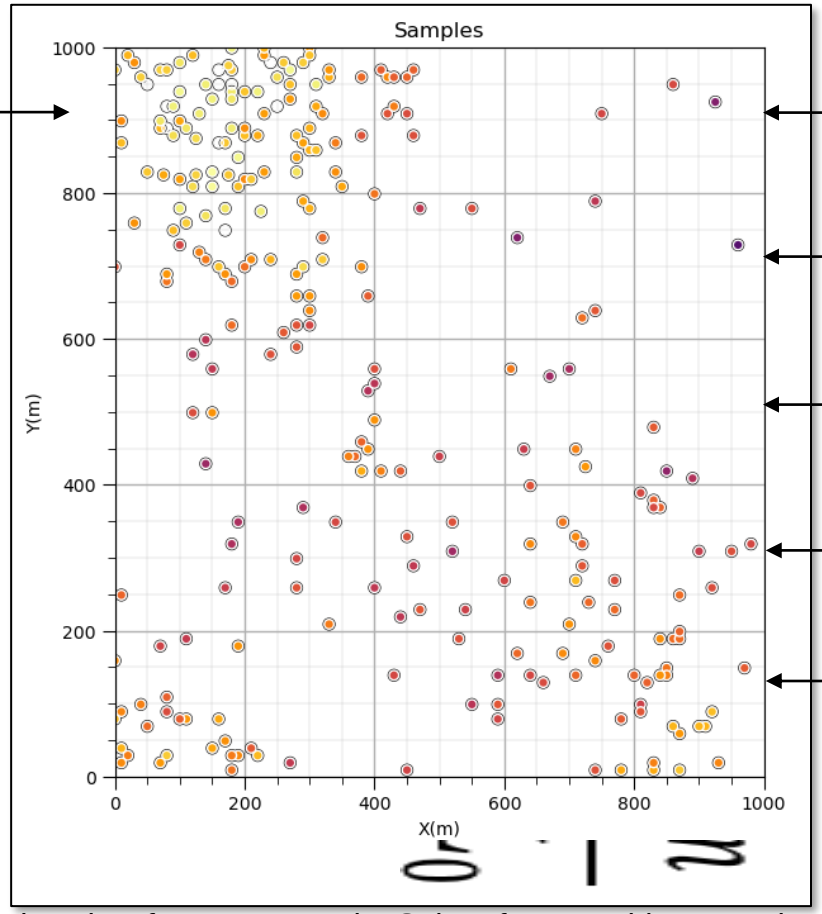
Cell-based Declustering

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 →

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

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

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

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

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

All data in the same cell get the same weight.

Clustered, biased data
GeostatsPy_declustering.ipynb

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

Cell-based Declustering

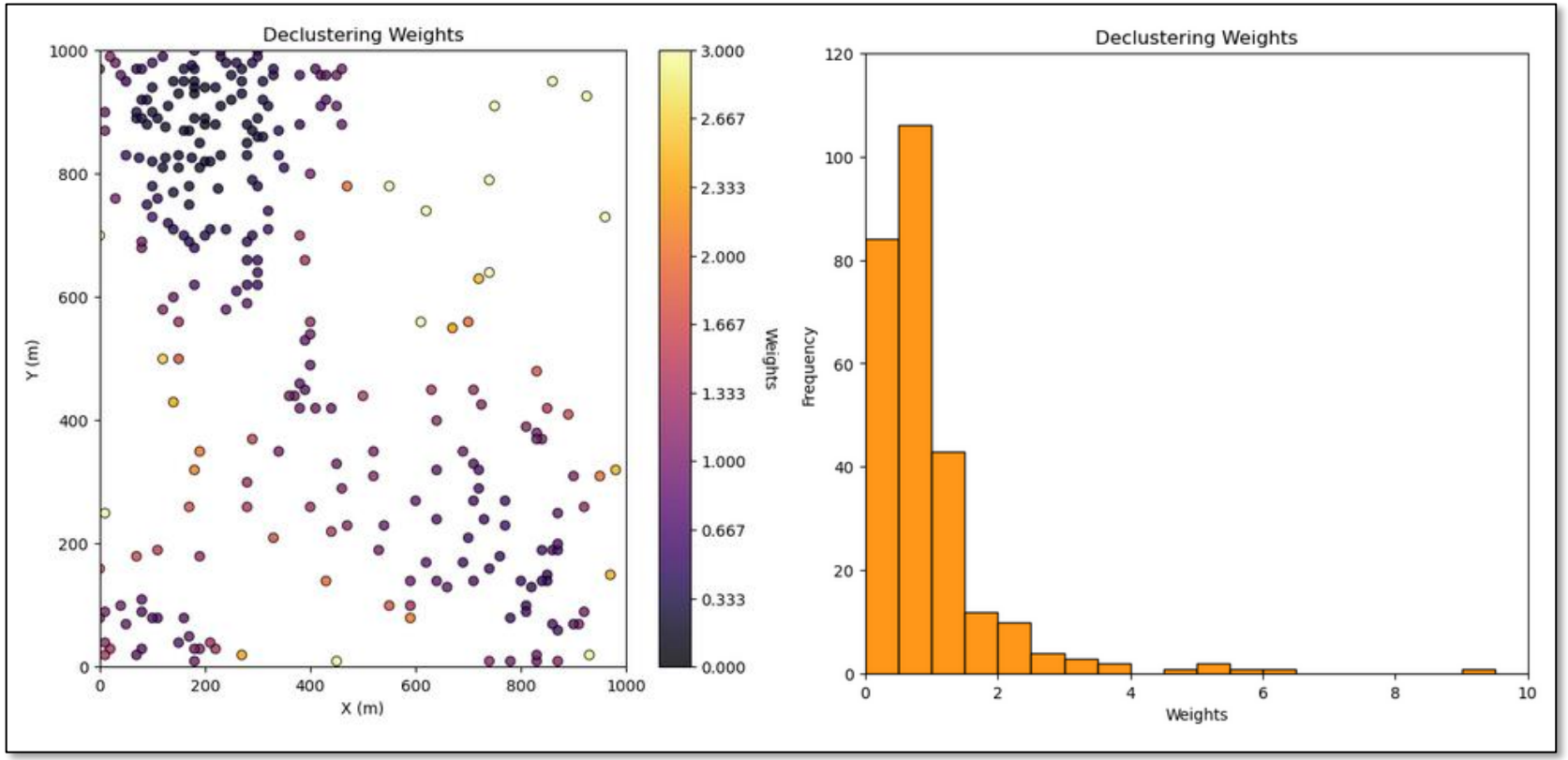
Declustering weights

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 (left) and location map (right) (GeostatsPy_declustering.ipynb).

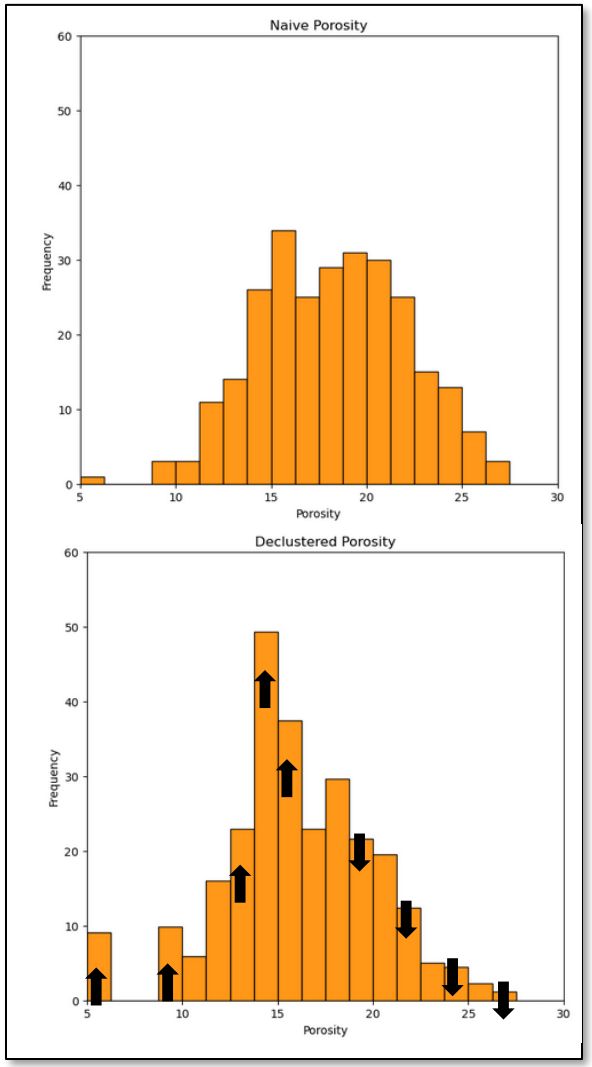
Declustered Distribution

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 command allows for a vector of weights.
  - Many statistics, algorithms and visualizations allow for data weighting
  - Commonly applied in hydrology and other scientific disciplines



