PGE 337 Lecture 1: Statistics

Introduction

General Concepts

Statistics

Probability

Univariate

Bivariate

Spatial Analysis

Machine Learning

Lecture outline . . .

- **Statistical Methods**
- **Data Types**
- **Sampling Methods**

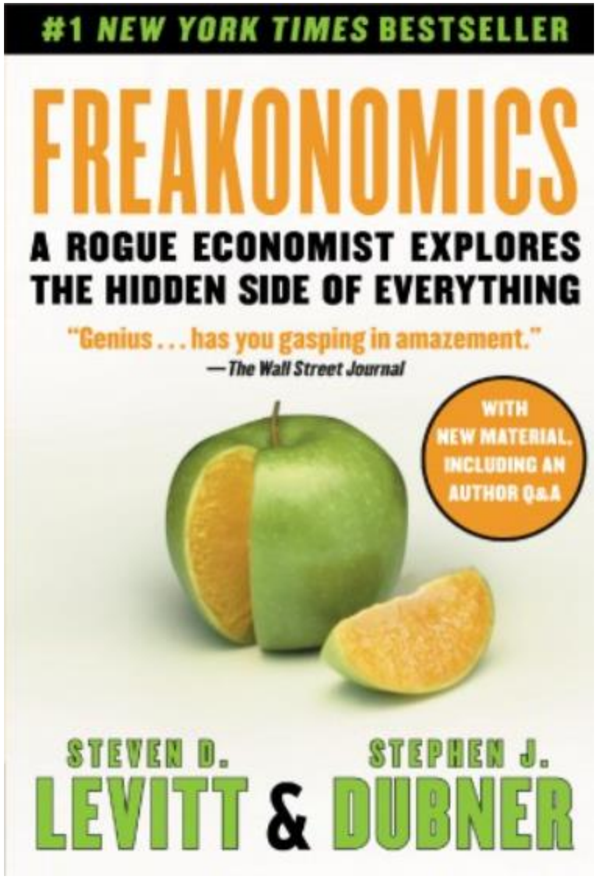
Statistics Minute

Freakonomics

Levitt and Dubner

On the unintended consequences of incentives:

- Sumo wrestlers cheat
  - Wrestlers that met their required number of wins threw matches
- Teachers in Atlanta were cheating
  - Correcting answers on students' benchmark tests
- Remember this when you:
  - Motivate your asset team, unit or division
  - Work with kids
  - Assess your own actions professionally for bias



Statistics
What should you learn from this lecture?

- **Fundamentals of Statistics**
  - Background on Statistics and its Importance
  - Data in Earth Sciences
  - Sampling Biases and Concepts for Mitigation

(Geo)statistics

Statistics *is the science of* collecting, pooling samples and making inferences.

Geostatistics is a branch of statistics with a focus on:

- Geologic context
- Spatial context with spatial correlation
- Account for scales / size / accuracy of the measurements

Statistics
The Method

Steps to answer a question about the subsurface?

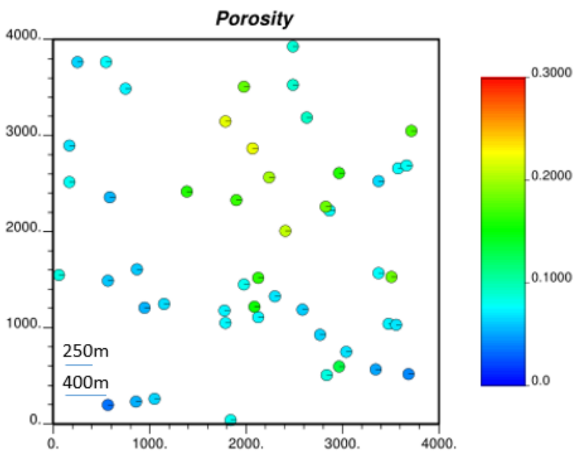
1. **Design:** Sample required to answer the questions of interest
2. **Description:** Summarizing and analyzing the obtained sample data
3. **Modeling**: Use physics, interpretation, proxies, geostatistical descriptions and modeling decisions to build geostatistical models
4. **Inference:** learning about the relationships between the various variables (*multivariate*) that are sampled and over locations (*spatial*)
5. **Prediction:** forecasting at unsampled locations variables of interest
6. **Uncertainty:** developing models of uncertainty for the variables of interest
7. **Decision Making:** optimum decisions in the presence of uncertainty

These steps only add value when it impacts a decision!

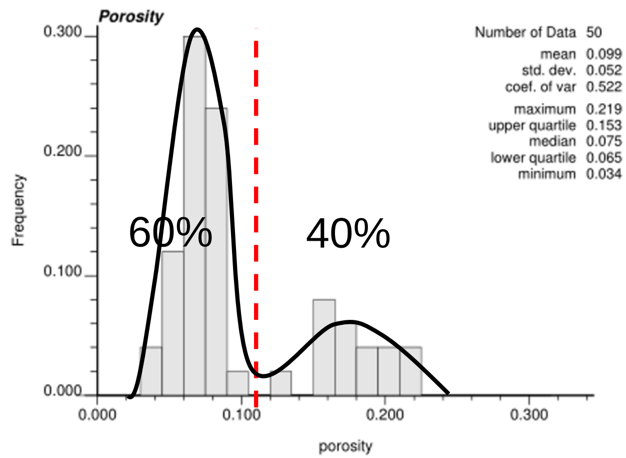
- For example: how many wells and where? what injection rate? for natural resources like water, oil and gas, environmental remediation etc.

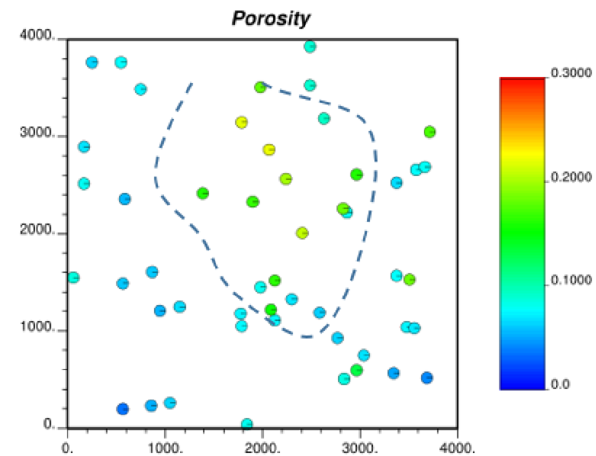
Statistics
The Method

What do we need to be able to answer a question about the subsurface?



Rock Type 1 **Rock Type 2**





Sampling Design

Test hypothesis
Control variables
Pooling available porosity samples

Description / Analysis / Modeling

e.g. Determination of multimodal distribution with natural break in porosity distribution.

Inference

Mapping of 2 distinct reservoir facies to model separately.

Statistics Moments
Share the Impact of Statistics

One stories / lecture on the role of statistics in society, natural resource industry, other sectors.

