PGE 337 Data Analytics and Geostatistics

Lecture 16b: Model Checking

Short Summary of:

- Model Checking
- Checking Reproduction of Model Inputs
- Cross Validation of Estimates
- Cross Validation of Uncertainty Models

Michael Pyrcz, The University of Texas at Austin

Motivation

- We must check the performance of our models
- Bad models will lead to bad decisions
- There are many modeling decisions, model inputs; therefore, opportunities for blunders!

We must check the final product.

PGE 337 Data Analytics and Geostatistics

Lecture 16b: Model Checking

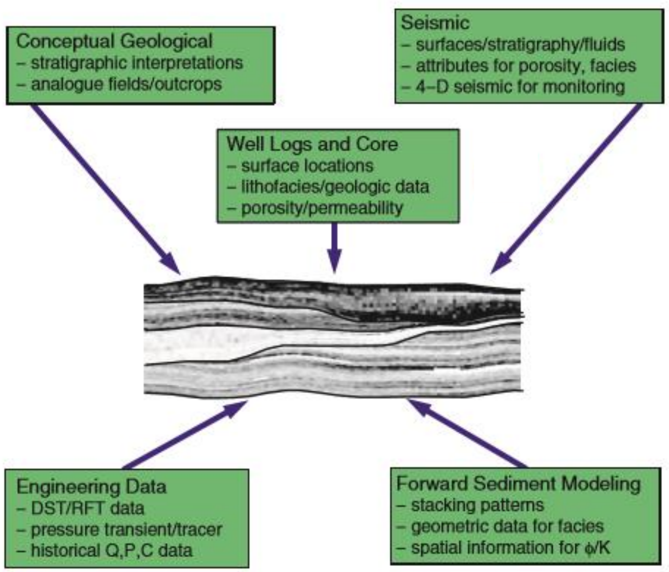
Short Summary of:

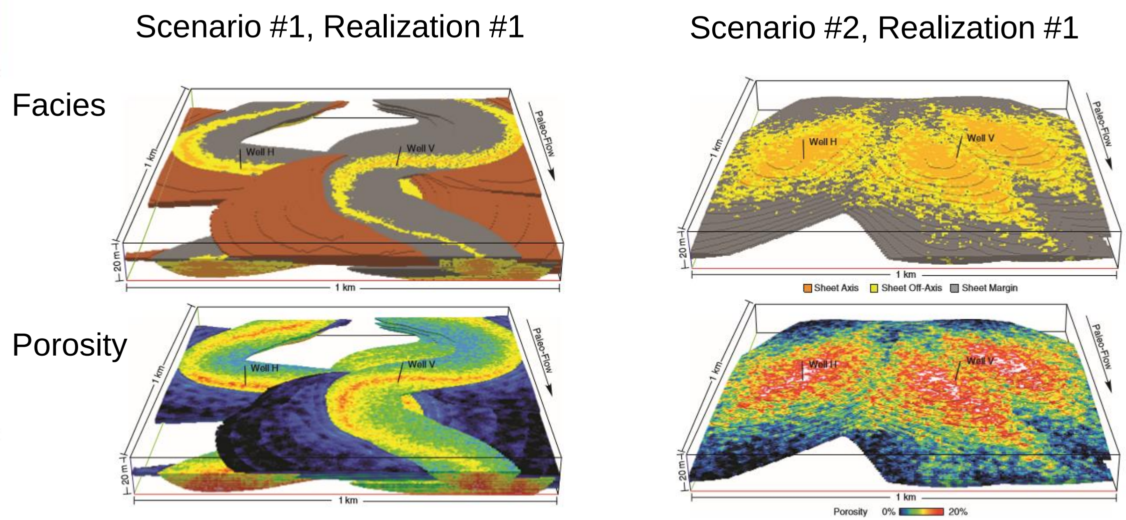
- Model Checking

Michael Pyrcz, The University of Texas at Austin

What is the Subsurface Model?

- Integration of all data sources
- Informed by statistics calculated from local data and analogs
- The results of many decisions, often result of complicated workflows
- Suite of models to represent uncertainty





Data Integration to build a spatial feature/realization.

Multiple spatial feature realizations/scenarios.

Spatial Model Checking

- Model Inputs: Data and Statistics Integration
  - Test the model to ensure the model inputs are honored in the models
  - E.g. input histogram and output histogram
- Accurate Spatial Estimates
  - Check the ability of the model to accurately predict away from the available sample data, over a variety of configurations, with accuracy
- Accurate Uncertainty Models
  - The uncertainty model is fair given the amount of information available and various sources of uncertainty

PGE 337 Data Analytics and Geostatistics

Lecture 16b: Model Checking

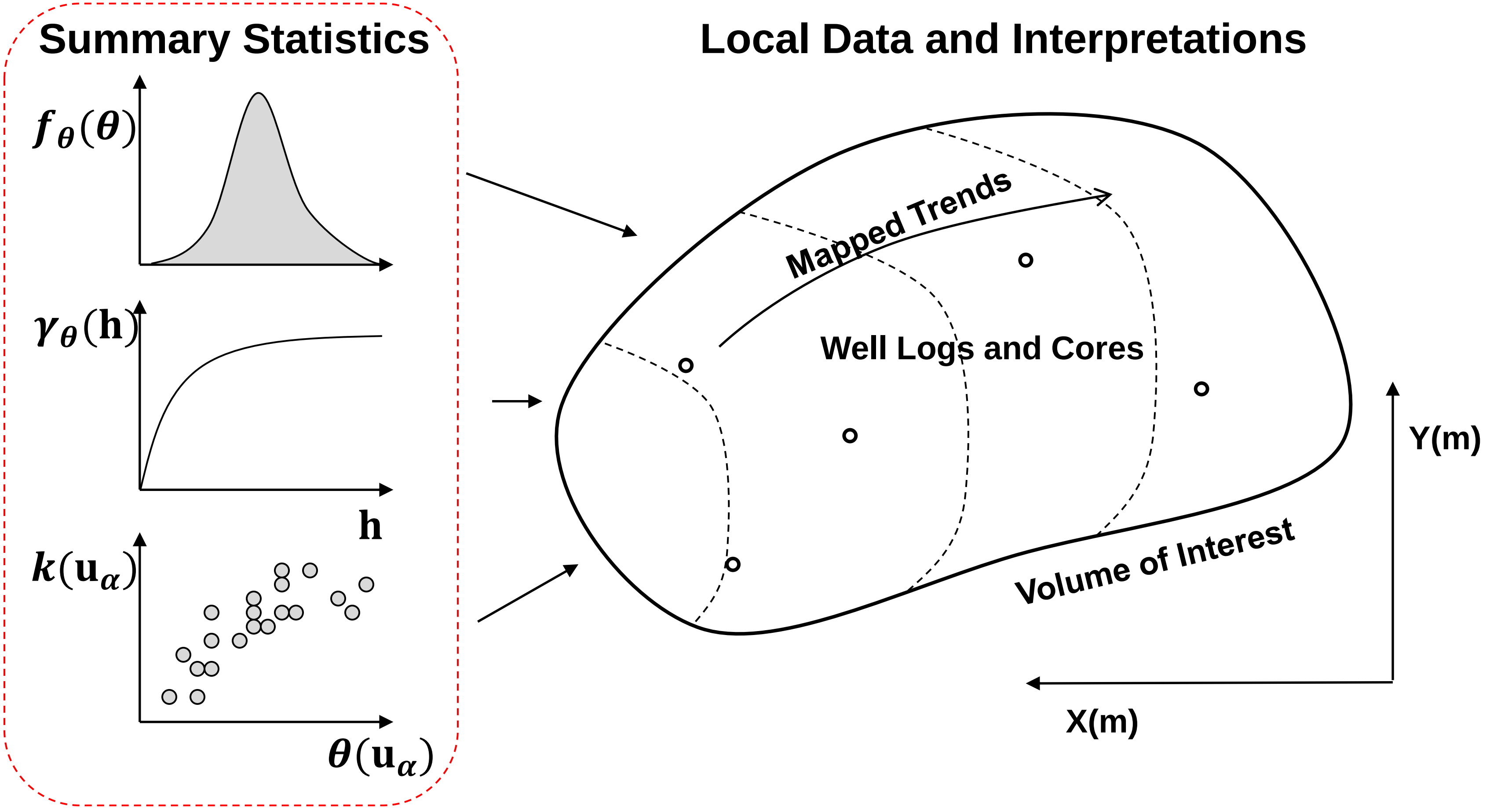
Short Summary of:

- Checking Reproduction of Model Inputs

Michael Pyrcz, The University of Texas at Austin

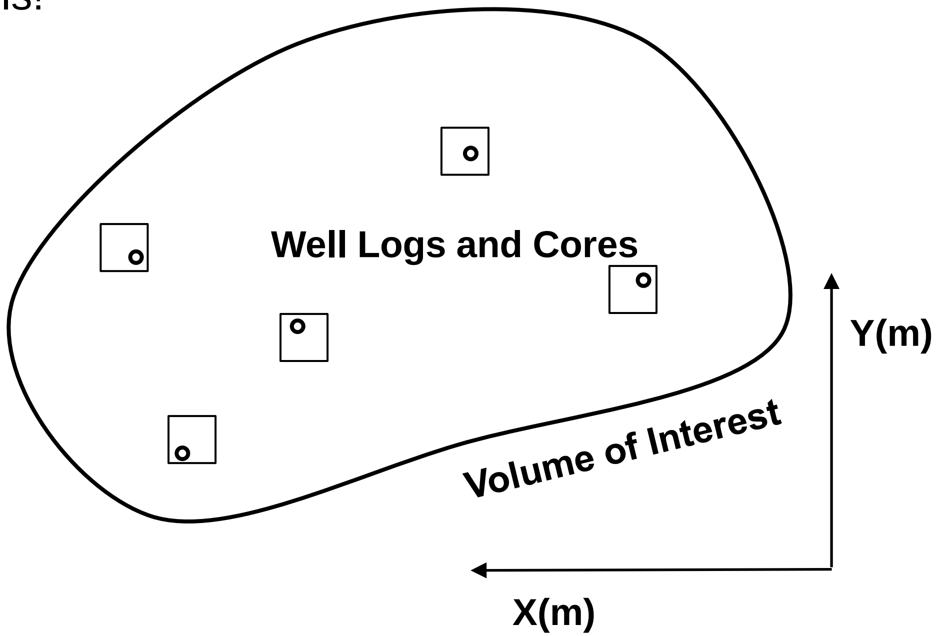
Subsurface Model Inputs

- Model Inputs
  - Local data and interpretations and input (geo)statistics



Checking Model Input (Geo)Statistics

- Check the data at the data locations!
  - Subsurface data is expensive, model lose credibility and accuracy if wrong at the wells!

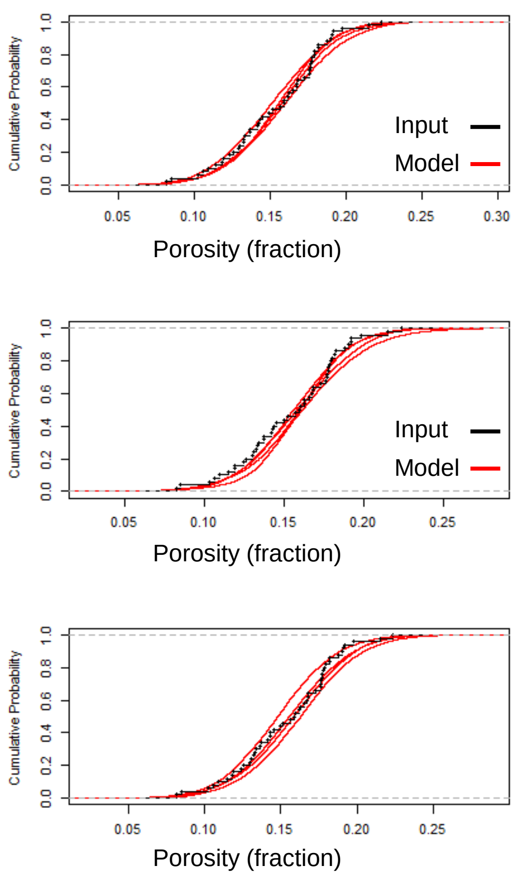


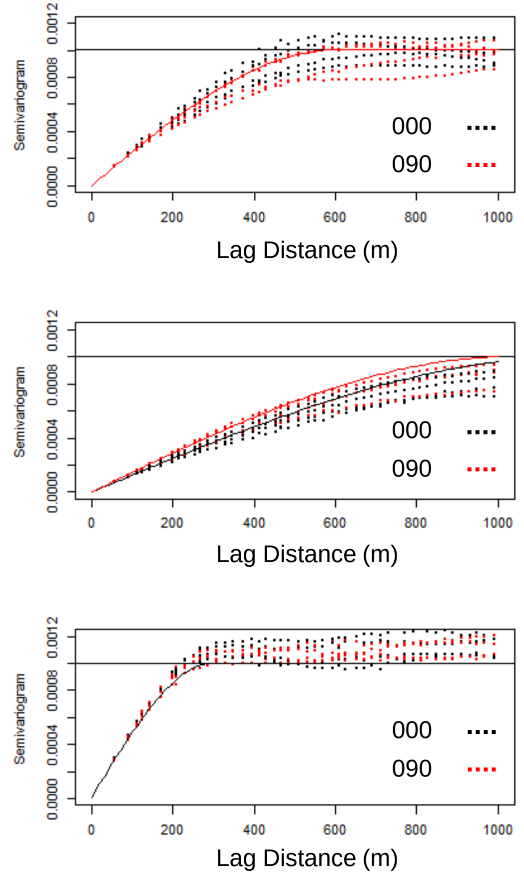
Spatial data and collocated model cells.