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

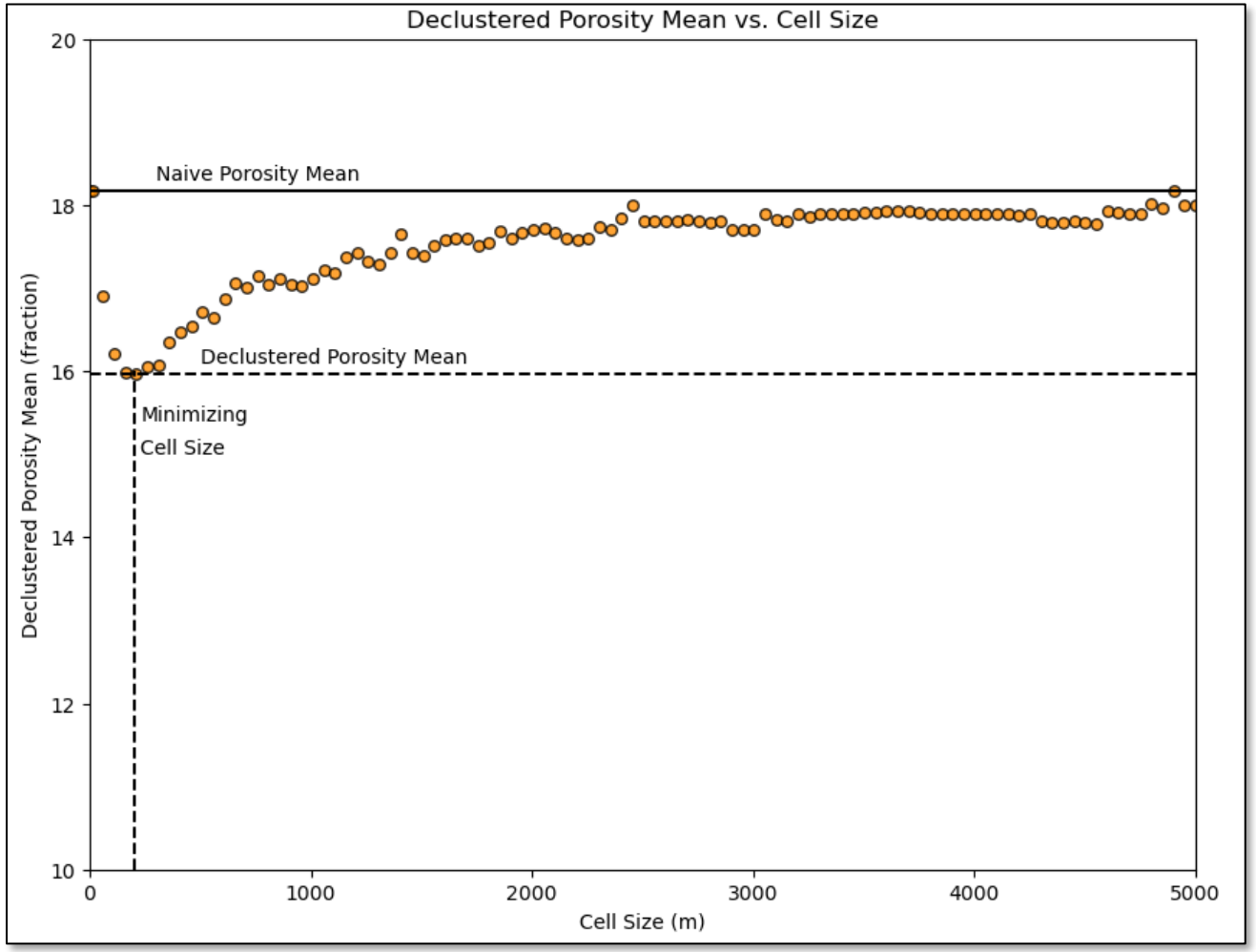
Cell Size Selection

- 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

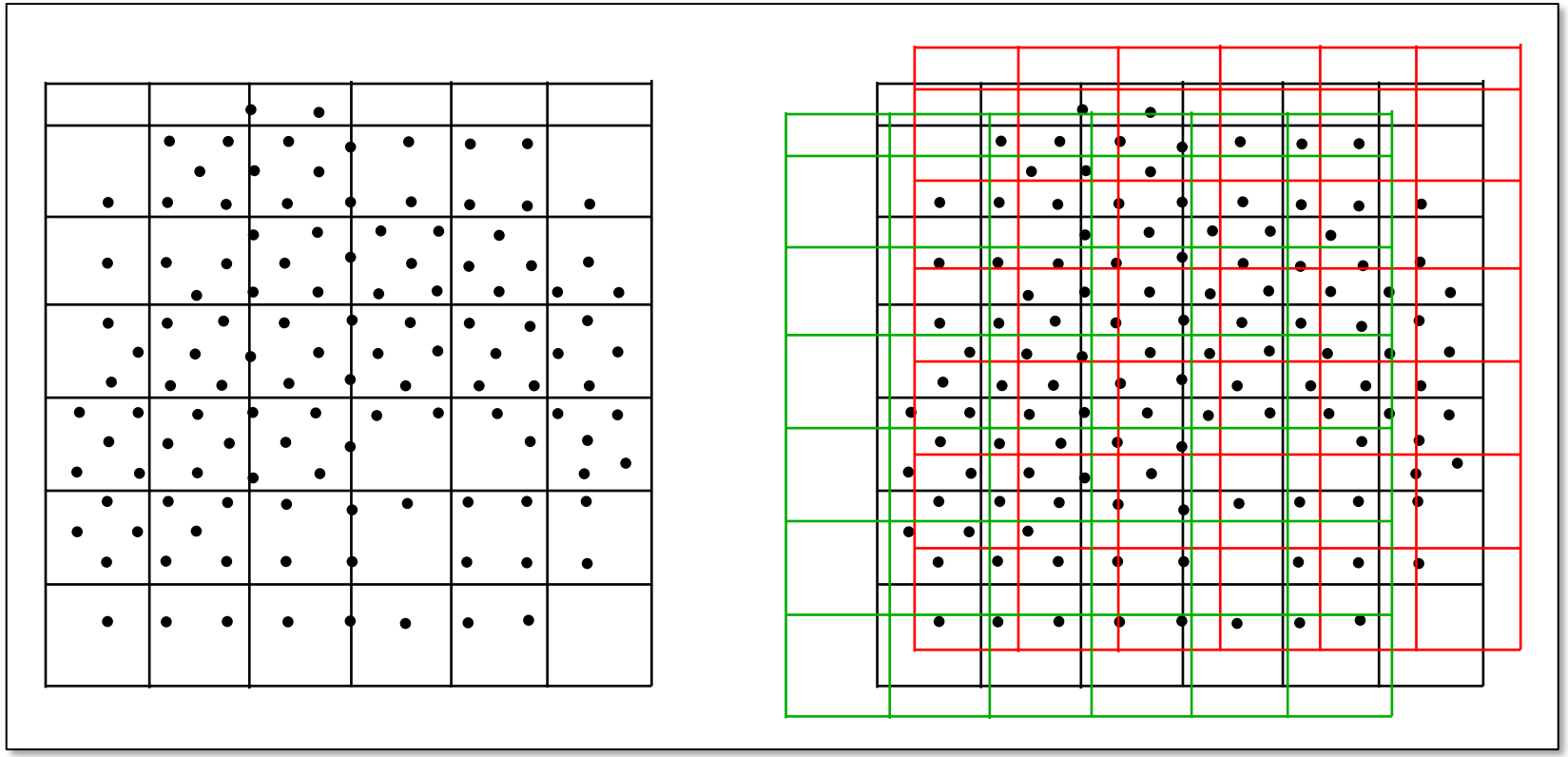
- Could choose the cell size so that there is **approximately one datum per cell in the sparsely sampled areas**, the nominal spacing



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 of 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

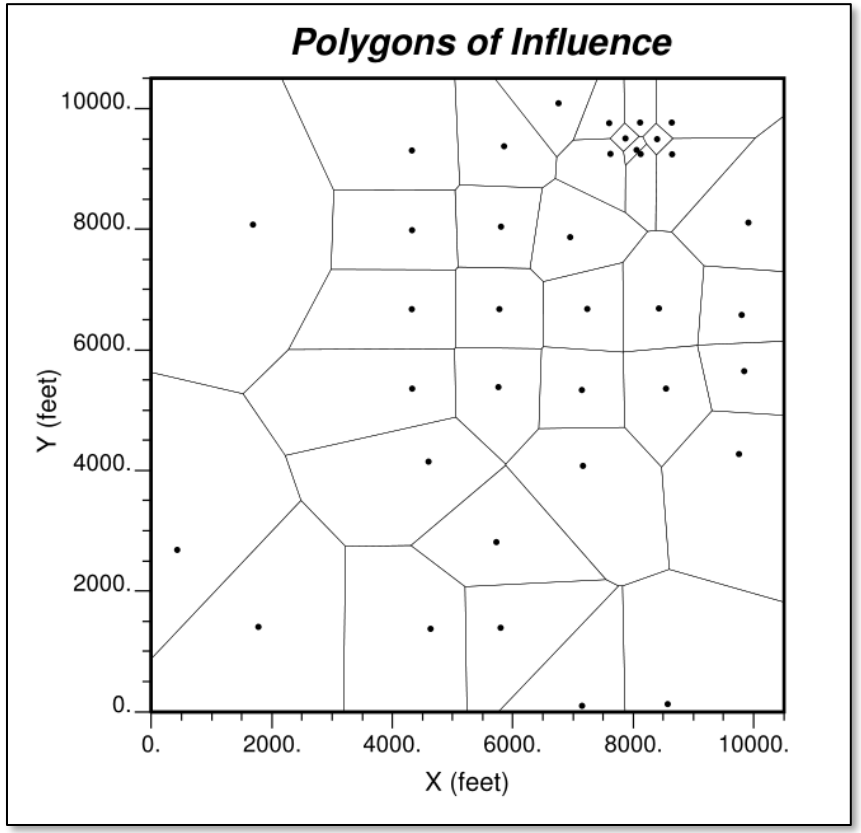
Polygonal Declustering

- 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.

Thiessen Polygons approach from hydrology, calculating representative rain fall from biased rain gauge locations (Thiessen, 1911).

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:

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

- 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

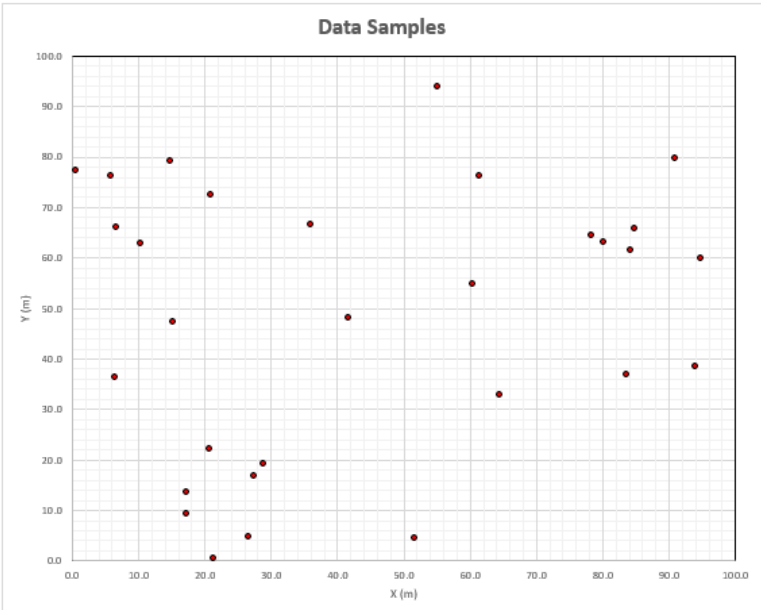
About: This demonstration includes cell-based declustering applied on a random sample set from a truth model.
Dataset: The truth model is a simple 2D geometric function with the high in the center of the area of interest (at 50m,50m).
Objective: Provide an opportunity to experiment with declustering 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	34.1	55	35	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

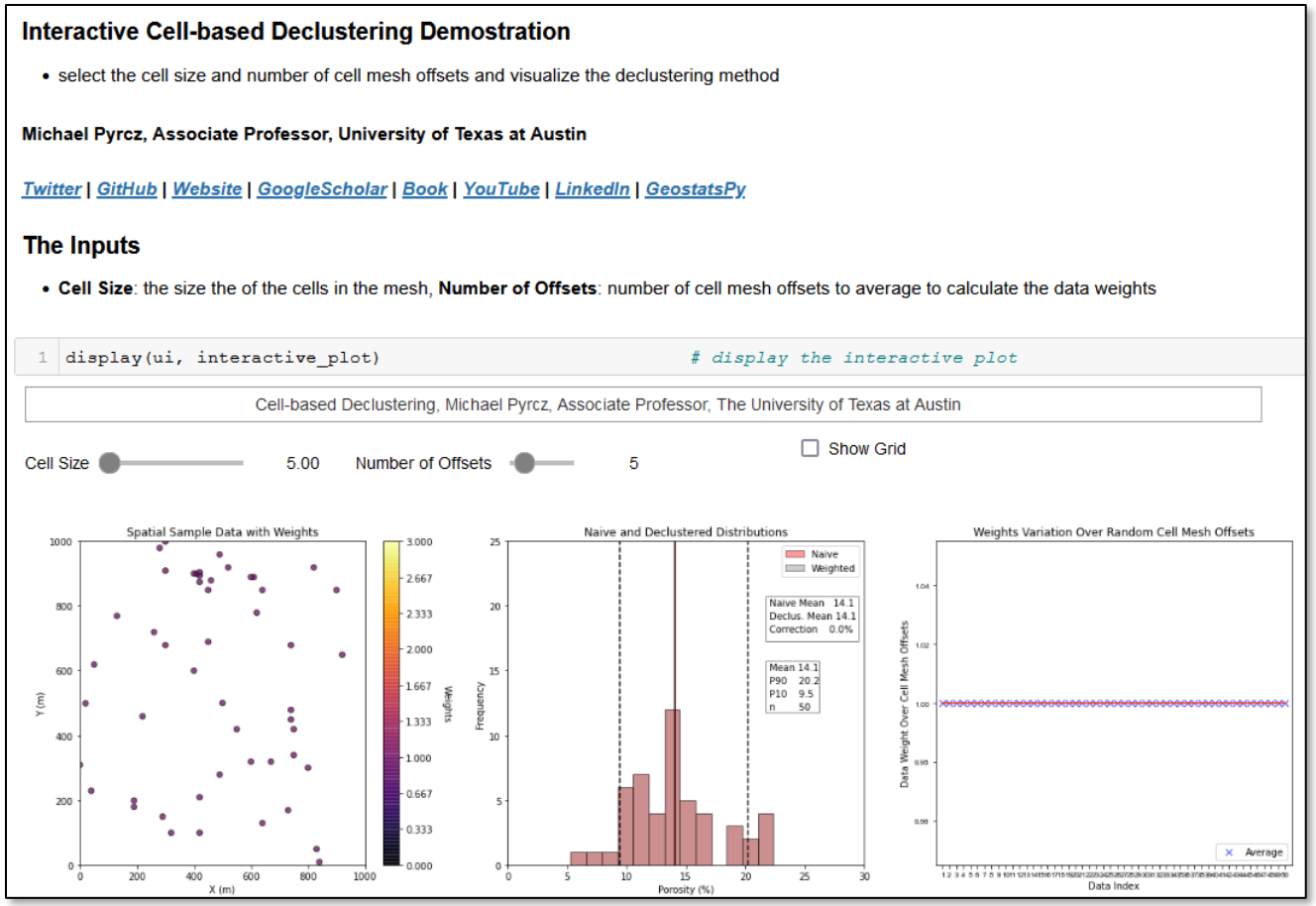
Sum wt: 30.00



Declustering in Excel demonstration, file is Decluster_Debias_Demo.xlsx.