- If there is no volunteer, we will have a "random" selection! **Be Prepared!**
- If you like, you can e-mail a single slide to support your study and I'll include in the lecture slide. Send the day before the lecture.
- Two minutes maximum.

Statistics Moments
My Favourite From Last Semester

Uddhav shared on Survivorship Bias and I tweeted this with linkage to subsurface.

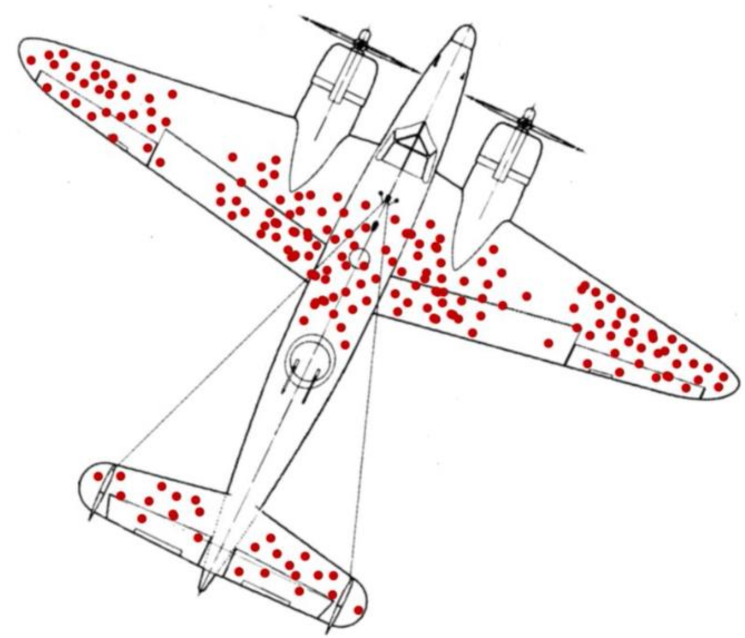
More on Bias: Survivorship Bias in Subsurface Modeling?

Michael Pyrcz, University of Texas at Austin (@GeostatsGuy)

Example shared in my Introduction to Geostatistics class by @uddhav_marwaha (Twitter).

Survivorship Bias: a form of selection bias resulting from selecting samples that "survived" some previous selection process. This often leads to false conclusions. For example, in WWII the Center for Naval Analyses (@CNA_org Twitter) compiled a dataset of bomber damage to assess where reinforcement was needed. Statistician Abraham Wald recognized this was a case of survivorship bias. The plans shot in critical locations did not return to base. Wald suggested reinforcement of locations that were not damaged in planes that safely returned to base! (https://en.wikipedia.org/wiki/Survivorship_bias#In_the_military)

Is there preselection in our subsurface datasets? For our subsurface projects do we only sample: success cases, producing wells, drill holes with economic ore grades, large fields, clastic depositional settings, marine seismic surveys, high resolution 3D seismic surveys, shallow reservoirs etc. When we pool samples, check for preselection and ensure this is considered in the resulting inferences and decision to export these results. The samples must be representative of the population to which we will apply our model. Of course, this applies to any datasets.



Hypothetical dataset of aircraft damage for planes that returned to based. Source https://en.wikipedia.org/wiki/Survivorship_bias#/media/File:Survivorship-bias.png

Safe-Stats
Why We Use R / Python?

Hadley Wickham, Chief Scientist at RStudio, known for development of open-source statistical packages for R to make statistics accessible and fun (http://hadley.nz/).

Read Hadley Wickham's, **Teaching Safe-Stats, Not Statistical Abstinence** (https://nhorton.people.amherst.edu/mererenovation/17_Wickham.PDF)

- **Teaching:** We need to rethink statistics curriculum – we risk becoming irrelevant!
- **Practice:** Stats tends to be taught as avoid, unless you are an "statistician" or with one
  - Otherwise you will cause great harm
  - But there are not enough professional statisticians
  - Rather than stigmatize amateur, new tools should be safer to use
- **Tools**: New tools should be easy and fun to use to encourage use
  - Flexible grammars, minimal set of independent components to build workflows
- **Coding:** Go for it! Teaching some programming even in a first course is achievable
- **My Job:** Teach safe methods for using (geo)statistics. So we will use R (e.g. ggplot2, dplyr, tidyr) and Python (e.g. numpy, pandas, statsmodels) packages during this class.
- **For Next Class: install Anaconda 3.6, R and RStudio on your laptops for next class.**

Hadley Wickham photograph from https://en.wikipedia.org/wiki/Hadley_Wickham

Statistics
Sampling Definitions

- **Variable:** any property measured / observed in a study
  - e.g. porosity, permeability, mineral concentrations, saturations, contaminant concentration
  - in data mining / machine learning this is known as a **feature**

- **Population:** Exhaustive, finite list of property of interest over area of interest. Generally the entire population is not accessible.
  - e.g. porosity at each location within a reservoir.

- **Sample:** The set of data that have actually been measured
  - e.g. porosity data from measured by well-logs within a reservoir.

- **Parameters:** summary measure of a population
  - e.g. population mean, population standard deviation, we rarely have access to this.

- **Statistics**: summary measure of a sample
  - e.g. sample mean, sample standard deviation, we use statistics as estimates of the parameters.