  - Many geostatistical modeling methods enforce data reproduction, paint the sample data values on the collocated grid cells.
  - Note, scale up to model cells [that we have not covered] will may result in mismatch, best practice is to compare the scaled-up well data to model cell

Checking Model Input (Geo)Statistics

- Comparison of Input Statistics and Model Statistics
  - It is straightforward to compare the model statistics vs. inputs statistics





Input and model cumulative distribution function Input and model empirical variograms (output is dashed)

  - Some level of variation is expected, ergodic fluctuations, but should be unbiased

PGE 337 Data Analytics and Geostatistics

Lecture 16b: Model Checking

Short Summary of:

- **Cross Validation of Estimates**

Introduction

General Concepts

Univariate

Bivariate

Spatial

- Calculation
- Variogram Modeling
- Kriging
- Simulation

Time Series

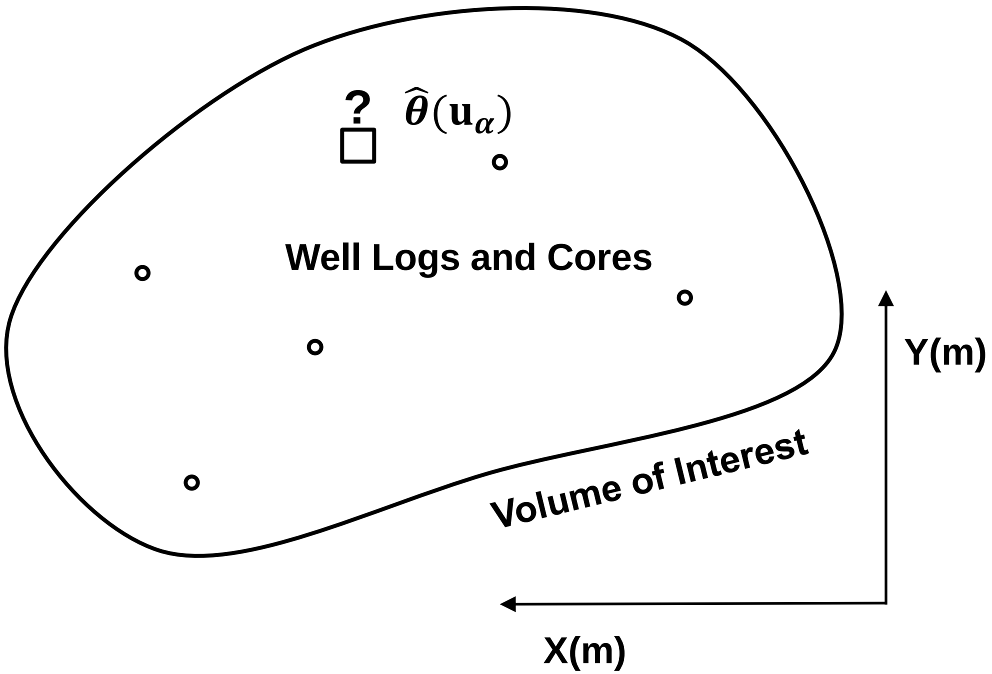
Machine Learning

Uncertainty Analysis

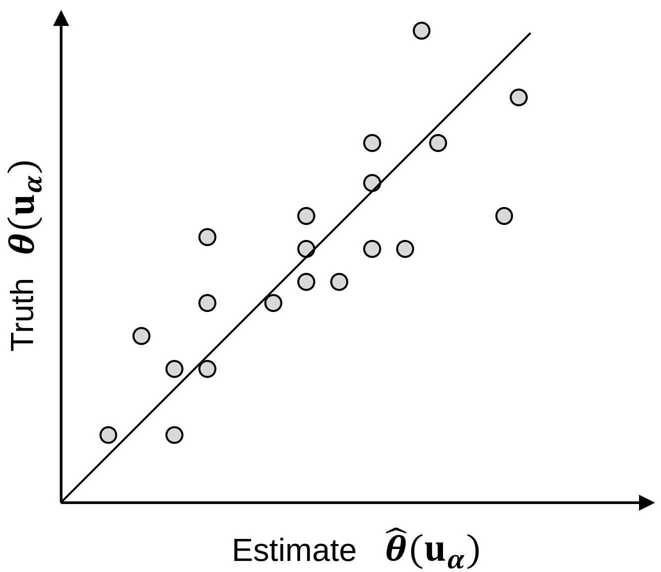
Michael Pyrcz, The University of Texas at Austin

Checking Local Accuracy

- Check the ability of the model to estimate away from data
  - We need to assess the accuracy for estimates away from wells, sample data



Well data over the volume of interest and an estimate at an unsampled location.

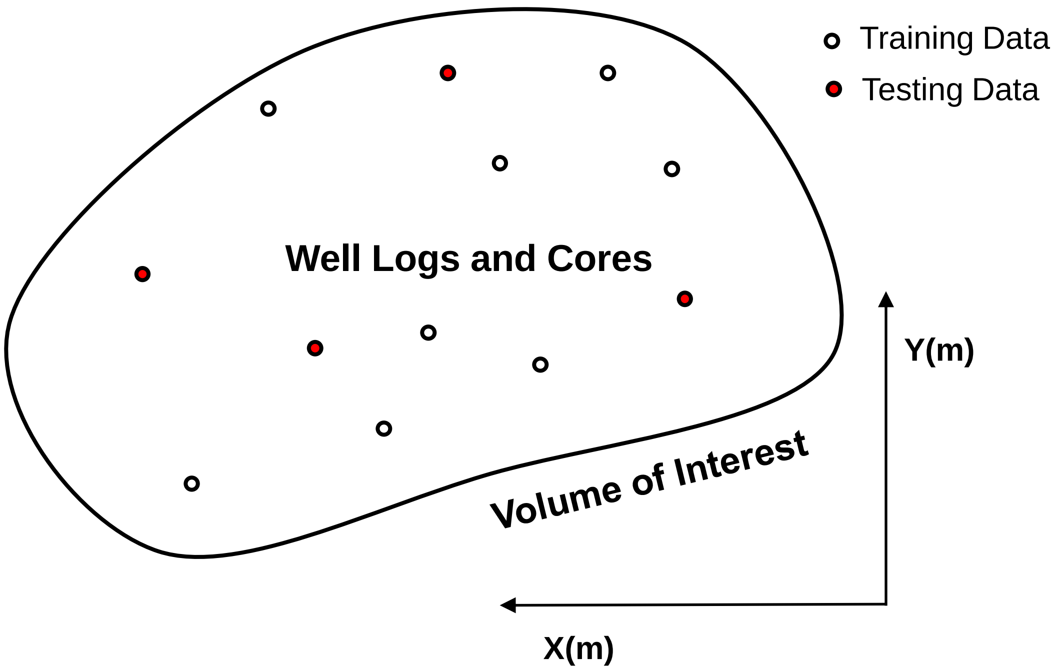


Withheld testing data vs. estimates.