Python Interactive Demonstration

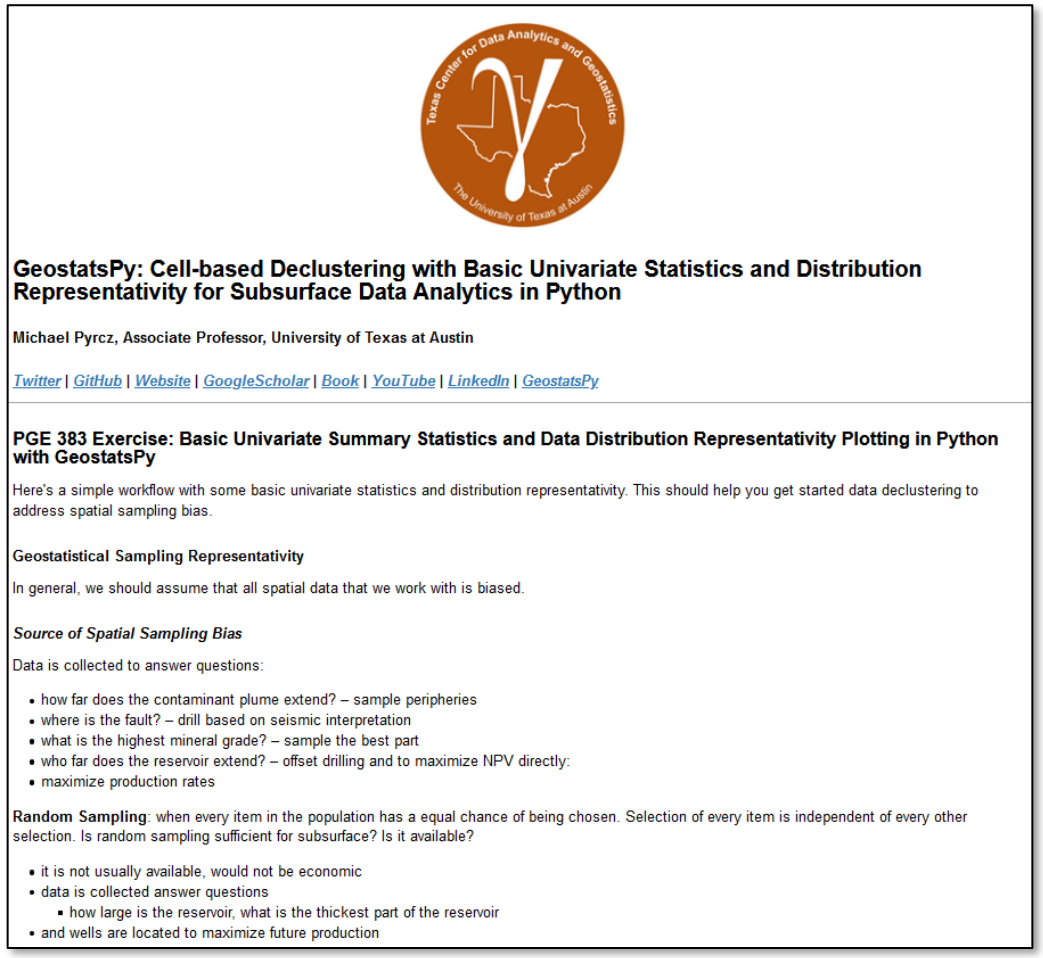
Here's interactive cell-based declustering in Python.



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.

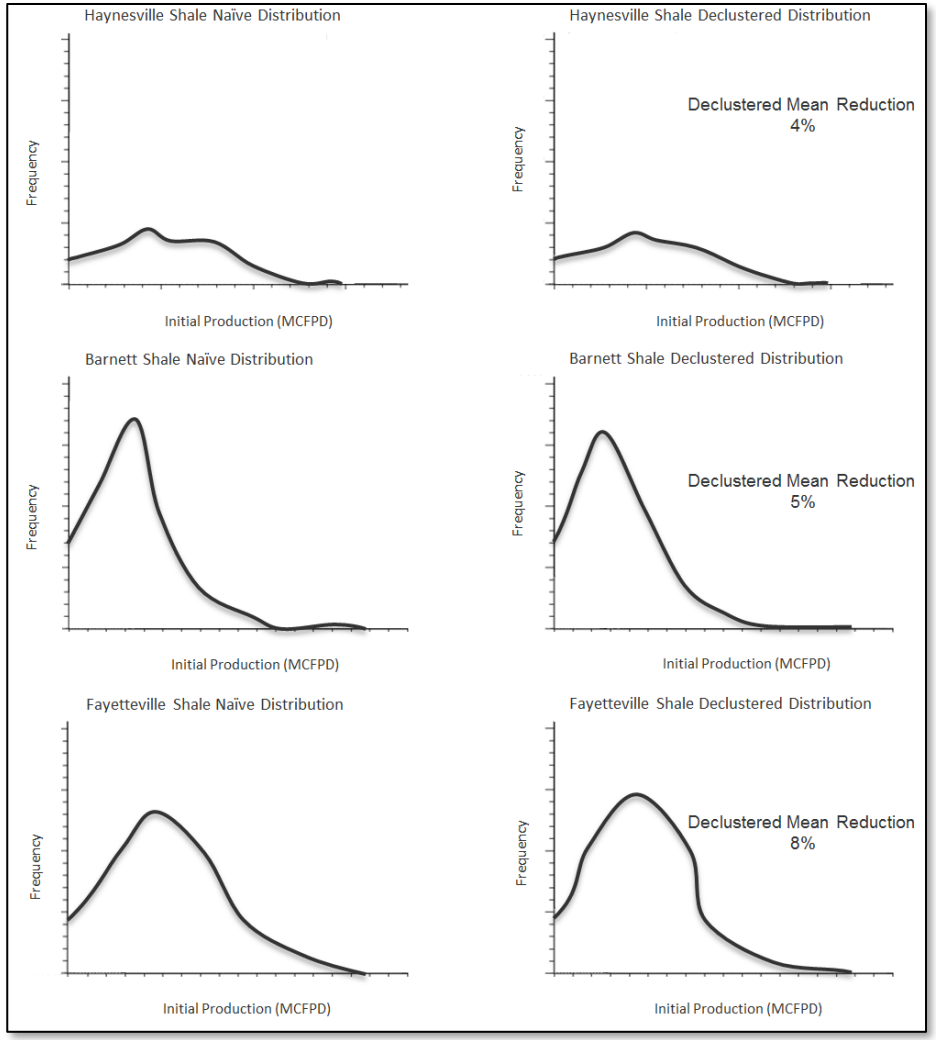
Comments on Cell Declustering

- 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

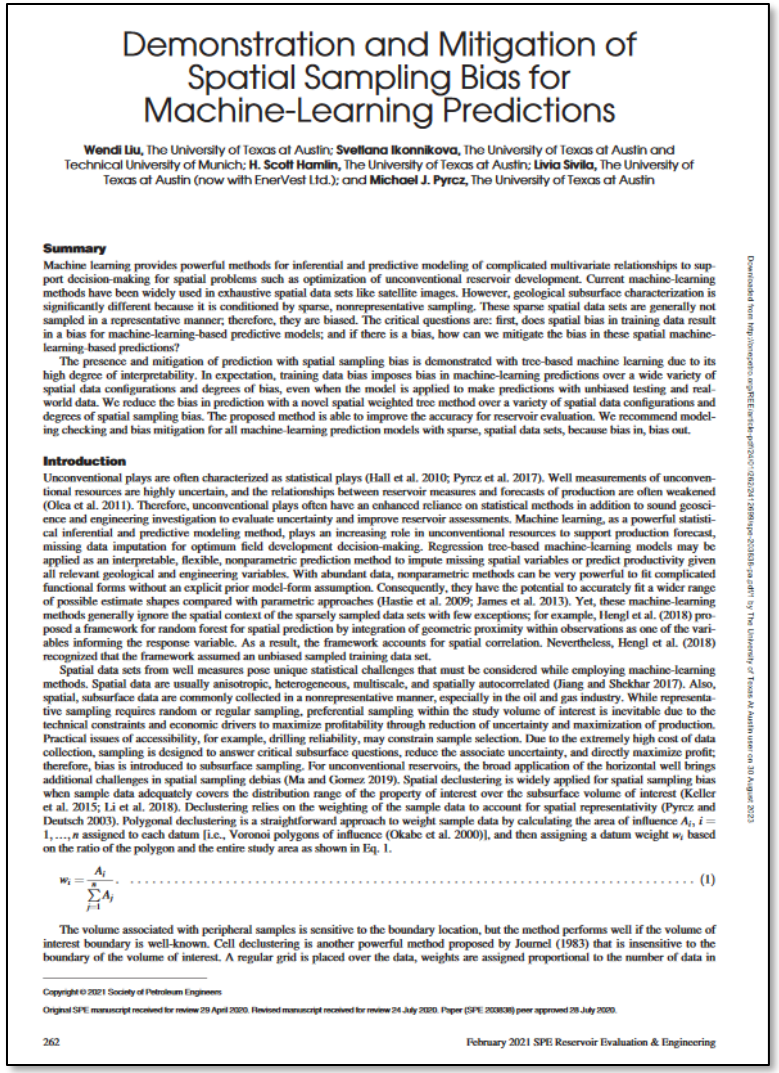
Representative Distributions for Production

- 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).

Impact of Spatial Bias on Machine Learning Models



Demonstration and Mitigation of Spatial Sampling Bias for Machine-Learning Predictions

Wendi Liu, The University of Texas at Austin; **Svetlana Ikonnikova**, The University of Texas at Austin and Technical University of Munich; **H. Scott Hamlin**, The University of Texas at Austin; **Livia Sivila**, The University of Texas at Austin (now with EnerVest Ltd.); and **Michael J. Pyrcz**, The University of Texas at Austin

Summary

Machine learning provides powerful methods for inferential and predictive modeling of complicated multivariate relationships to support decision-making for spatial problems such as optimization of unconventional reservoir development. Current machine-learning methods have been widely used in exhaustive spatial data sets like satellite images. However, geological subsurface characterization is significantly different because it is conditioned by sparse, nonrepresentative sampling. These sparse spatial data sets are generally not sampled in a representative manner; therefore, they are biased. The critical questions are: first, does spatial bias in training data result in a bias for machine-learning-based predictive models; and if there is a bias, how can we mitigate the bias in these spatial machine-learning-based predictions?

The presence and mitigation of prediction with spatial sampling bias is demonstrated with tree-based machine learning due to its high degree of interpretability. In expectation, training data bias imposes bias in machine-learning predictions over a wide variety of spatial data configurations and degrees of bias, even when the model is applied to make predictions with unbiased testing and real-world data. We reduce the bias in prediction with a novel spatial weighted tree method over a variety of spatial data configurations and degrees of spatial sampling bias. The proposed method is able to improve the accuracy for reservoir evaluation. We recommend modeling checking and bias mitigation for all machine-learning prediction models with sparse, spatial data sets, because bias in, bias out.