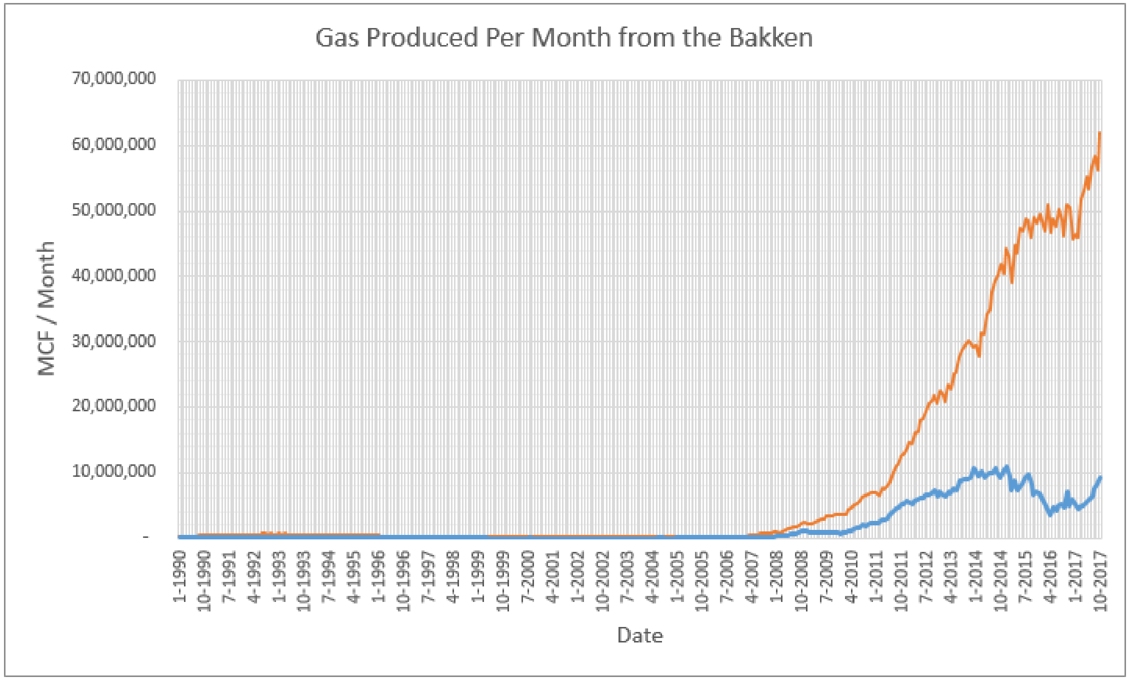
m variables

n samples

Sample ID	Location	Var 1		Var m
1	x_1,y_1,z_1	$v_{1,1}$	…	$v_{1,m}$
2	x_2,y_2,z_2	$v_{2,1}$		$v_{2,m}$
…				…
n	x_n,y_n,z_n	$v_{n,1}$	…	$v_{n,m}$

Data table, part of tidy data from Wickham.

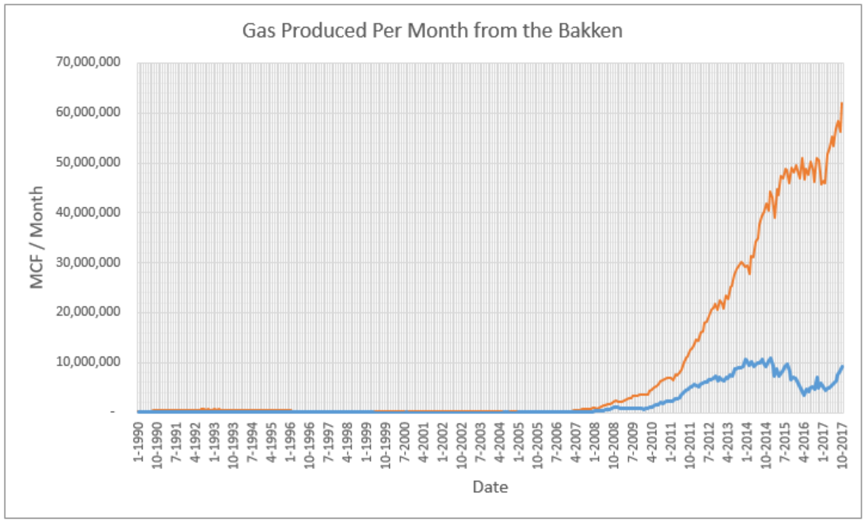
Data Cleaning and Preparation Example

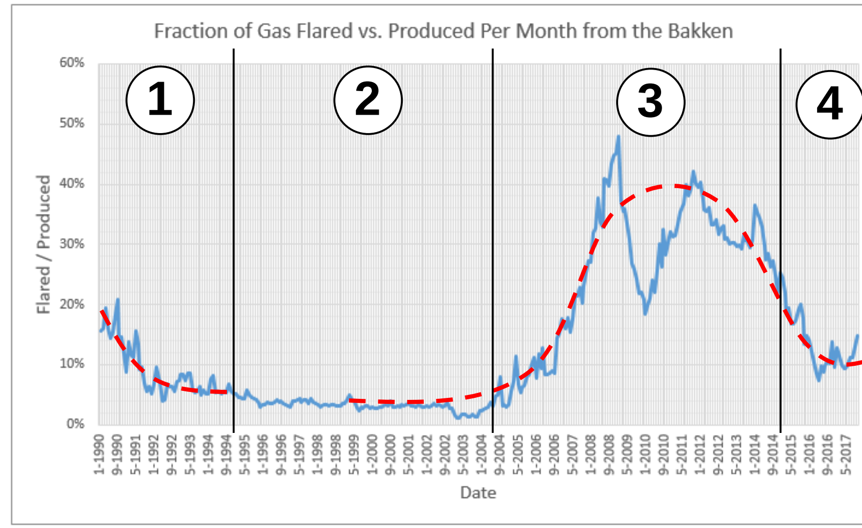


Let's look at the production (orange) and flaring (blue) from the Bakken, North Dakota (https://www.dmr.nd.gov/oilgas/stats/statisticsvw.asp). This is a nice temporal data set.

What types of questions could we ask?

Data Cleaning and Preparation Example



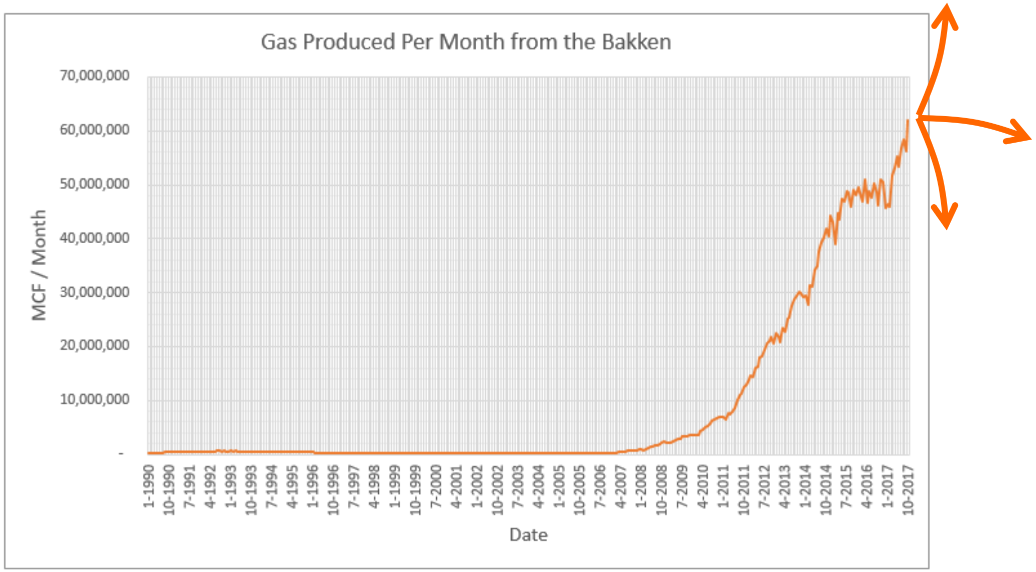


How is the volume flared related to production over time? Calculated the fraction of gas flared vs. gas produced (right).

1. Early utilization
2. Stable low level production
3. Sudden increase in production
4. Infrastructure catches up

Data Cleaning and Preparation is often 80% of project effort.

(Geo)statistics
Some Definitions



Forecasting future production – moving beyond the sample.

Do we have enough data? What else do we need?