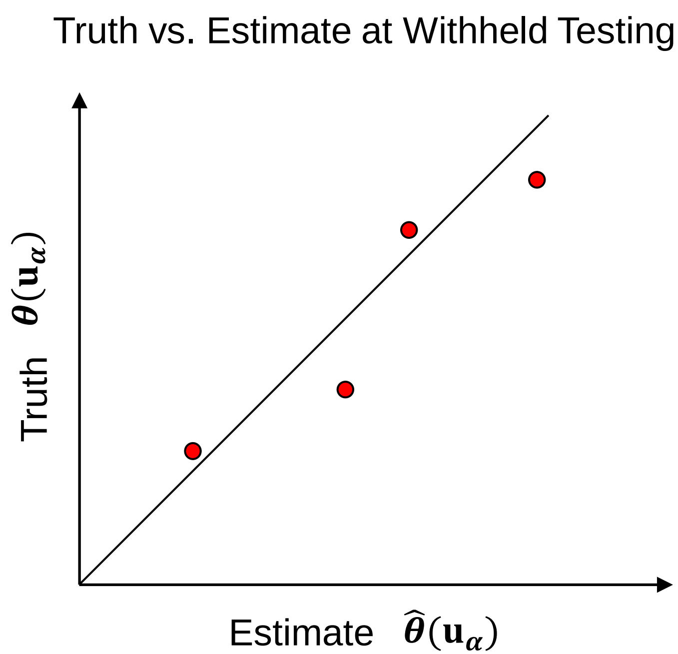
  - This is critical to our assessment of resource in place, and development decisions such as well locations and enhanced recovery methods
  - But we don’t have data away from the data! Cross validation methods.

Cross Validation

- Cross Validation Method
  - Split the data into train and test (15-30%) subsets, mutually exclusive, exhaustive groups



Training and testing data for Jackknife cross validation

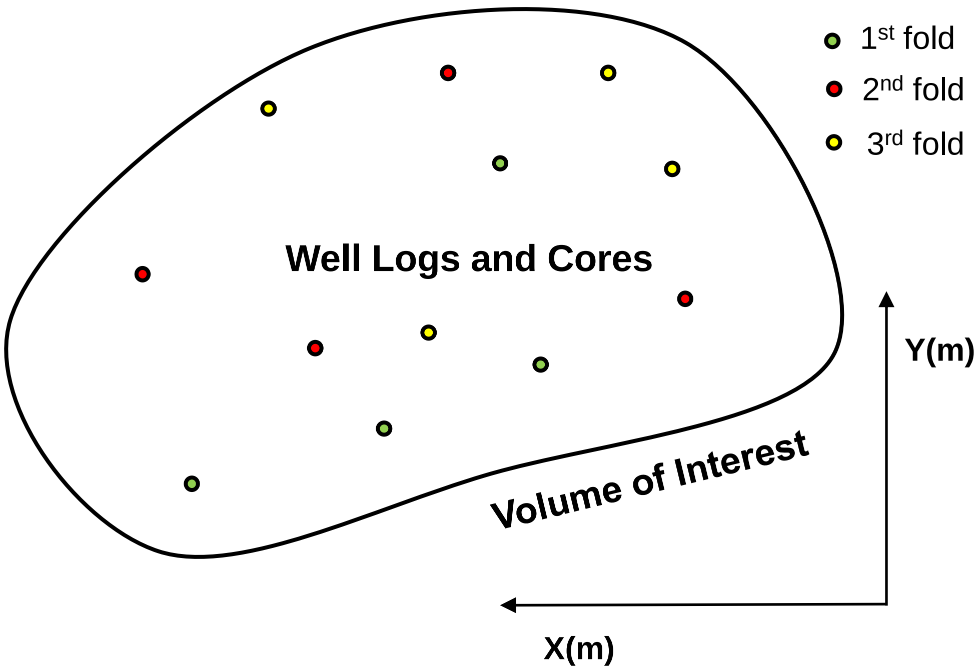


Estimated and truth values at testing data

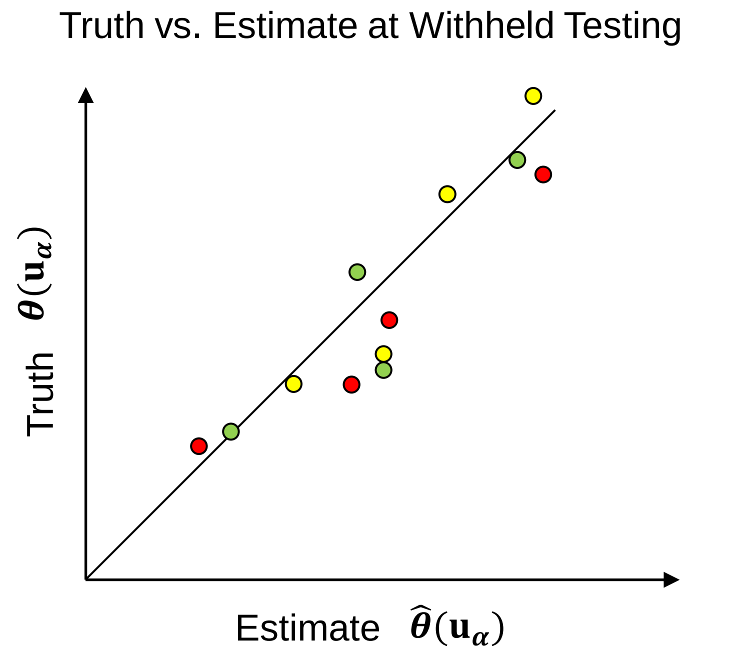
  - To be a fair test, the test data cannot inform any part of the model, e.g. variogram, distribution and trends.
  - The difficulty of the estimates should be similar to the planned use of the model

Cross Validation

- Cross Validation Method: k-fold Cross Validation
  - Like Jackknife, but we repeat over multiple folds, withheld subsets to test all data
  - We get to test at all data, and an error score for each fold, that we average



Three data folds for k=3 fold Cross Validation.