Introduction

Unconventional plays are often characterized as statistical plays (Hall et al. 2010; Pyrcz et al. 2017). Well measurements of unconventional resources are highly uncertain, and the relationships between reservoir measures and forecasts of production are often weakened (Olea et al. 2011). Therefore, unconventional plays often have an enhanced reliance on statistical methods in addition to sound geoscience and engineering investigation to evaluate uncertainty and improve reservoir assessments. Machine learning, as a powerful statistical inferential and predictive modeling method, plays an increasing role in unconventional resources to support production forecast, missing data imputation for optimum field development decision-making. Regression tree-based machine-learning models may be applied as an interpretable, flexible, nonparametric prediction method to impute missing spatial variables or predict productivity given all relevant geological and engineering variables. With abundant data, nonparametric methods can be very powerful to fit complicated functional forms without an explicit prior model-form assumption. Consequently, they have the potential to accurately fit a wider range of possible estimate shapes compared with parametric approaches (Hastie et al. 2009; James et al. 2013). Yet, these machine-learning methods generally ignore the spatial context of the sparsely sampled data sets with few exceptions; for example, Hengl et al. (2018) proposed a framework for random forest for spatial prediction by integration of geometric proximity within observations as one of the variables informing the response variable. As a result, the framework accounts for spatial correlation. Nevertheless, Hengl et al. (2018) recognized that the framework assumed an unbiased sampled training data set.

Spatial data sets from well measures pose unique statistical challenges that must be considered while employing machine-learning methods. Spatial data are usually anisotropic, heterogeneous, multiscale, and spatially autocorrelated (Jiang and Shekhar 2017). Also, spatial, subsurface data are commonly collected in a nonrepresentative manner, especially in the oil and gas industry. While representative sampling requires random or regular sampling, preferential sampling within the study volume of interest is inevitable due to the technical constraints and economic drivers to maximize profitability through reduction of uncertainty and maximization of production. Practical issues of accessibility, for example, drilling reliability, may constrain sample selection. Due to the extremely high cost of data collection, sampling is designed to answer critical subsurface questions, reduce the associate uncertainty, and directly maximize profit; therefore, bias is introduced to subsurface sampling. For unconventional reservoirs, the broad application of the horizontal well brings additional challenges in spatial sampling debias (Ma and Gomez 2019). Spatial declustering is widely applied for spatial sampling bias when sample data adequately covers the distribution range of the property of interest over the subsurface volume of interest (Keller et al. 2015; Li et al. 2018). Declustering relies on the weighting of the sample data to account for spatial representativity (Pyrcz and Deutsch 2003). Polygonal declustering is a straightforward approach to weight sample data by calculating the area of influence A_i, $i = 1, \ldots, n$ assigned to each datum [i.e., Voronoi polygons of influence (Okabe et al. 2000)], and then assigning a datum weight w_i based on the ratio of the polygon and the entire study area as shown in Eq. 1.

$$w_i = \frac{A_i}{\sum_{j=1}^{n} A_j}. \quad \ldots\ldots (1)$$

The volume associated with peripheral samples is sensitive to the boundary location, but the method performs well if the volume of interest boundary is well-known. Cell declustering is another powerful method proposed by Journel (1983) that is insensitive to the boundary of the volume of interest. A regular grid is placed over the data, weights are assigned proportional to the number of data in

Original SPE manuscript received for review 29 April 2020. Revised manuscript received for review 24 July 2020. Paper (SPE 203838) peer approved 28 July 2020.

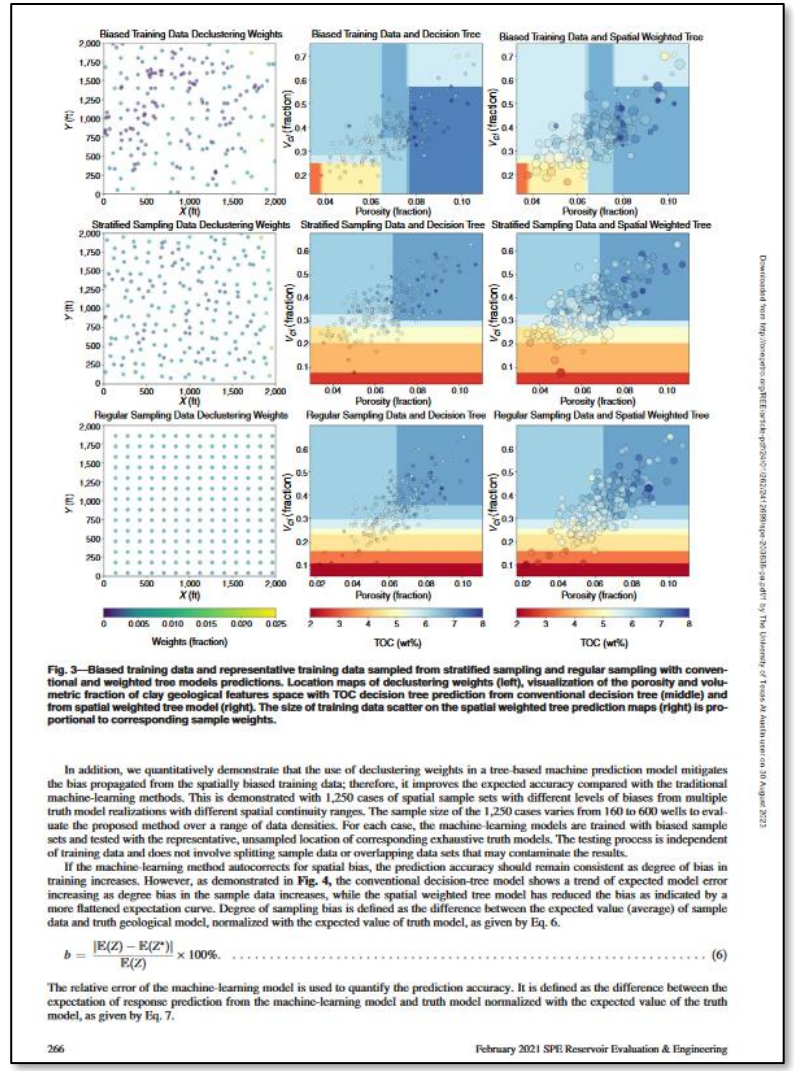


Fig. 3—Biased training data and representative training data sampled from stratified sampling and regular sampling with conventional and weighted tree models predictions. Location maps of declustering weights (left), visualization of the porosity and volumetric fraction of clay geological features space with TOC decision tree prediction from conventional decision tree (middle) and from spatial weighted tree model (right). The size of training data scatter on the spatial weighted tree prediction maps (right) is proportional to corresponding sample weights.

In addition, we quantitatively demonstrate that the use of declustering weights in a tree-based machine prediction model mitigates the bias propagated from the spatially biased training data; therefore, it improves the expected accuracy compared with the traditional machine-learning methods. This is demonstrated with 1,250 cases of spatial sample sets with different levels of biases from multiple truth model realizations with different spatial continuity ranges. The sample size of the 1,250 cases varies from 160 to 600 wells to evaluate the proposed method over a range of data densities. For each case, the machine-learning models are trained with biased sample sets and tested with the representative, unsampled location of corresponding exhaustive truth models. The testing process is independent of training data and does not involve splitting sample data or overlapping data sets that may contaminate the results.

If the machine-learning method autocorrects for spatial bias, the prediction accuracy should remain consistent as degree of bias in training increases. However, as demonstrated in **Fig. 4**, the conventional decision-tree model shows a trend of expected model error increasing as degree bias in the sample data increases, while the spatial weighted tree model has reduced the bias as indicated by a more flattened expectation curve. Degree of sampling bias is defined as the difference between the expected value (average) of sample data and truth geological model, normalized with the expected value of truth model, as given by Eq. 6.

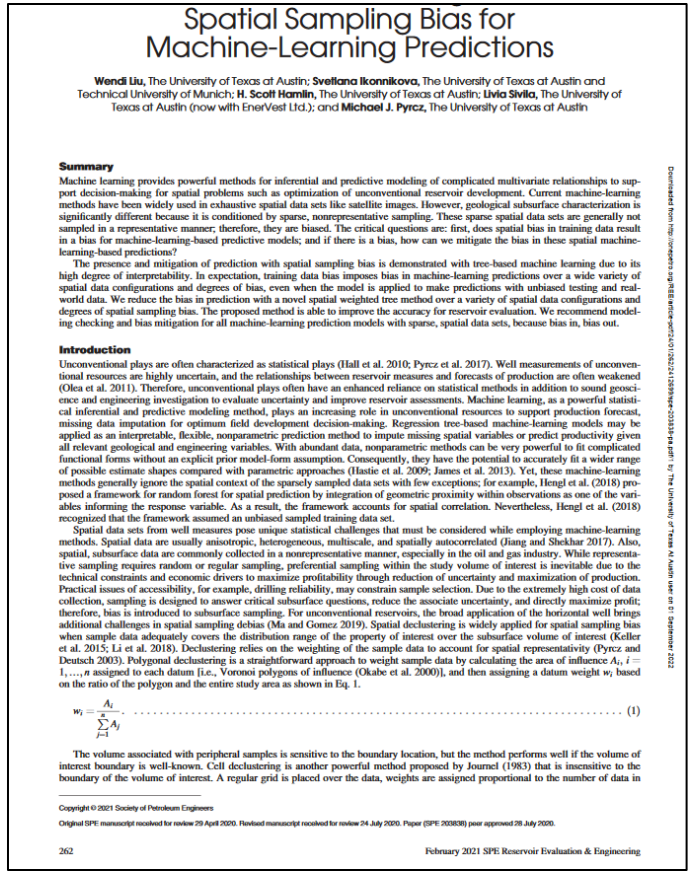
$$b = \frac{|E(Z) - E(Z^*)|}{E(Z)} \times 100\%. \quad \ldots\ldots (6)$$

The relative error of the machine-learning model is used to quantify the prediction accuracy. It is defined as the difference between the expectation of response prediction from the machine-learning model and truth model normalized with the expected value of the truth model, as given by Eq. 7.

Impact of spatial sampling bias on a predictive machine learning model and the use of declustering weights to correct the model (Liu et al., 2021).

Data Debiasing 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.*



Spatial Sampling Bias for Machine-Learning Predictions

Wendi Liu, The University of Texas at Austin; **Svetlana Ikonnikova,** The University of Texas at Austin and Technical University of Munich; **H. Scott Hamlin,** The University of Texas at Austin; **Livia Sivila,** The University of Texas at Austin (now with EnerVest Ltd.); and **Michael J. Pyrcz,** The University of Texas at Austin

Summary

Machine learning provides powerful methods for inferential and predictive modeling of complicated multivariate relationships to support decision-making for spatial problems such as optimization of unconventional reservoir development. Current machine-learning methods have been widely used in exhaustive spatial data sets like satellite images. However, geological subsurface characterization is significantly different because it is conditioned by sparse, nonrepresentative sampling. These sparse spatial data sets are generally not sampled in a representative manner; therefore, they are biased. The critical questions are: first, does spatial bias in training data result in a bias for machine-learning-based predictive models; and if there is a bias, how can we mitigate the bias in these spatial machine-learning-based predictions?

The presence and mitigation of prediction with spatial sampling bias is demonstrated with tree-based machine learning due to its high degree of interpretability. In expectation, training data bias imposes bias in machine-learning predictions over a wide variety of spatial data configurations and degrees of bias, even when the model is applied to make predictions with unbiased testing and real-world data. We reduce the bias in prediction with a novel spatial weighted tree method over a variety of spatial data configurations and degrees of spatial sampling bias. The proposed method is able to improve the accuracy for reservoir evaluation. We recommend modeling checking and bias mitigation for all machine-learning prediction models with sparse, spatial data sets, because bias in, bias out.