- number and locations of new wells?
  - predict production at new locations.
- production profiles of current wells?
- scheduled downtime / reworking

Context and Domain Knowledge are essential!

(Geo)statistics
What Do We Sample?

Analyze the 1D data recorded in a sequence of distance or time

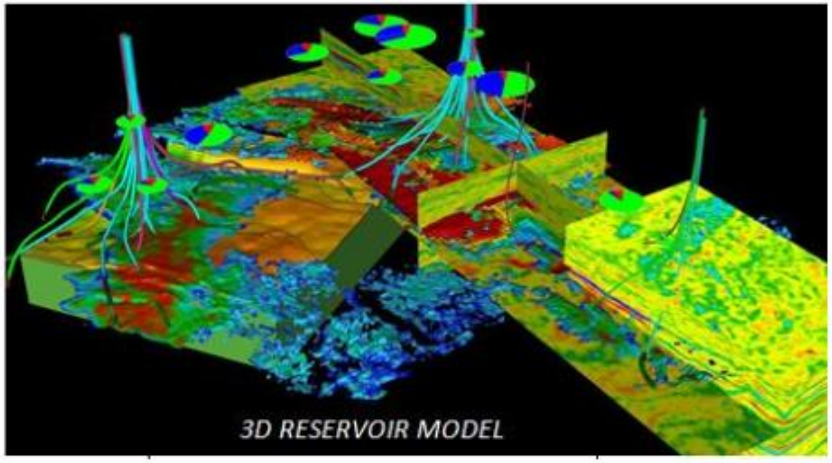
- Variation in vertical distribution of mineralogy or mechanical properties
- Variation in a single well log, gamma ray (shale indicator)

Analyze 2D sampling for spatial interpretation

- Geologic maps, spatial analysis of thin sections, …

Analyze 3D sampling for spatial interpretation

- 3D seismic volumes, sets of correlated well logs



3D reservoir model with various data sources from http://www.oil-gasportal.com/reservoir-management/integrated-reservoir-modeling.

(Geo)statistics
Sampling Types of Measures

Measurement Types:

- **Categorical / Nominal (Classes)**
  - Example: Grains in sandstones can belong to categories including quartz, feldspar, …
- **Categorical / Ordinal:** the ordering of the categories are important
  - Example: Geologic age, hardness
- **Continuous / Interval:** the intervals between numbers are equal
  - Example: Celsius scale of temperature (arbitrary zero)
- **Continuous / Ratio**: numerical value truly indicate the quantity being measured
  - Example: Kelvin scale of temperature, porosity, permeability, saturation

Continuous Data

- Interval scale
- Ratio scale

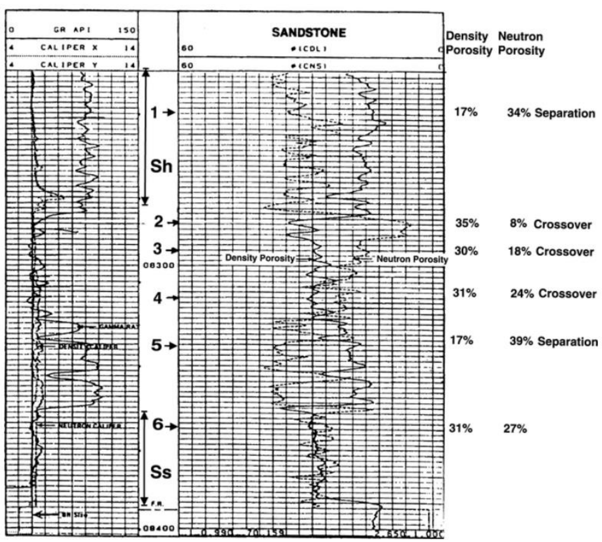
Discrete Data

- Nominal scale
- Ordinal scale

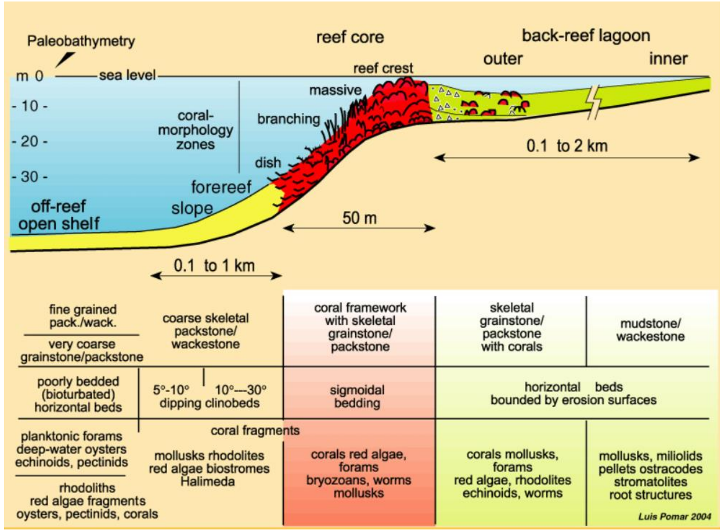
(Geo)statistics
Sampling Types of Measures

Types of Data:

- **Quantitative Data** – information about quantities that can be written in numbers.
  - Example: age, porosity, saturation
- **Qualitative Data** – information about quantities that you cannot directly measure, require interpretation of measurement
  - Example: rock types, facies



Density and neutron log for measuring porosity from a sandstone unit. Alberty, http://wiki.aapg.org/Density-neutron_log_porosity



Interpretation of lagoon carbonate setting from Pomar (2004) taken from http://www.sepmstrata.org/page.aspx?&pageid=54&3

(Geo)statistics
Sampling Types of Measures

Types of Data:

- **Hard Data** – data that has a high degree of certainty. Usually based on a direct measurement.
  - Example: well core- and log-based porosity, lithofacies assessed from well logs.

- **Soft Data** – data that provides indirect measures of the property of interest with a significant degree of uncertainty
  - Example: probability density function for local porosity calibrated from acoustic impedance