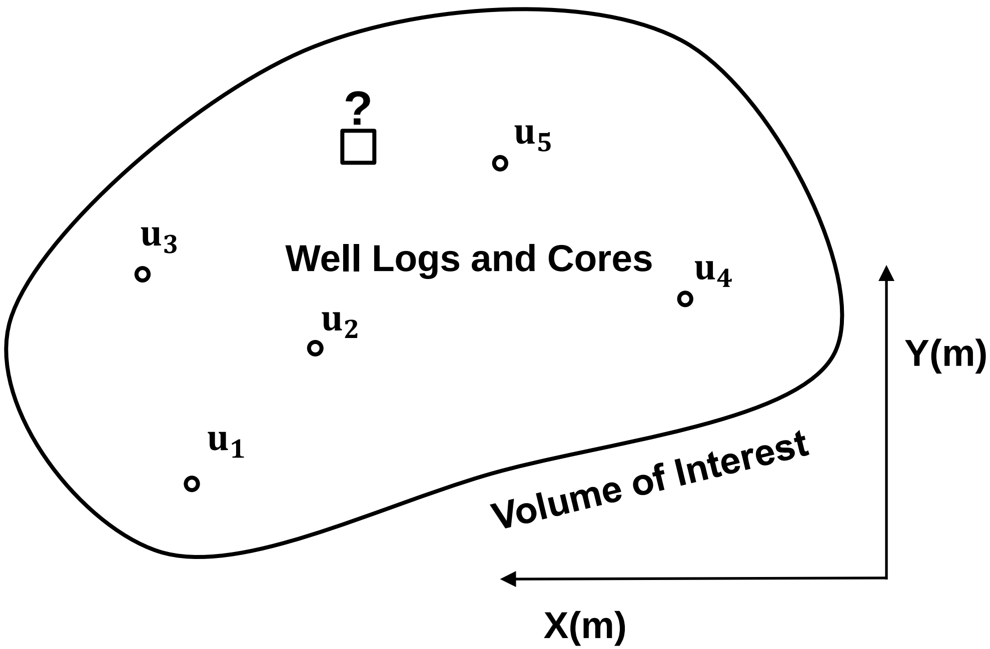


Estimated and truth values at testing data over 3 folds.

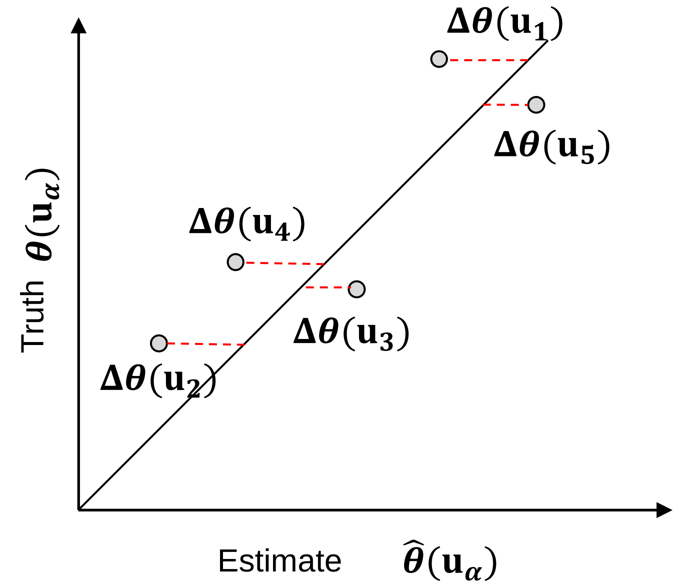
  - To be a fair test, the test data cannot inform any part of the model, e.g. variogram, distribution and trends.
  - This requires k models to be calculated, i.e. 1^{st} fold as test, …, k^{th} fold as test.

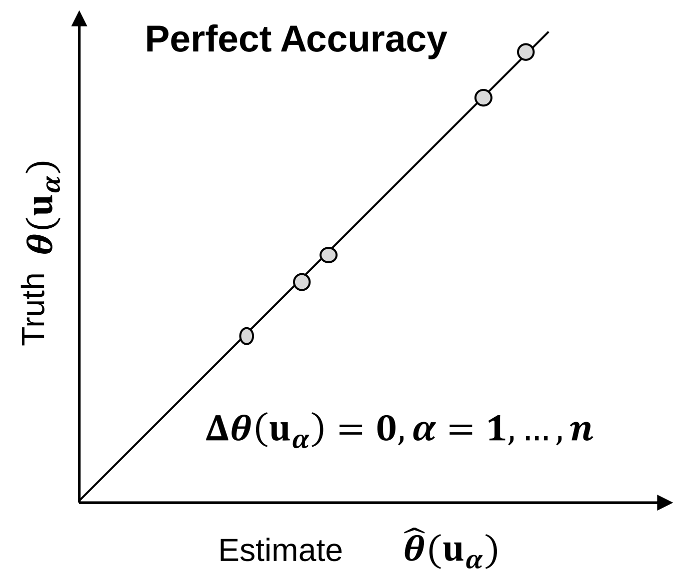
Summarizing Accuracy

- We will need a measure to summarize the accuracy
  - We need to go beyond a scatter plot.



Well data over the volume of interest.



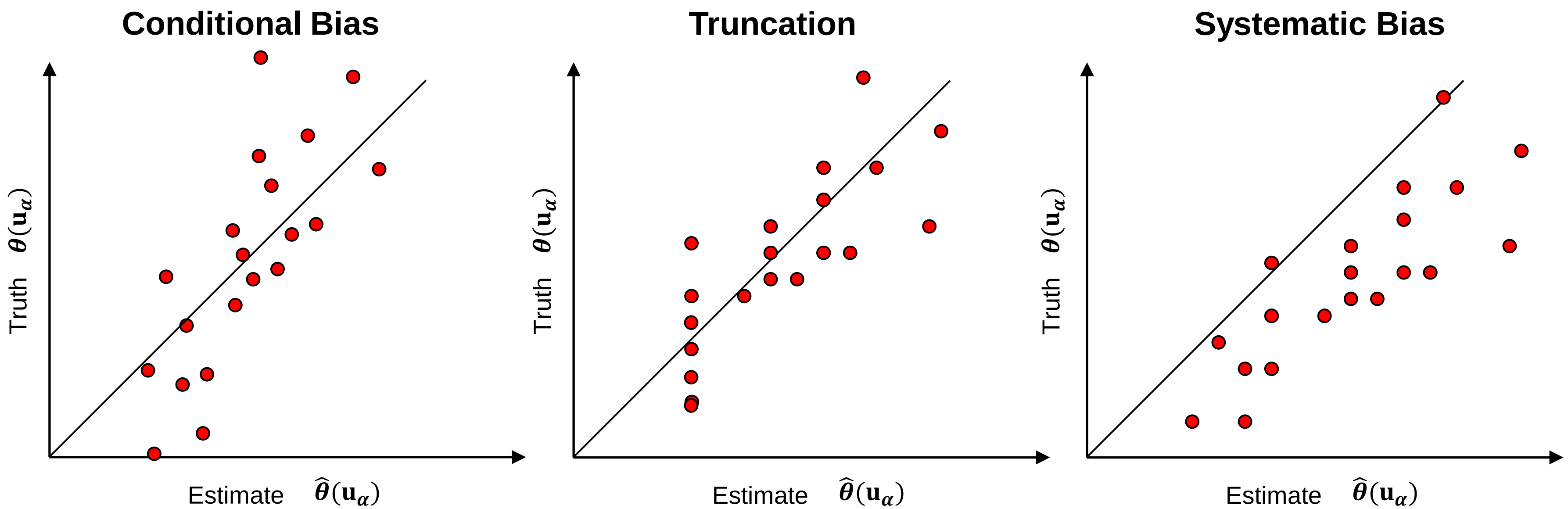


  - A common measure is the Mean Square Error:

$$MSE = \frac{1}{n}\sum_{\alpha=1}^{n} \Delta\theta(u_\alpha)^2$$

Checking Local Accuracy

- Some interpretations of cross validation plots
  - Here are some examples of poor results from cross validation with interpretations.