Introduction

Unconventional plays are often characterized as statistical plays (Hall et al. 2010; Pyrcz et al. 2017). Well measurements of unconventional resources are highly uncertain, and the relationships between reservoir measures and forecasts of production are often weakened (Olea et al. 2011). Therefore, unconventional plays often have an enhanced reliance on statistical methods in addition to sound geoscience and engineering investigation to evaluate uncertainty and improve reservoir assessments. Machine learning, as a powerful statistical inferential and predictive modeling method, plays an increasing role in unconventional resources to support production forecast, missing data imputation for optimum field development decision-making. Regression tree-based machine-learning models may be applied as an interpretable, flexible, nonparametric prediction method to impute missing spatial variables or predict productivity given all relevant geological and engineering variables. With abundant data, nonparametric methods can be very powerful to fit complicated functional forms without an explicit prior model-form assumption. Consequently, they have the potential to accurately fit a wider range of possible estimate shapes compared with parametric approaches (Hastie et al. 2009; James et al. 2013). Yet, these machine-learning methods generally ignore the spatial context of the sparsely sampled data sets with few exceptions; for example, Hengl et al. (2018) proposed a framework for random forest for spatial prediction by integration of geometric proximity within observations as one of the variables informing the response variable. As a result, the framework accounts for spatial correlation. Nevertheless, Hengl et al. (2018) recognized that the framework assumed an unbiased sampled training data set.

Spatial data sets from well measures pose unique statistical challenges that must be considered while employing machine-learning methods. Spatial data are usually anisotropic, heterogeneous, multiscale, and spatially autocorrelated (Jiang and Shekhar 2017). Also, spatial, subsurface data are commonly collected in a nonrepresentative manner, especially in the oil and gas industry. While representative sampling requires random or regular sampling, preferential sampling within the study volume of interest is inevitable due to the technical constraints and economic drivers to maximize profitability through reduction of uncertainty and maximization of production. Practical issues of accessibility, for example, drilling reliability, may constrain sample selection. Due to the extremely high cost of data collection, sampling is designed to answer critical subsurface questions, reduce the associate uncertainty, and directly maximize profit; therefore, bias is introduced to subsurface sampling. For unconventional reservoirs, the broad application of the horizontal well brings additional challenges in spatial sampling debias (Ma and Gomez 2019). Spatial declustering is widely applied for spatial sampling bias when sample data adequately covers the distribution range of the property of interest over the subsurface volume of interest (Keller et al. 2015; Li et al. 2018). Declustering relies on the weighting of the sample data to account for spatial representativity (Pyrcz and Deutsch 2003). Polygonal declustering is a straightforward approach to weight sample data by calculating the area of influence A_i, $i = 1,\ldots,n$ assigned to each datum [i.e., Voronoi polygons of influence (Okabe et al. 2000)], and then assigning a datum weight w_i based on the ratio of the polygon and the entire study area as shown in Eq. 1.

$$w_i = \frac{A_i}{\sum_{j=1}^{n} A_j}. \quad \ldots\ldots (1)$$

The volume associated with peripheral samples is sensitive to the boundary location, but the method performs well if the volume of interest boundary is well-known. Cell declustering is another powerful method proposed by Journel (1983) that is insensitive to the boundary of the volume of interest. A regular grid is placed over the data, weights are assigned proportional to the number of data in

Original SPE manuscript received for review 29 April 2020. Revised manuscript received for review 24 July 2020. Paper (SPE 203838) peer approved 28 July 2020.

Recent paper on the use of declustering weights in machine learning-based prediction models.

PGE 383 Subsurface Machine Learning

Lecture 4: Data Preparation

Lecture outline:

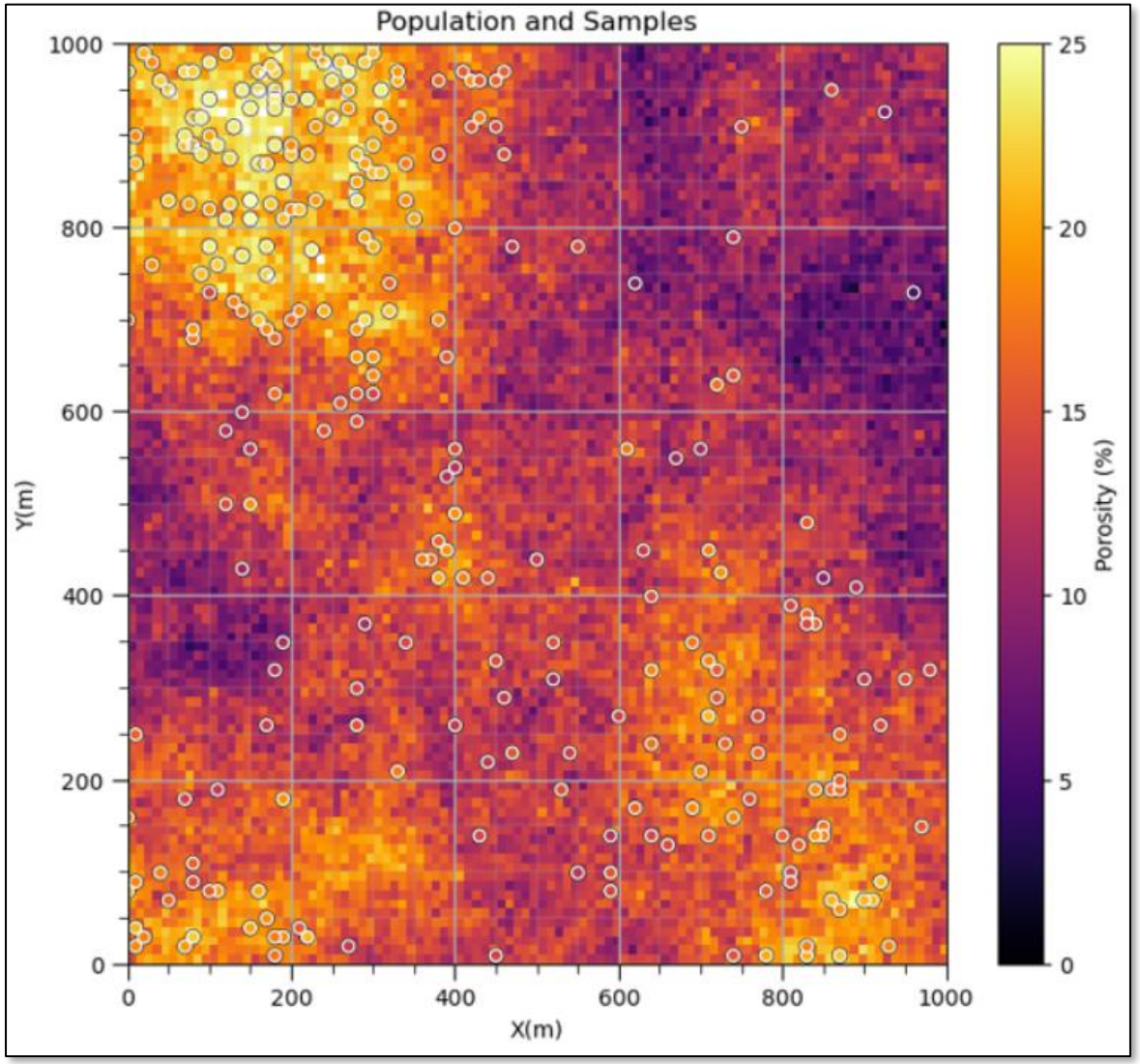
- **Quantifying Uncertainty**

Instructor: Michael Pyrcz, the University of Texas at Austin

Bootstrap

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



Samples and population (left), population distribution (upper right) and sample distribution (lower right), from Bootstrap chapter of Applied Geostatistics in Python e-book.

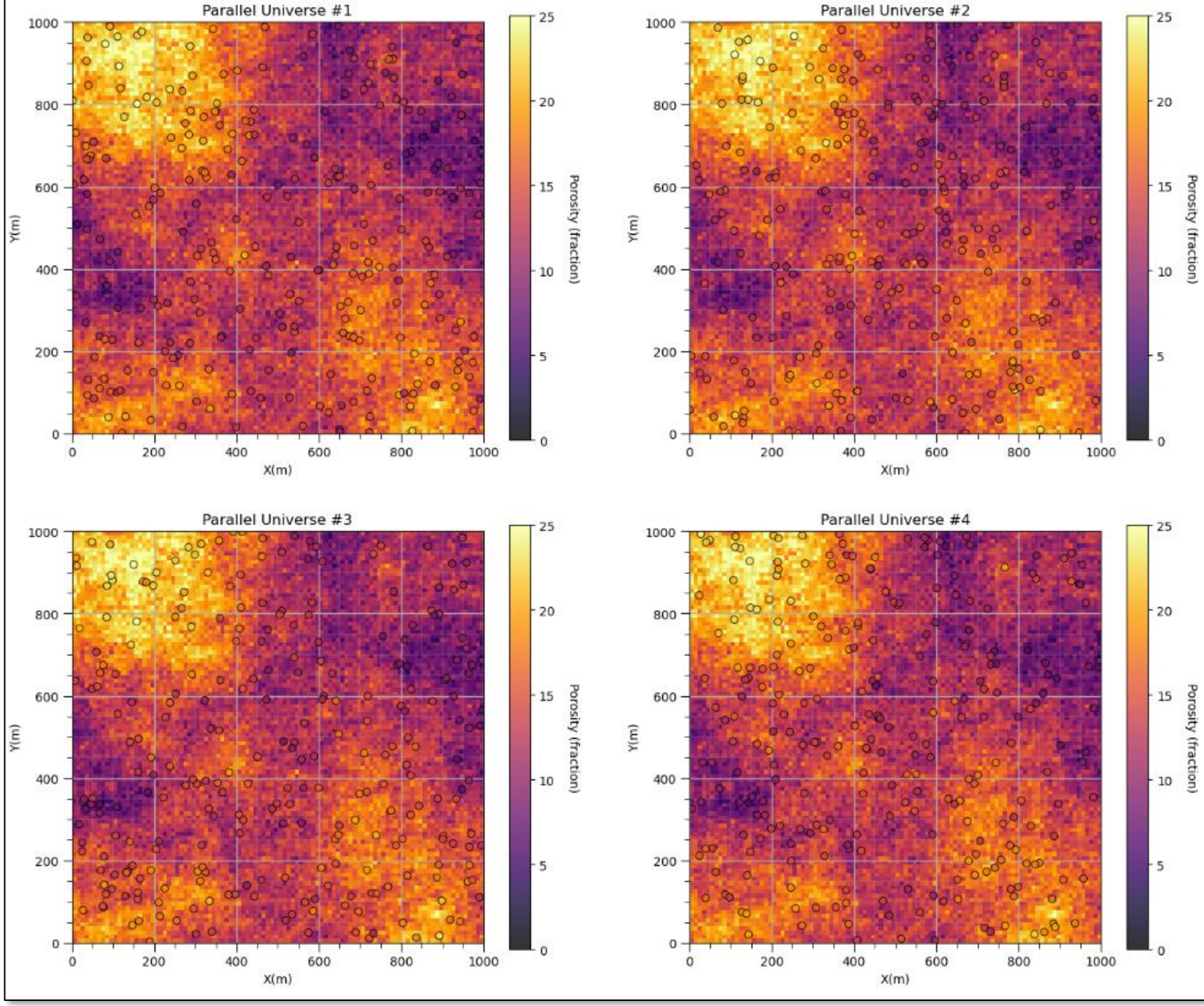
Bootstrap

Uncertainty Due to Data Paucity

What if we had ‘L’ different datasets? L parallel universes where we collected n samples from the inaccessible truth (the population).

We only exist in 1 universe.

- this is not possible.