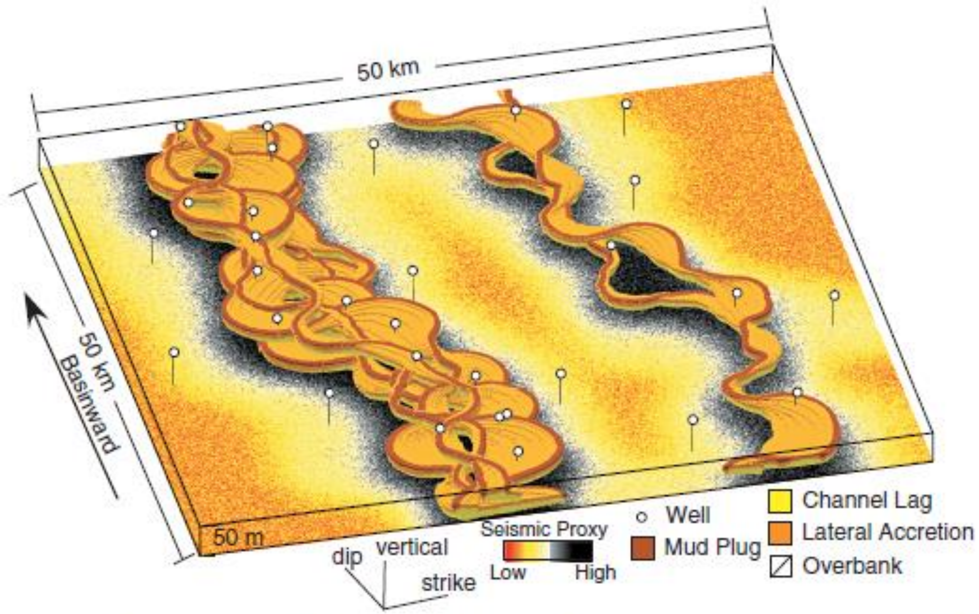


FIGURE 3.43: Oblique View of the Fluvial Reservoir Truth Model with Facies Painted on the Reservoir net Region, Well Locations, and a Seismic Attribute Painted on a 2-D Plane.

From Pyrcz and Deutsch (2014)

(Geo)statistics
Sampling Types of Measures

Types of Data:

- **Primary Data** – the variable of interest. The target for building a model.
  - Example: porosity measures from cores and logs used to build a full 3D porosity model.

- **Secondary Data** – another variable / feature that provides information about the primary data through a relationship / calibration.
  - Example: acoustic impedance to support modeling porosity and porosity to support modeling permeability.

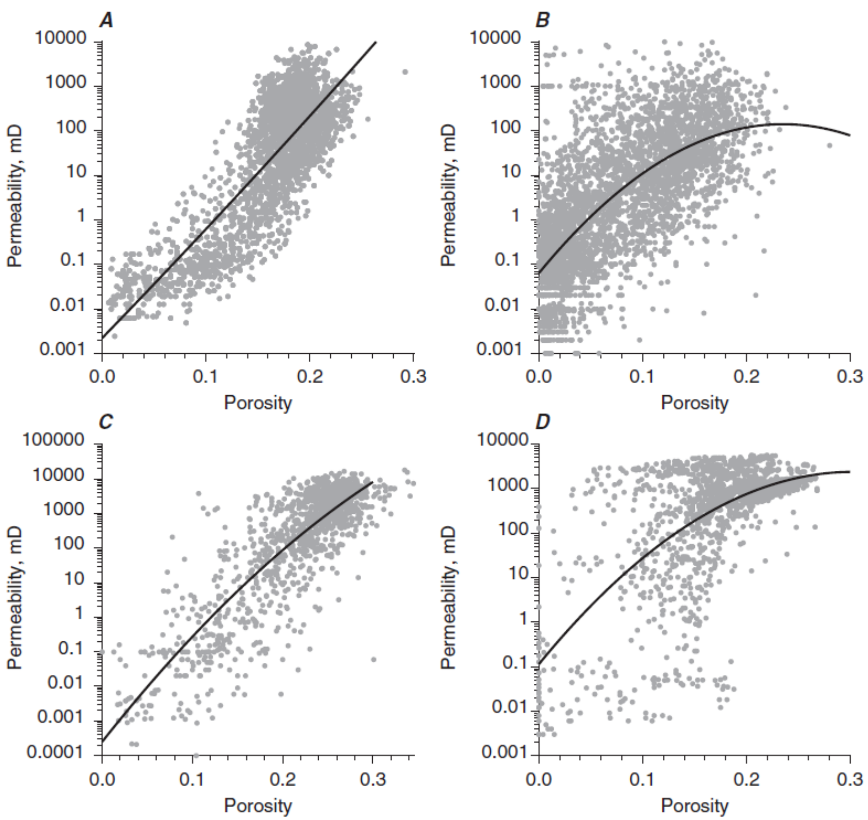


FIGURE 4.77: Four Examples of Porosity/Permeability Cross Plots with Second-Order Regression Curves.

Permeability and porosity relationships (Pyrcz and Deutsch, 2014).

(Geo)statistics
Types of Measures

Types of Measures

- The following discussion is a very cursory general treatment.
- Multiple classes would be required to cover each
- We just explain what they are and summarize their coverage, scale and information type

Coverage

- What proportion of the reservoir has this data available typically?
- e.g. a couple of meters around wells, everywhere etc.

Scale / Support Size

- What is the scale of the individual data measures?
- e.g. pore scale, cm^3 scale, m^3 scale, reservoir unit scale etc.

Information Type

- What does the data tell us about the subsurface?
- e.g. grain size, fluid type, layering etc.

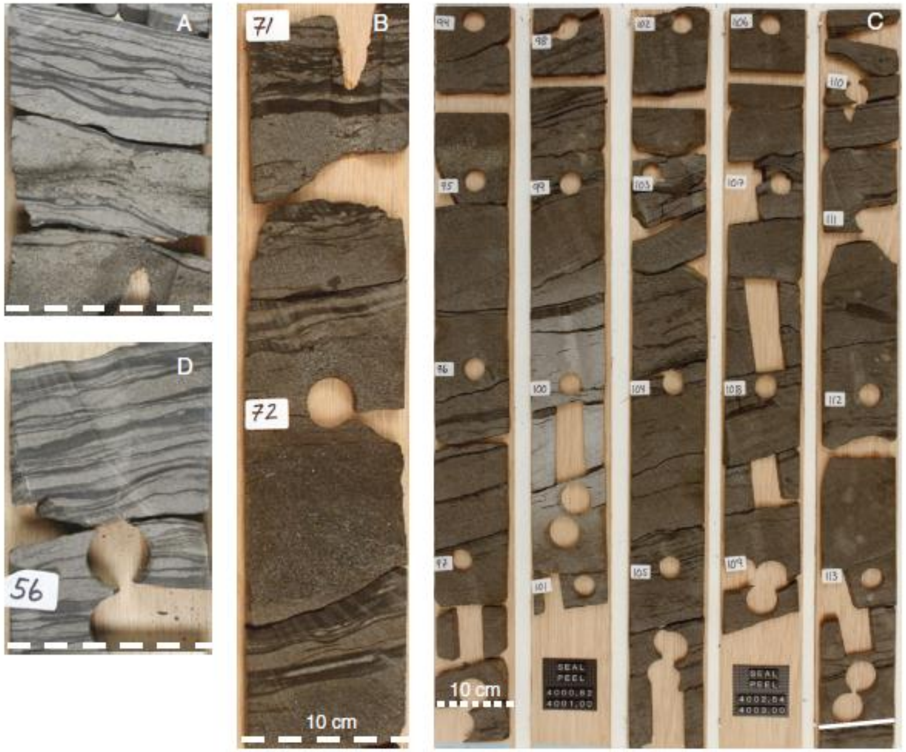
(Geo)statistics Example Measures

Core Data

- Expensive / Time Consuming to Collect
  - infrequent / incomplete coverage of well
  - at select locations

Petrology, Stratigraphy

- Excellent for quantitative measures such as grain size and porosity
- Interpretations are critical to support the entire reservoir concept / framework for prediction
- Integration of facies, porosity important calibration for all well logs



Sectioned Core Photographs of the Cook Formation, a Shallow Marine Sandstone Reservoir from the North Sea. The core data have been interpreted as a fluvial / deltaic depositional setting with general progradation upward Folkestad et al. (2012).