Three examples of poor cross validation results with interpretation.

  - **Conditional Bias** – systematic overestimation of lows and underestimation of highs
  - **Truncation** – the range of estimates is artificially truncated
  - **Systematic Bias** – mean of estimates is too low or too high over the entire model

PGE 337 Data Analytics and Geostatistics

Lecture 16b: Model Checking

Short Summary of:

- Cross Validation of Uncertainty Models

Introduction

General Concepts

Univariate

Bivariate

Spatial

Calculation

Variogram Modeling

Kriging

Simulation

Time Series

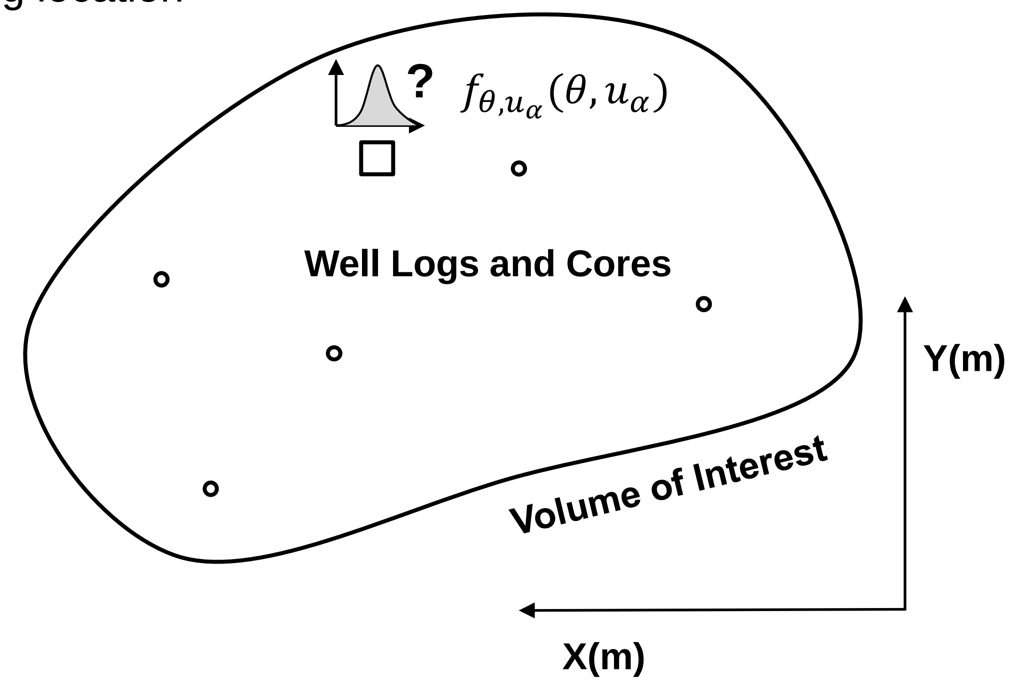
Machine Learning

Uncertainty Analysis

Michael Pyrcz, The University of Texas at Austin

Checking Local Uncertainty

- Our subsurface models provide the entire uncertainty distribution
  - We need to check the entire distribution, not just a single estimate at each testing location



Well data over the volume of interest and an uncertainty model at an unsampled location.

  - We need to determine if our uncertainty model performs well, fair uncertainty
  - We use a modified form of cross validation (Deutsch, 1996, Pyrcz and Deutsch, 2014)

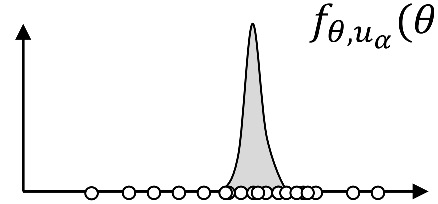
Checking Local Uncertainty

- What can go wrong with our uncertainty model?
  - We need to check the entire distribution, not just a single estimate at each testing location

Too Low Uncertainty

– Too many truth values outside our confidence intervals

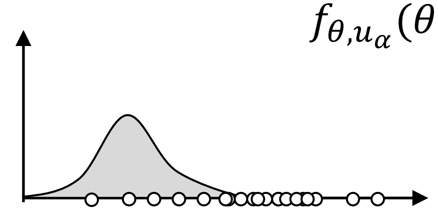
$f_{\theta,u_\alpha}(\theta, u_\alpha), \;\; \alpha = 1, \ldots, n$



Biased Estimates

– Too many truth values outside side our confidence intervals

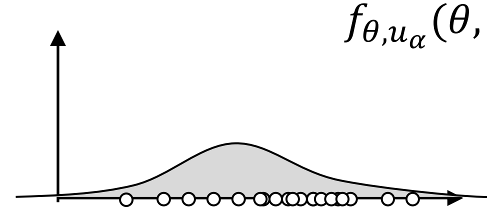
$f_{\theta,u_\alpha}(\theta, u_\alpha), \;\; \alpha = 1, \ldots, n$



Too High Uncertainty

– Too many truth values inside our confidence intervals

$f_{\theta,u_\alpha}(\theta, u_\alpha), \;\; \alpha = 1, \ldots, n$



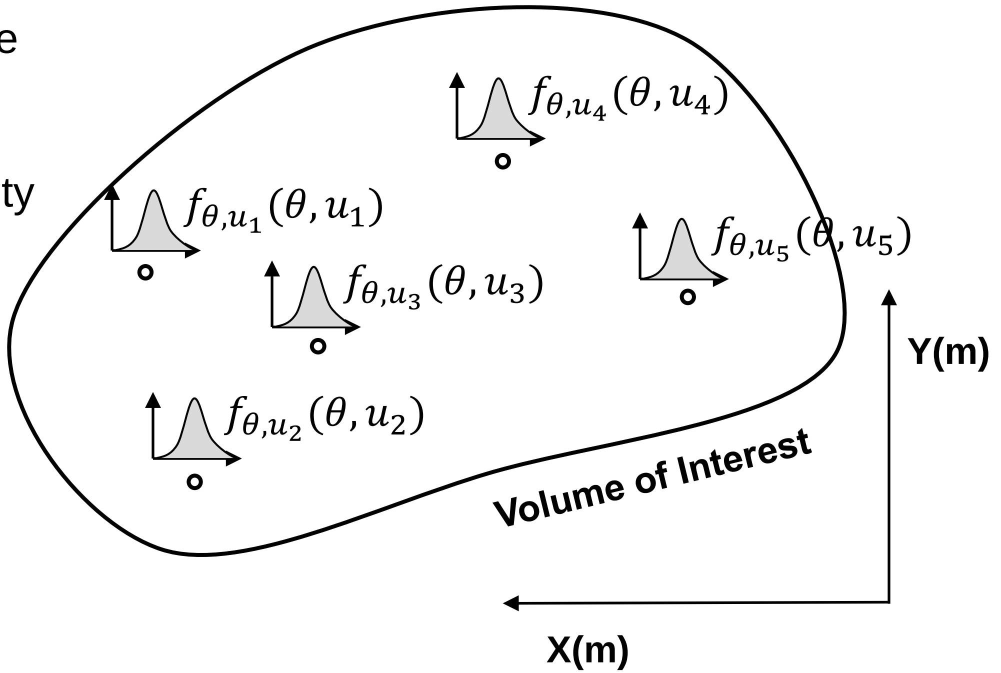
Uncertainty distribution vs. data examples.