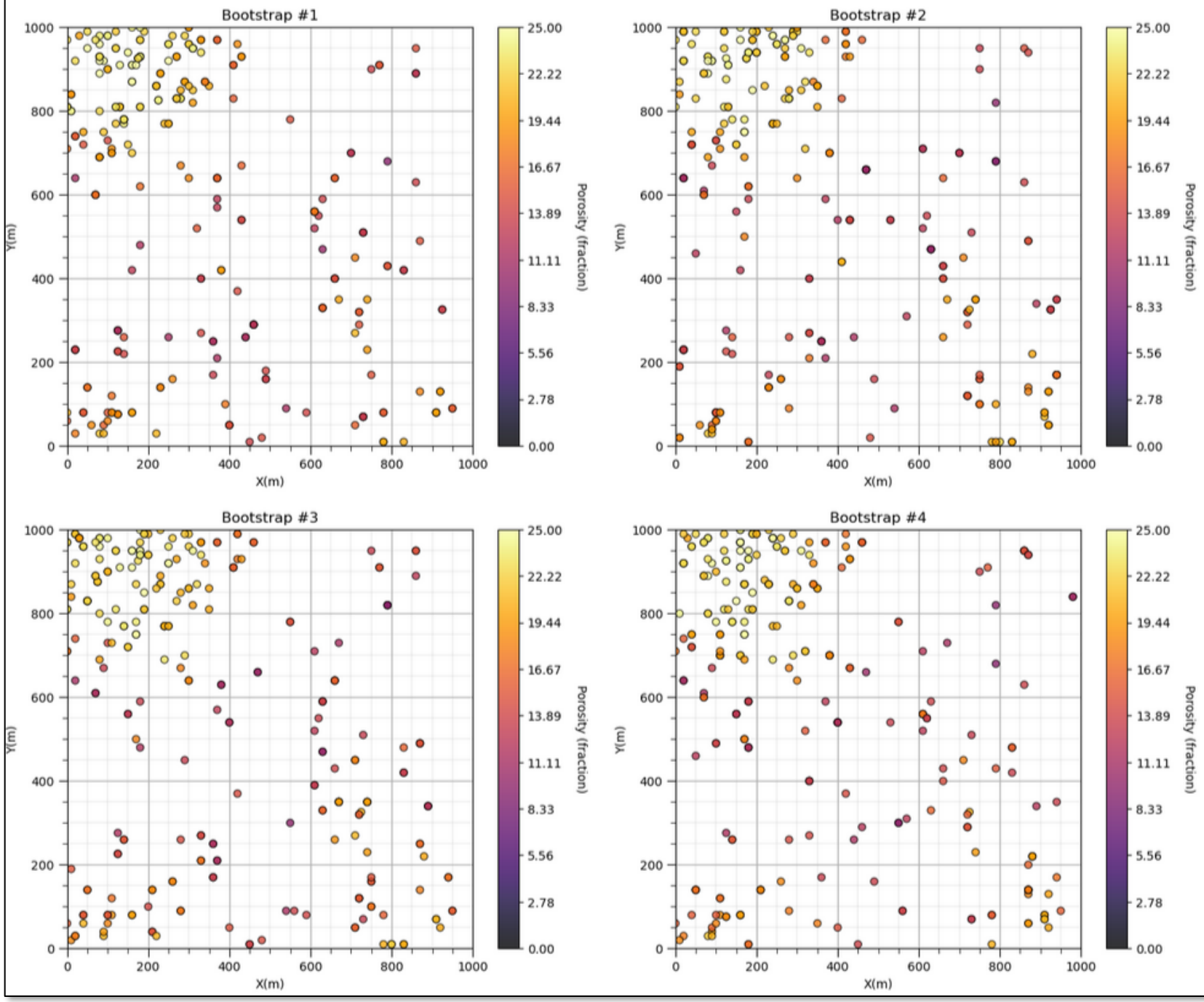
Multiple dataset realizations from the truth population, from Bootstrap chapter of Applied Geostatistics in Python e-book.

Bootstrap

Uncertainty Due to Data Paucity

Instead we sample n times from the dataset with replacement

- bootstrap realizations of the data
- vary by due to some samples being left out and others sampled multiple times.



Multiple dataset bootstrap realizations, from Bootstrap chapter of Applied Geostatistics in Python e-book.

Bootstrap Definition

Bootstrap
- method to assess the uncertainty in a sample statistic by repeated random sampling with replacement
- simulating the sampling process to acquire dataset realizations

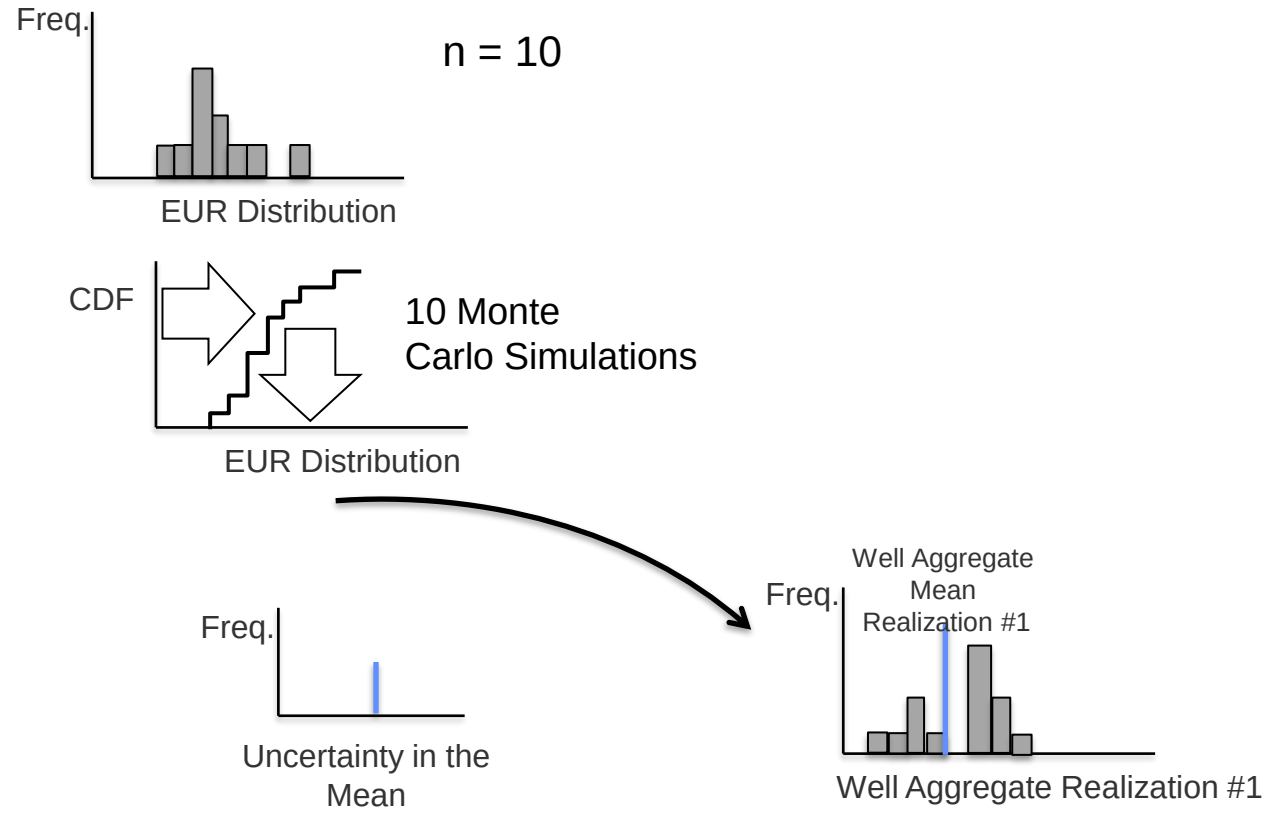
Assumptions
- sufficient, representative sampling