(Geo)statistics
Example Measures

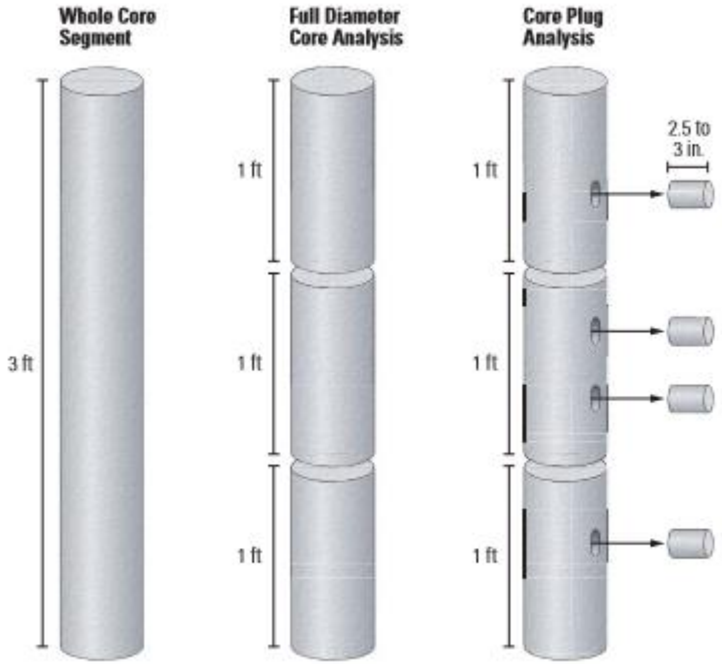
More on Core Data

Routine Core Analysis

- Porosity, permeability and saturation
- Core gamma logging for calibration to well logs
- Core tomography (CT) scans to assess pore structure

Special Core Analysis

- Electrical measurements for calibration of spontaneous potential (SP) and nuclear magnetic resonance (NMR) well logs.
- Mercury injection for pore throat distribution
- Relative permeability for multiphase flow character



^ Divided cores. At the wellsite, whole cores are typically cut into smaller segments for ease of shipping. At the laboratory, the whole core segments may be cut and subsampled.

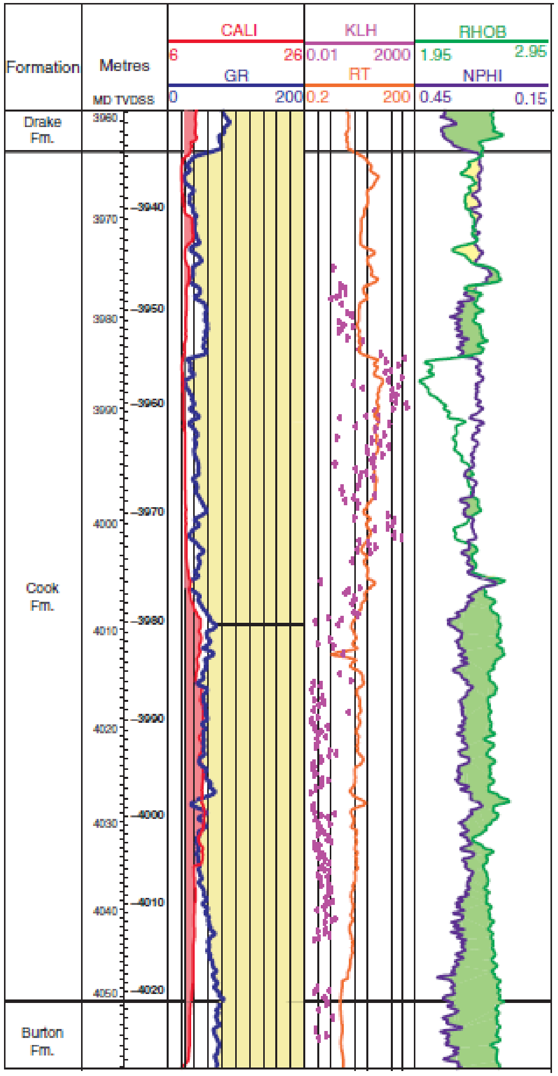
Whole core, full diameter and core plug analysis http://www.slb.com/~/media/Files/resources/oilfield_review/ors13/sum13/02_core_truth.pdf

(Geo)statistics Example Measures

Well Log Data

- Common / Wide Coverage / Suite of Logs
- Examples:
  - Multiple indirect measures of near bore
    - Resistivity and spontaneous (SP)
      - porous / permeability vs. shales
      - bed boundaries
      - fraction of shale
    - Gamma ray
      - Gamma ray counter to detect organic rich shale
    - Nuclear magnetic resonance
      - Use for medical imaging
      - Respond to presence of hydrogen protons
      - Quantity and type of fluids

Online Source: http://petrowiki.org/Types_of_logs



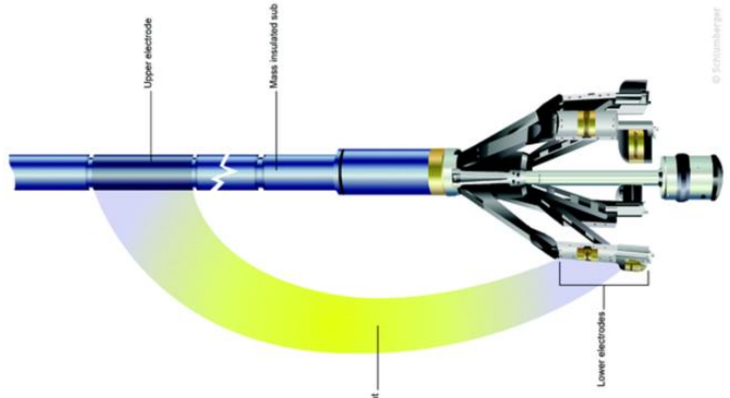
Suite of Well Logs with Interpreted Structures from the Core Data and Stratigraphic Units Form the Cook Formation, a Shallow Marine Sandstone Reservoir from the North Sea. The core data have been interpreted as a fluvial / deltaic depositional setting with general progradation upward Folkestad et al. (2012).