We are comparing many locations for which out model would give the same distribution

Checking Local Uncertainty

- The accuracy plot method to cross validate uncertainty
  - The is the workflows to calculate an ‘accuracy plot’

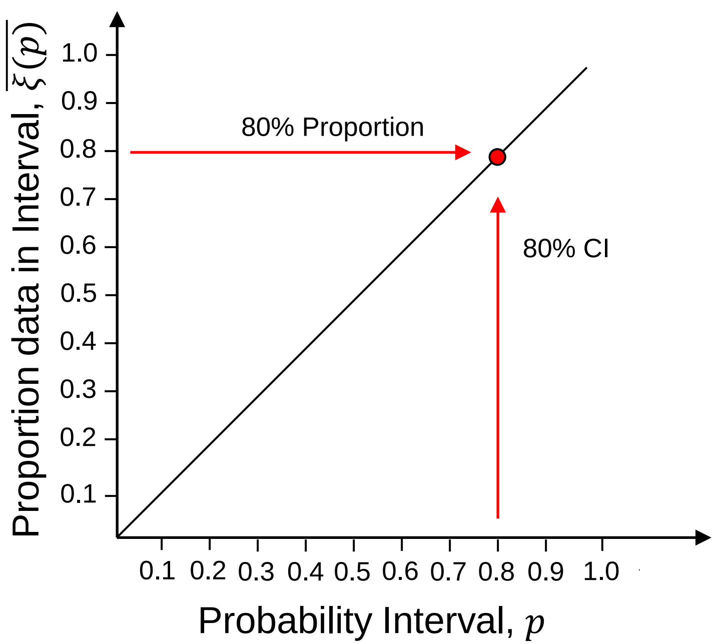
1. Withhold testing data and estimate the uncertainty distributions at the testing data locations.
2. Calculate the cumulative probability of the withheld testing data.
3. For a set of symmetric probability intervals calculate the proportion of testing data in the interval.
4. Plot the proportion of data in the interval vs. the probability interval



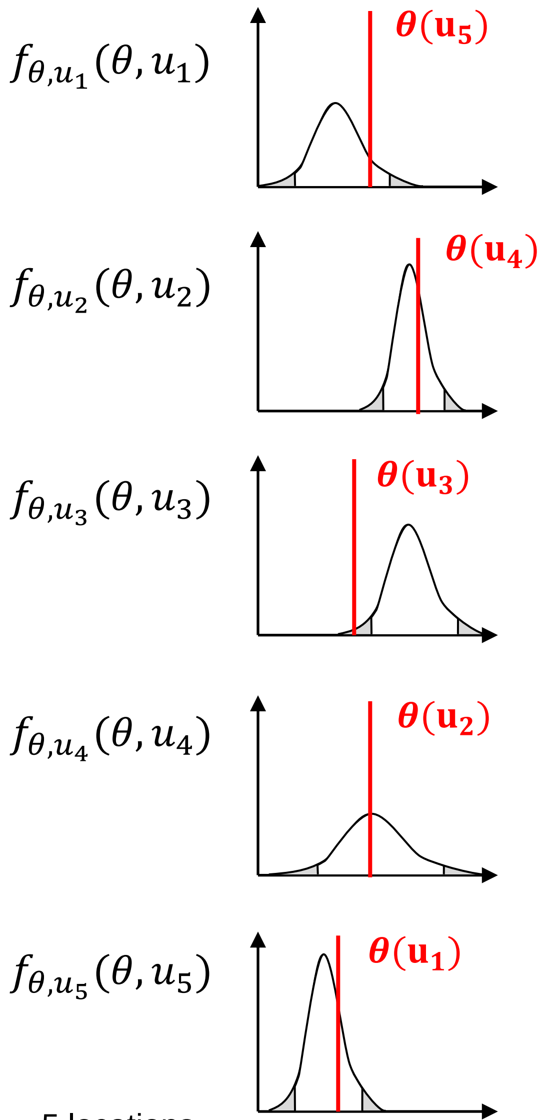
Testing data locations and estimated uncertainty distributions.

Checking Local Uncertainty Example

- We have $n_{test} = 5$ and CI = 80%
  - We plot the withheld data values on the uncertainty distributions, calculate cumulative p-values
  - In 4 of the 5 locations the true value with within the 80% symmetric confidence interval of the data



Accuracy plot with a single point plotted for 80% CI.

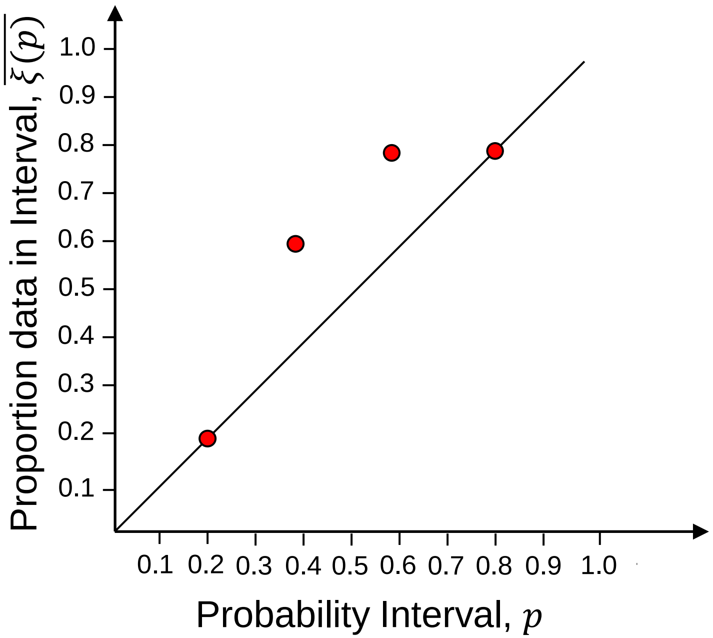


5 locations, uncertainty distributions, 80% CI and true values.