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

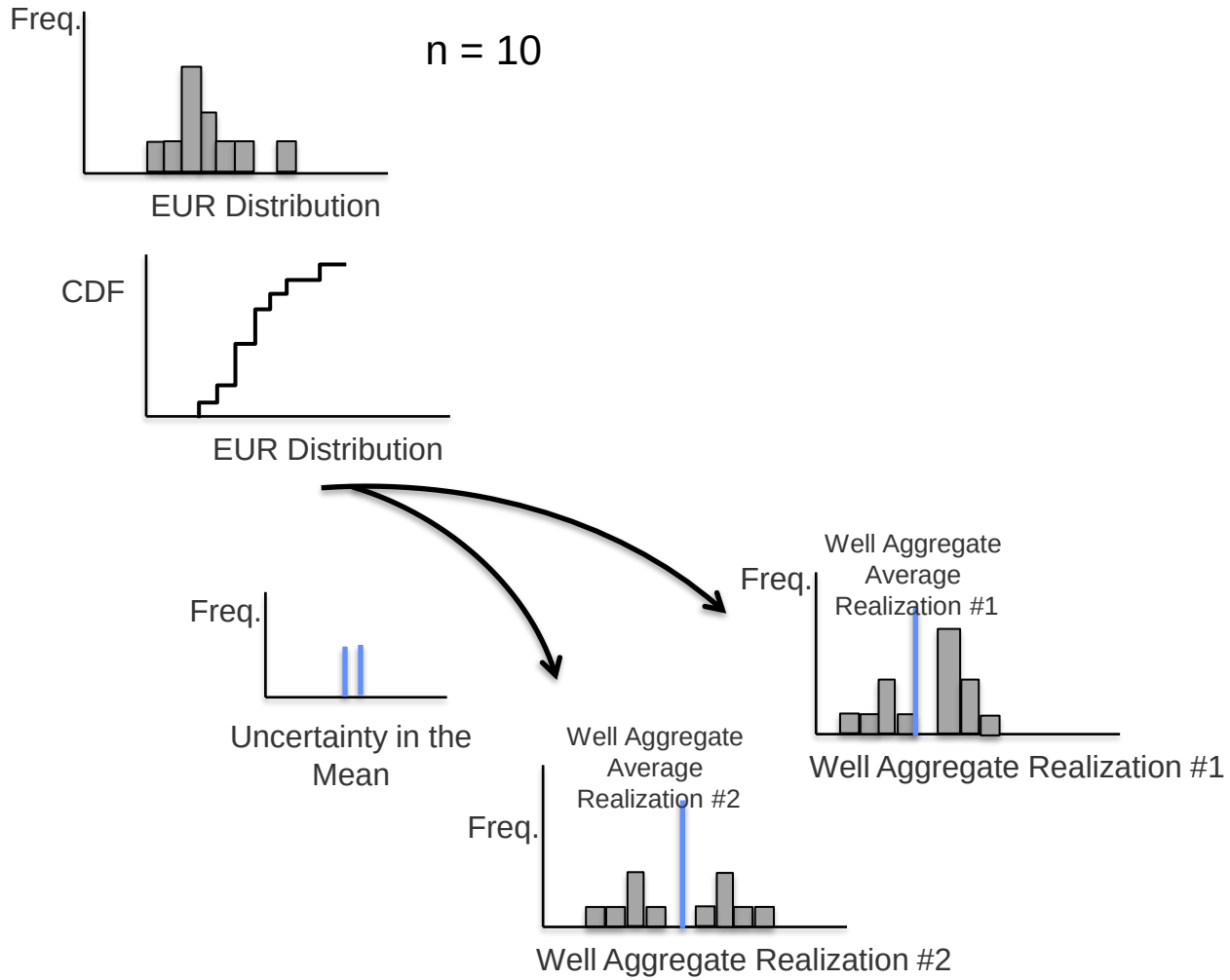
Bootstrap

Bootstrap for Uncertainty in EUR Mean



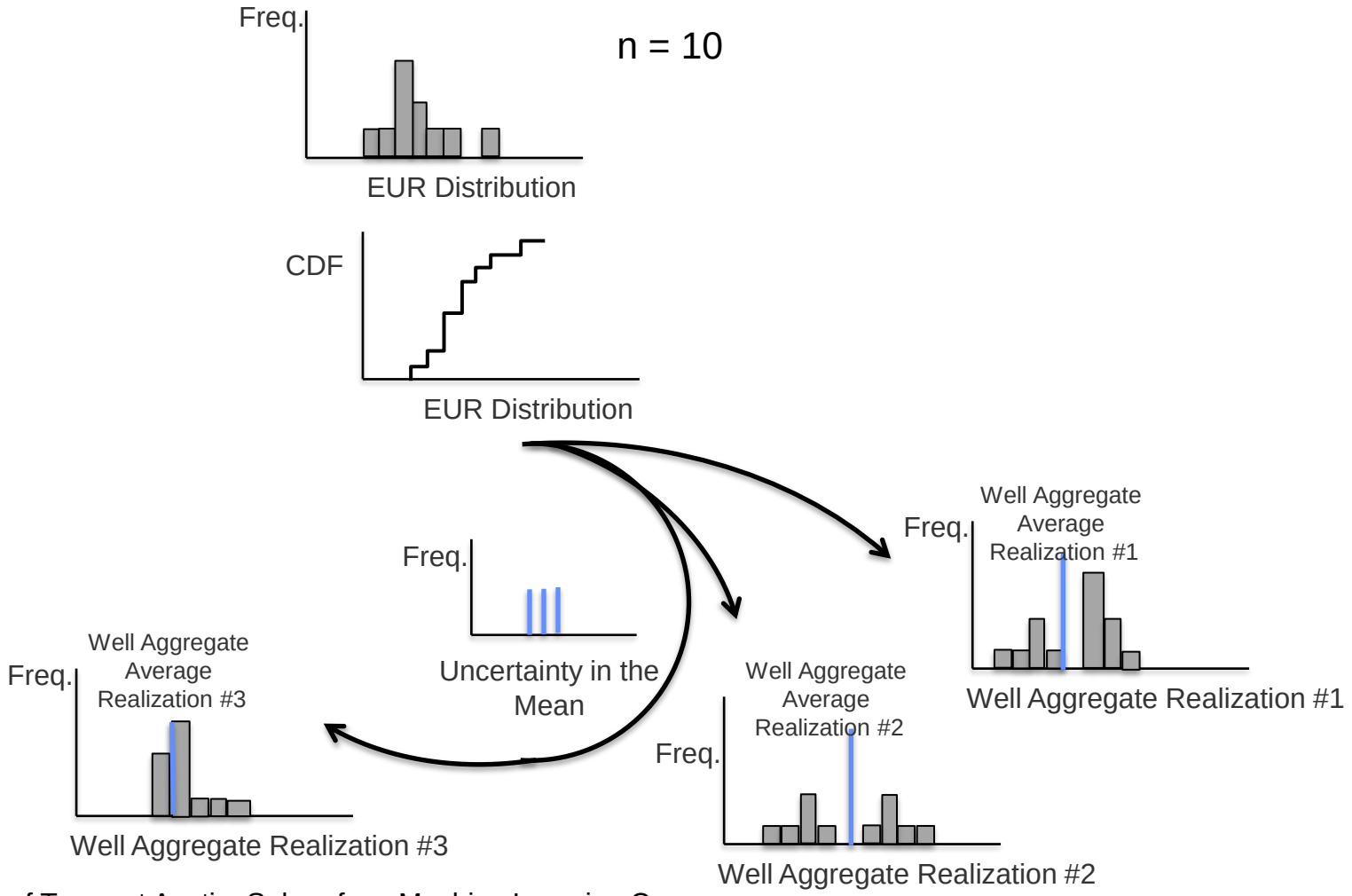
Bootstrap

Bootstrap for Uncertainty in EUR Mean



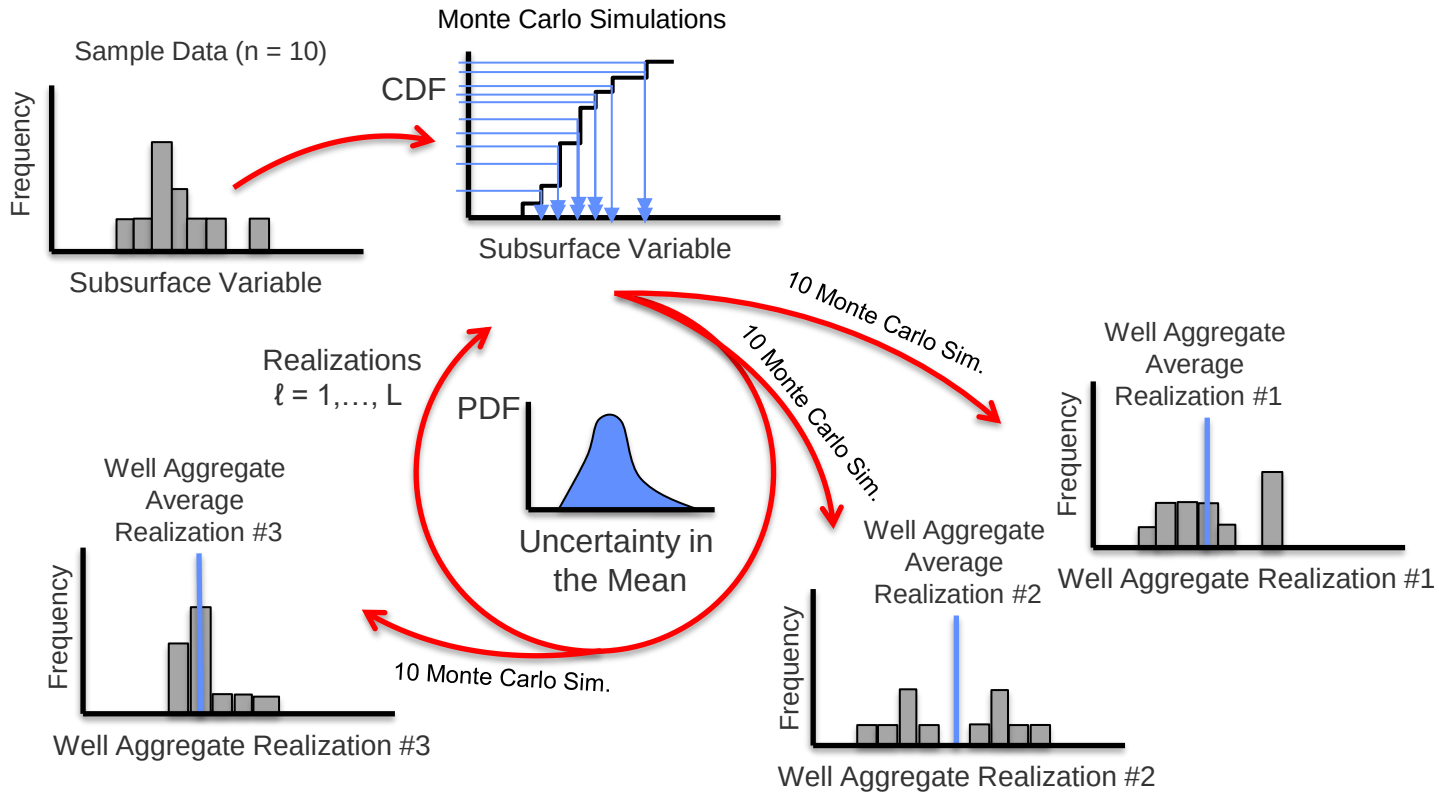
Bootstrap

Bootstrap for Uncertainty in EUR Mean



Bootstrap

Bootstrap for Uncertainty in EUR Mean



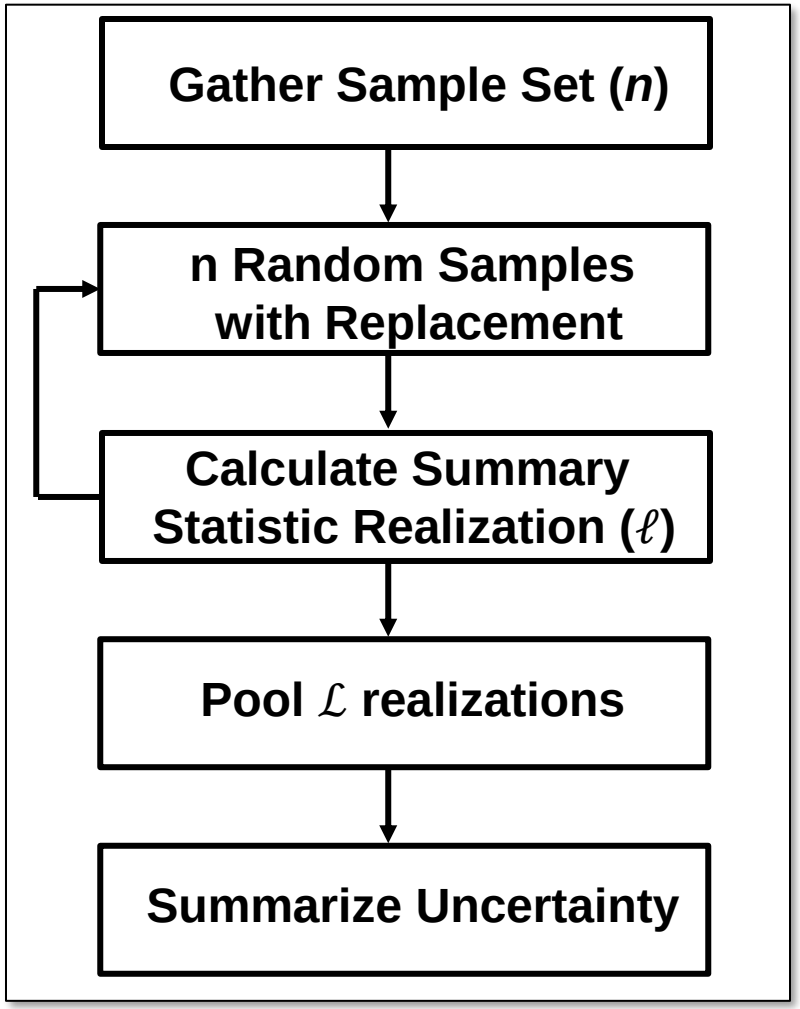
Bootstrap

Bootstrap Approach developed by 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).



Bootstrap workflow for uncertaint in statistic.

Bootstrap

You now know about one of the most powerful statistical tools!

- 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 even do bootstrap in Excel.

Boostrap Demonstration

Michael Pyrcz, the University of Texas at Austin, @GeostatsGuy

Method: Bootstrap uncertainty in distribution statistics by resampling with replacement from the initial sample dataset.