(Geo)statistics Example Measures

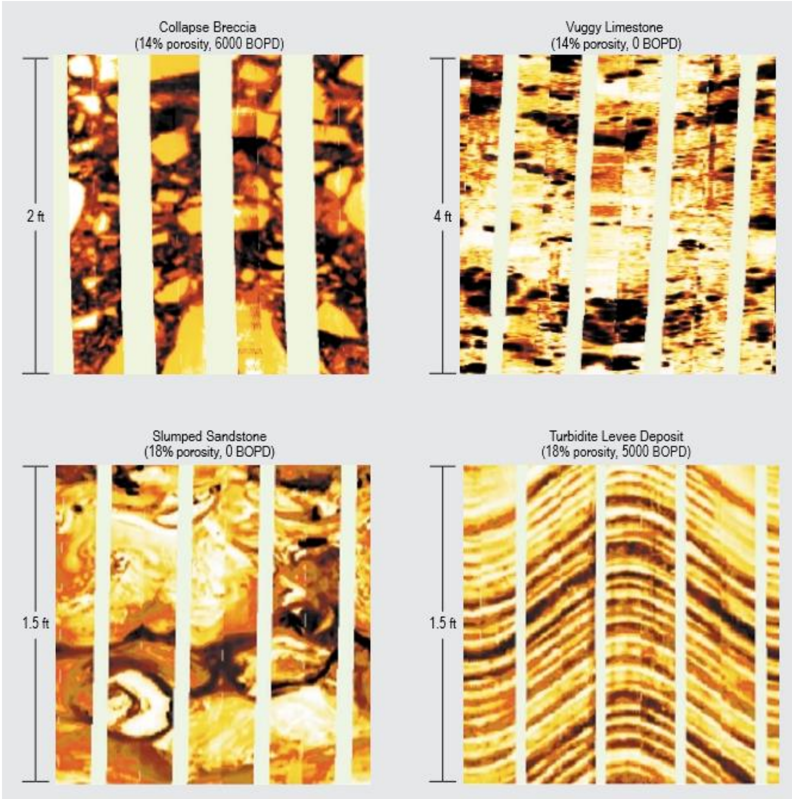
Well Log Data – Image Logs

- Variable coverage over wells
- Centimeter-scale microresistivity images of bore hole walls
- Example:
  - Formation Microimager (FMI)
    - 80% bore hole coverage 0.2”
    - resolution vertical and horizontal
    - 30” depth of investigation
    - Observe lithology change, bed dips and sedimentary structures.

Online Source: http://petrowiki.org/Types_of_logs

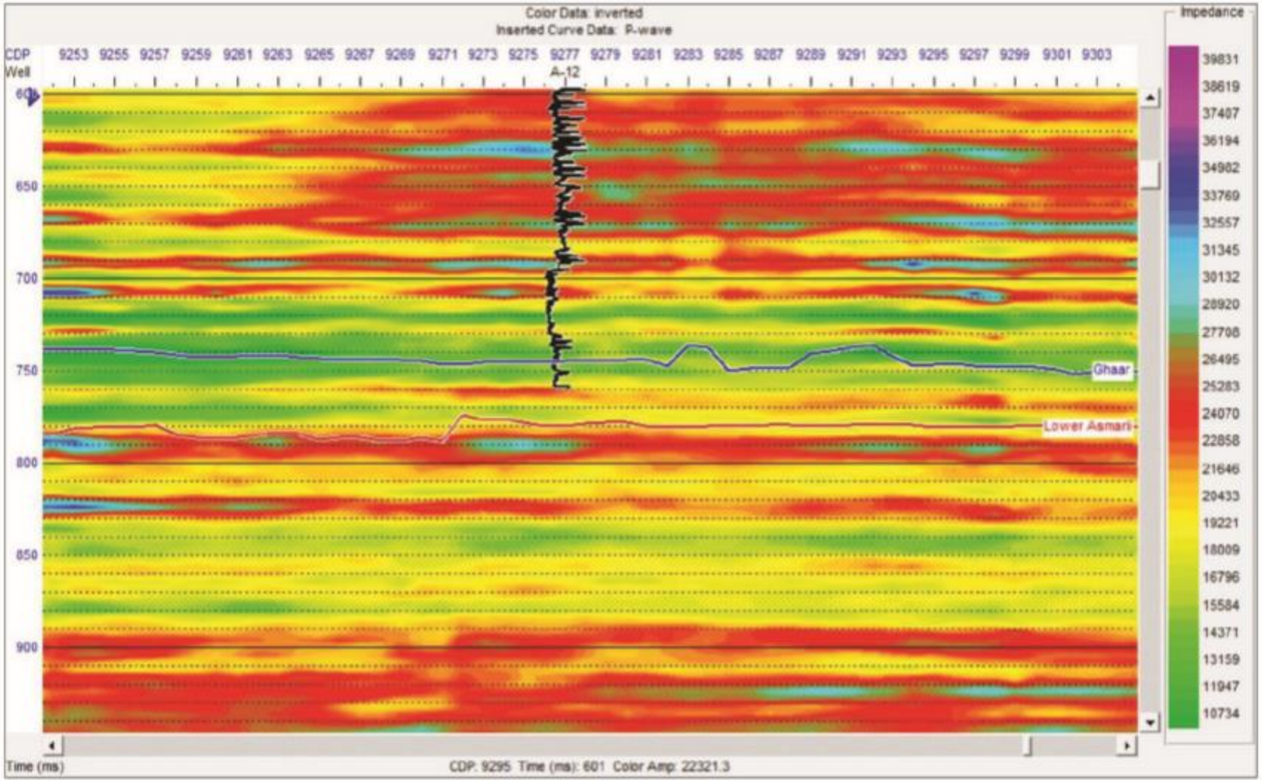


http://petrowiki.org/File:Vol5_Page_0403_Image_0001.png



FMI Image Log examples from: http://www.slb.com/~/media/Files/evaluation/brochures/wireline_open_hole/geology/fmi_br.ashx

(Geo)statistics
Example Measures



Acoustic-impedance section result of model-based inversion on the seismic section in A-1 well location. The black well-log curve is the sonic log. From Jafari et al. (2017) https://library.seg.org/doi/pdf/10.1190/tle36060487.1