- We will have a lot
- more testing data
- for improved resolution

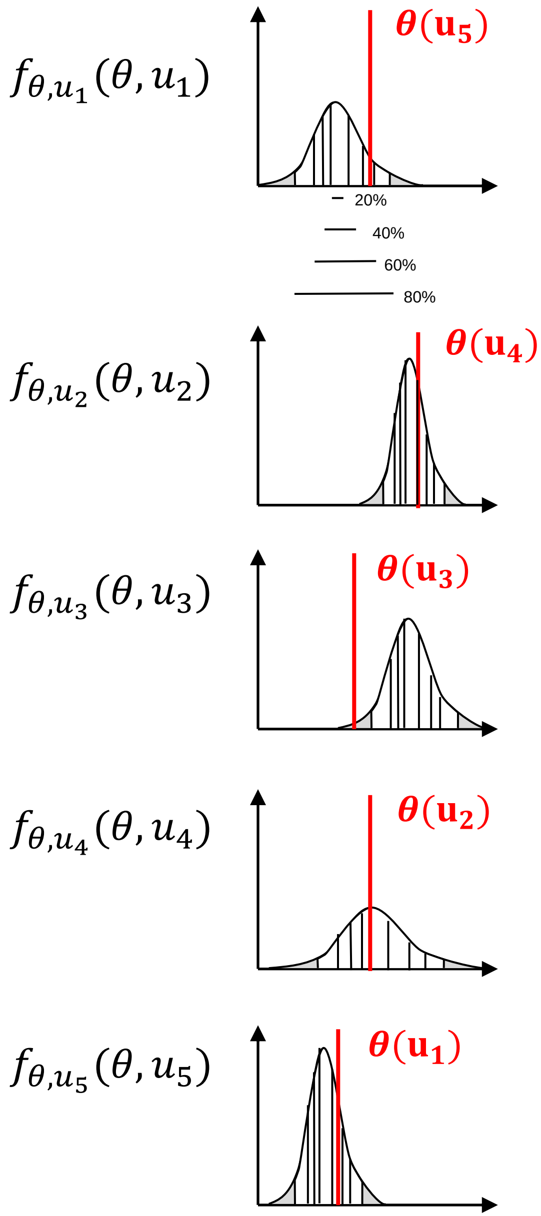
Checking Local Uncertainty Example

- We have $n_{test} = 5$ and CI = 80%
  - No we draw the 20%, 40% and 60% probability CI's
  - We can add the proportion of true data within vs. the probability interval



Accuracy plot with a 20%, 40%,…, 80% CI's.

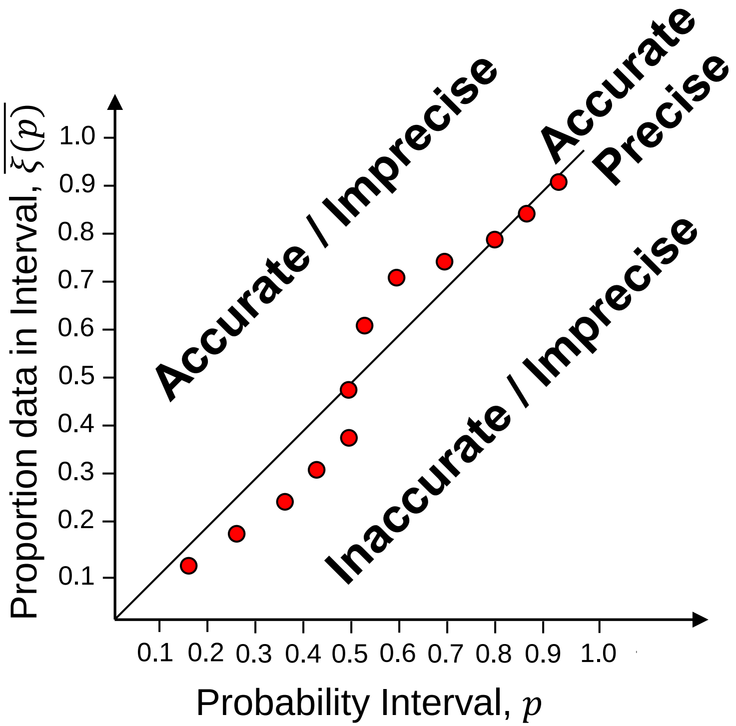
  - We have too many true data in the 40% and 60% probability intervals



- We will have a lot more testing data for more resolution

Checking Local Uncertainty Example

- Now we can add an interpretation to our plot
  - Above the 45-degree line, **accurate, but imprecise**, uncertainty too wide
  - On the 45-degree line, **accurate and precise**
  - Below the 45-degree line, **inaccurate and imprecise**, uncertainty too narrow or biased



Checking Local Uncertainty Example

- The Goodness Measure (Deutsch, 1996).

$$\boldsymbol{G} = \mathbf{1} - \int_0^1 [3a(p) - 2]\left[\overline{\xi(p)} - p\right] dp$$

Where $a(p)$ is an indicator transform

$a(p) \; = \; 1$, if $\overline{\xi(p)} \geq p$

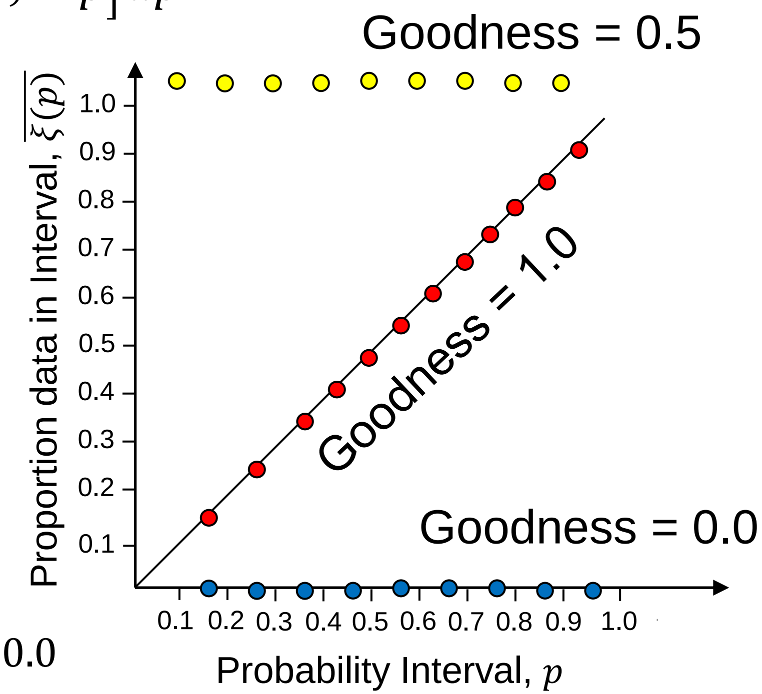
$a(p) \; = \; 0$, if $\overline{\xi(p)} < p$

Completely imprecise and inaccurate:

$$G = 1 - [-2] \int_0^1 \left[\overline{\xi(p)} - p\right] dp \; = 1 - [-2][-0.5] = 0.0$$

Completely imprecise but all accurate:

$$G = 1 - [1] \int_0^1 \left[\overline{\xi(p)} - p\right] dp \; = 1 - [1]0.5 = 0.5$$



PGE 337 Data Analytics and Geostatistics

Lecture 16b: Model Checking

Short Summary of:

- Model Checking
- Checking Reproduction of Model Inputs
- Cross Validation of Estimates
- Cross Validation of Uncertainty Models

Michael Pyrcz, The University of Texas at Austin