Details: Resample with replacement n times to build a realization. Repeat for L=1,000 realizations. Calculate various sum

Observations: Uncertainty will decrease as number of samples (n) increases. Bootstrap assumes no spatial continuity bet

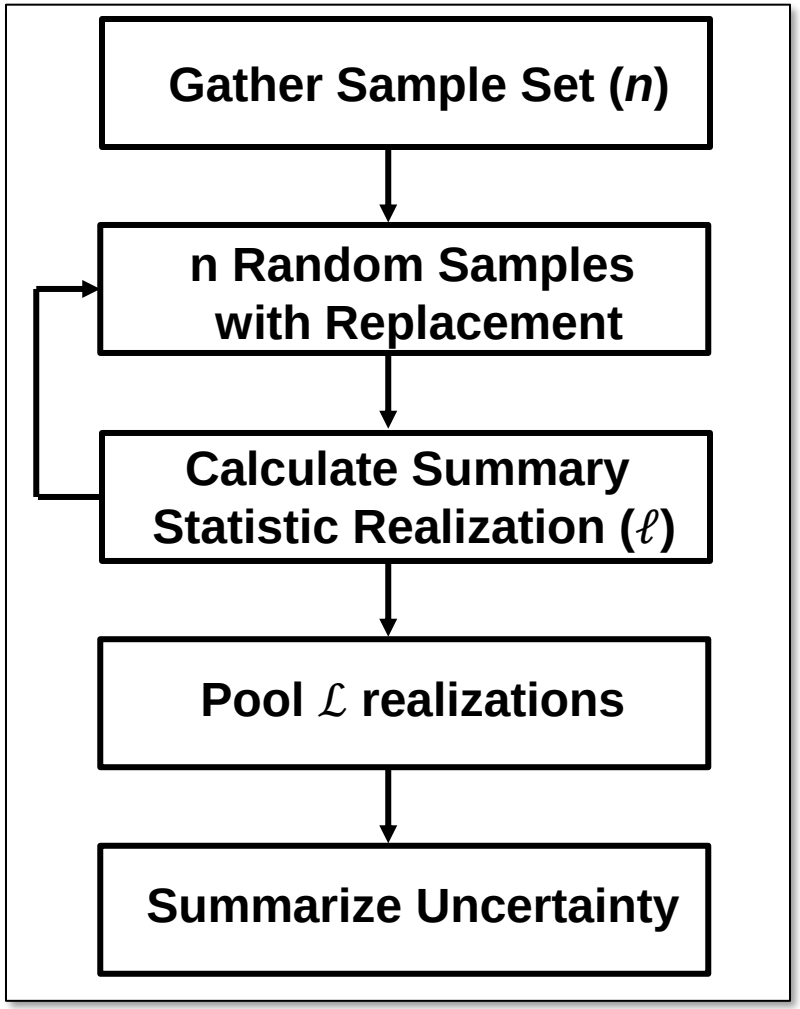
Sample Data Set

Index	P-value	Norm[0,1]
1	0.586	0.218
2	0.741	0.647
19	0.851	1.040
20	0.219	-0.776

Samples	Real1	Real2	Real3	Real4	Real5	Real6	Real7	Real8	Real9	Real10	Real11	Real12	Real13	Real14	Real15	Real16	Real17
1	-0.642	0.784	0.218	-0.023	-0.776	1.040	1.967	-0.025	-0.025	-0.570	1.040	-1.155	1.973	-1.652	-0.230	0.890	0.218
2	0.218	1.040	-1.652	0.890	-0.025	0.784	-1.155	-0.709	-1.155	0.167	-1.652	-0.023	-0.023	-0.776	0.890	-1.155	-0.025
19	0.890	-1.155	0.473	0.647	0.167	0.508	0.218	0.890	-0.776	0.473	0.473	-0.023	1.040	0.784	0.473	-0.709	-1.652
20	0.473	-0.025	0.218	0.890	-0.230	0.647	-0.025	0.647	1.967	0.167	0.708	0.708	1.040	0.218	-1.155	-1.155	-0.709

Statistic	Real1	Real2	Real3	Real4	Real5	Real6	Real7	Real8	Real9	Real10	Real11	Real12	Real13	Real14	Real15	Real16	Real17
Mean	0.30	0.45	0.37	0.16	0.25	0.10	0.03	0.13	-0.11	0.10	0.10	0.11	0.30	-0.12	-0.07	0.06	0.05
St.Dev.	0.99	0.95	0.71	0.96	0.87	0.93	0.78	0.84	1.09	0.52	0.82	0.83	0.95	0.87	0.87	0.85	1.06
Min	-1.65	-1.16	-1.65	-1.65	-1.65	-1.65	-1.16	-1.16	-1.65	-0.78	-1.65	-1.65	-1.65	-1.65	-1.65	-1.16	-1.65
Max	1.97	1.97	1.97	1.97	1.97	1.04	1.97	1.97	1.97	1.04	1.04	1.04	1.97	1.04	1.97	1.97	1.97
P10	-1.20	-0.72	-0.26	-0.81	-0.66	-1.20	-0.81	-0.72	-1.20	-0.64	-1.16	-1.20	-0.81	-1.65	-1.16	-1.16	-1.65
P50	0.49	0.58	0.47	0.07	0.32	0.51	0.10	-0.02	-0.13	0.17	0.35	0.32	0.58	0.22	-0.02	0.22	0.10
P90	1.13	1.97	0.90	1.00	1.00	1.04	0.73	1.04	1.97	0.65	0.90	0.90	1.13	0.78	0.79	0.90	1.13

Bootstrap in Excel with random and DataBase functions, Bootstrap_Demo.xlsx.

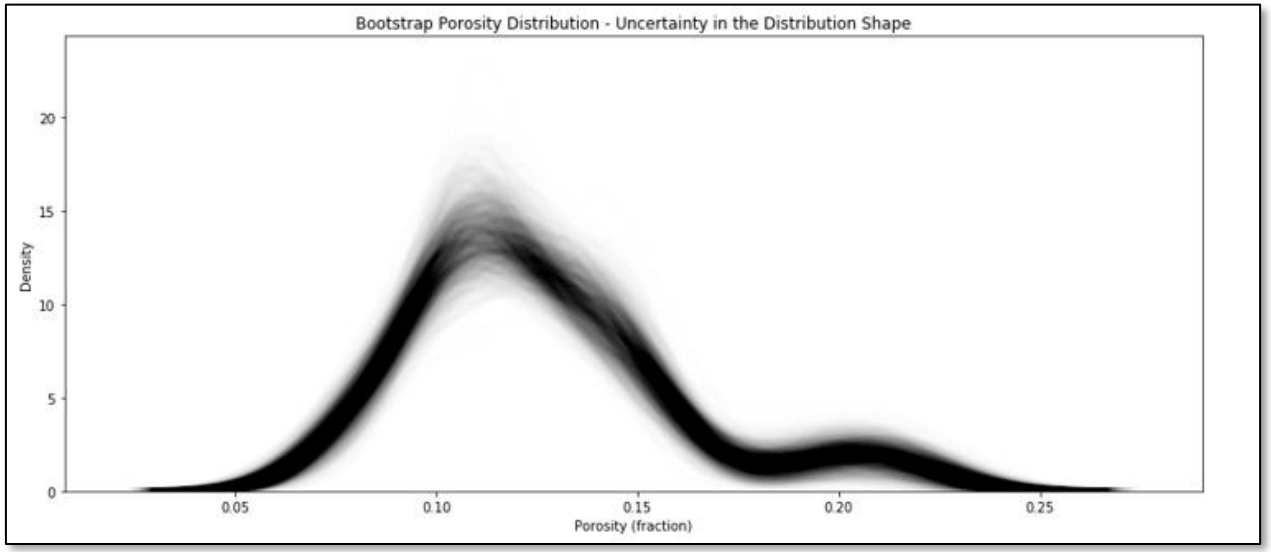


Bootstrap workflow for uncertaint in statistic.

Bootstrap Demonstration in Python

Demonstration workflow for boostrap for uncertainty in statistics and models,

- a variety of example statistics



Bootstrap for uncertainty in statistics, file is SubsurfaceDataAnalytics_bootstrap.ipynb.

Subsurface Data Analytics

Bootstrap for Subsurface Data Analytics in Python

Michael Pyrcz, Associate Professor, University of Texas at Austin

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

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

More on Bootstrap

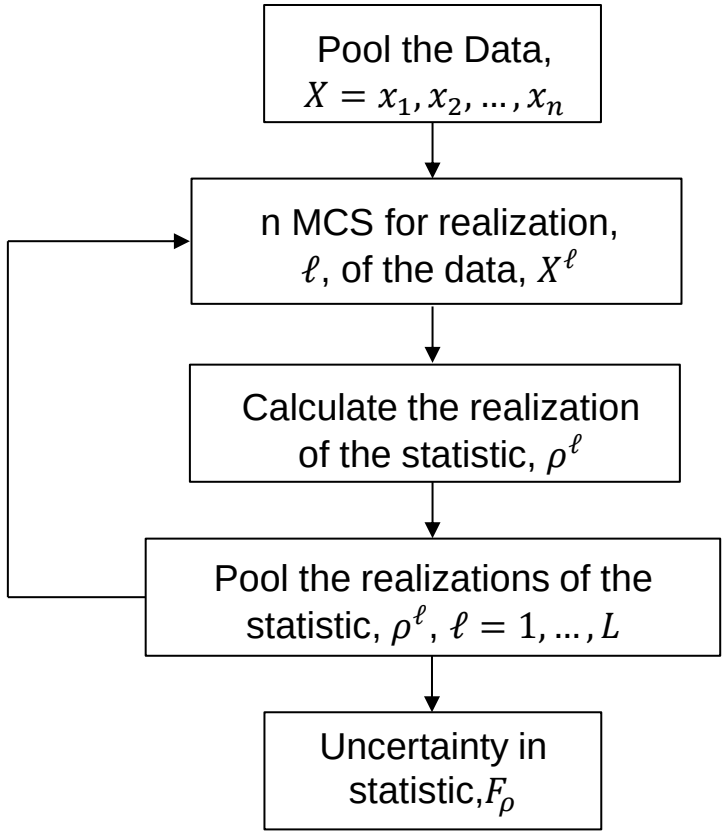
Let me reinforce, the bootstrap approach may be applied to calculate uncertainty in any statistic, from new realizations of the data.

- We can even bootstrap ML models, known as bagging. More later.

Notation,

- Note, $P_k(0.5)^\ell$ is the 50th percentile of permeability, k, bootstrap data realization, ℓ.

- and ${s_1^2}^\ell$ is the sample variance of the 1st bootstrap dataset's ℓ realization.



ℓ, realization of the data distribution(s)

50th Percentile $P_k(50)^\ell$

Dykstra Parsons $dp^\ell = \frac{P_k(50)^\ell - P_k(16)^\ell}{P_k(50)^\ell}$

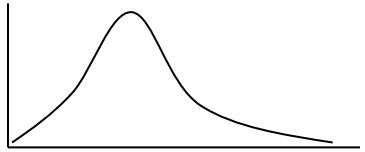
Welch's t-statistic

2 paired distributions $\hat{t}^\ell = \frac{{\bar{x}_1}^\ell - {\bar{x}_2}^\ell}{\sqrt{\left(\frac{{s_1^2}^\ell}{n_1} + \frac{{s_2^2}^\ell}{n_2}\right)}}$

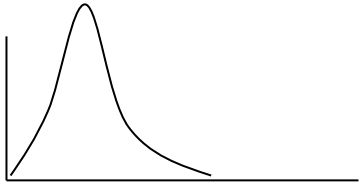
$P_k(50)^\ell$, dp^ℓ, $\hat{t}^\ell$ **realization of the statistic**

More on Bootstrap

For Dykstra Parsons, what if we did bootstrap on each input to our statistic?

$f_{P_K(50)}$ 

$P_k(50)^\ell, \ell = 1, \ldots, L_{boot}$

$f_{P_K(16)}$ 

$P_k(16)^\ell, \ell = 1, \ldots, L_{boot}$

Then applied Monte Carlo simulation from these distributions to:

$$dp^\ell = \frac{F^{-1}_{P_K(50)}(p^{\ell_1}) - F^{-1}_{P_K(16)}(p^{\ell_2})}{F^{-1}_{P_K(50)}(p^{\ell_1})} \qquad \ell_1 = 1, \ldots, L_{MCS}, \quad \ell_2 = 1, \ldots, L_{MCS}$$

This decorrelates the $P_k(16)^\ell$ and $P_k(50)^\ell$ realizations!

- Conventional MCS assumes independence, e.g. $OIP^\ell = \bar{\varphi}^\ell \cdot s_o^\ell \cdot V^\ell$, the average porosity, $\bar{\varphi}$, oil saturation, s_o and reservoir volume, V, are independent of each other.
- Why would we compare P16 and P50 from two different distributions? Make a new data set, then simulate the calculation of the statistic, that's bootstrap.

With Bootstrap Don't Stop Short! Calculate the Statistic with Each Dataset Realization!

Bootstrap
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*

Note, bootstrap is the applied for machine learning ensemble learning for reduced prediction model variance. This is known as **model bagging**, more on this later.

PGE 383 Subsurface Machine Learning

Lecture 4: Data Preparation

Lecture outline:

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

Instructor: Michael Pyrcz, the University of Texas at Austin