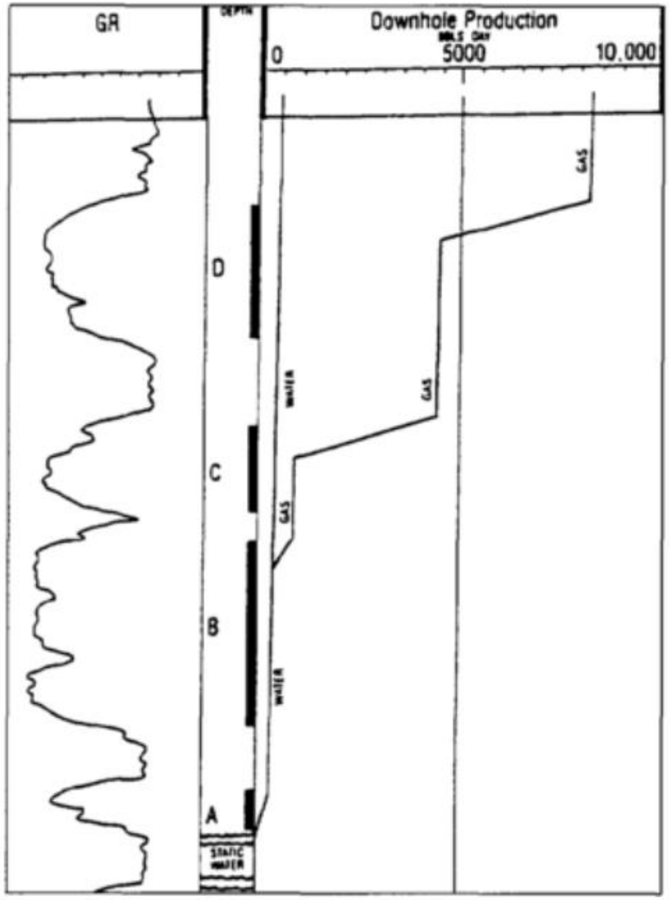
Seismic Data

- Seismic reflections (amplitude) data inverted to rock properties e.g. acoustic impedance.
- Consistent with well sonic logs
- Provides framework, soft information on reservoir properties such as porosity and facies.

(Geo)statistics
Example Measures

Production Data

- Bottom hole pressure, fluid production (rates, types, temperatures etc.)
- Production may be comingled over multiple producing intervals, unless production logging tool (PLT) results are available
- **Most important ground truth** to be matched with a reservoir model.



Production log from a producing logging tool (PLT) of a well from http://wiki.aapg.org/Production_logging.

(Geo)statistics
Summary of Measures

Type	Resolution	Coverage	Information Type
Core	$\simeq \infty$	In Well Bore	Lithology, pore and sedimentary structures
Well Log	10 cm	Near Bore	Facies, porosity, minerology
Image Log	5 mm	Near Bore	Sedimentary structures, faults
Seismic	10 m	Exhaustive	Framework, trends, facies, porosity
Production	10–100 m	Drainage Radius	Volumes, connectivity, permeability
Analog			
Mature Fields	10–100 m	≤ Complete	Validation, prior for all
Outcrop *Geomorphology*	$\simeq \infty$ $\simeq \infty$	none none	Concepts, input statistics Concepts
Shallow Seismic	≥ Element	none	Concepts, input statistics
Experimental Stratigraphy	$\simeq \infty$	none	Concepts
Numerical Process	≥ Complex	none	Concepts

A general summary of data types, resolution, coverage and information type.

(Geo)statistics
Sampling Representatively

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)

(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.

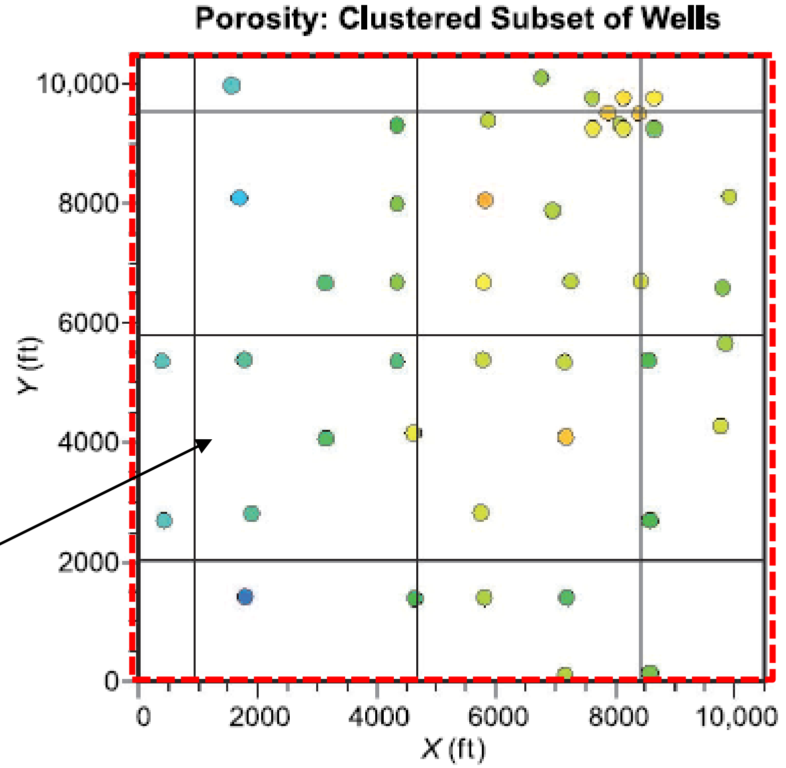
Routine core analysis from https://www.rigzone.com/training/insight.asp?insight_id=325.

(Geo)statistics
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?



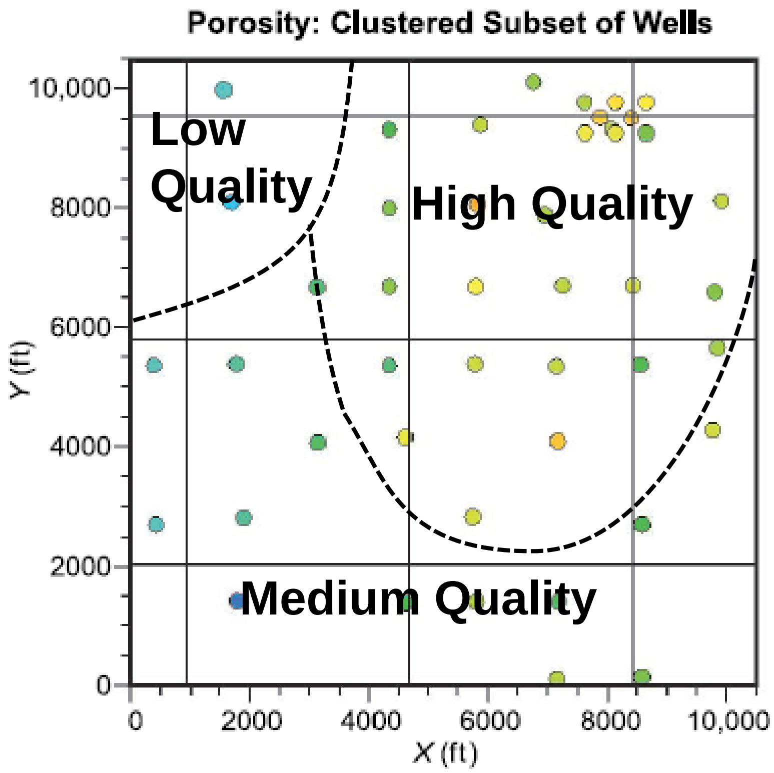
Porosity sample data for an example reservoir (Pyrcz and Deutsch, 2014).

(Geo)statistics Sampling Bias

Addressing Bias:

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

1. 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).

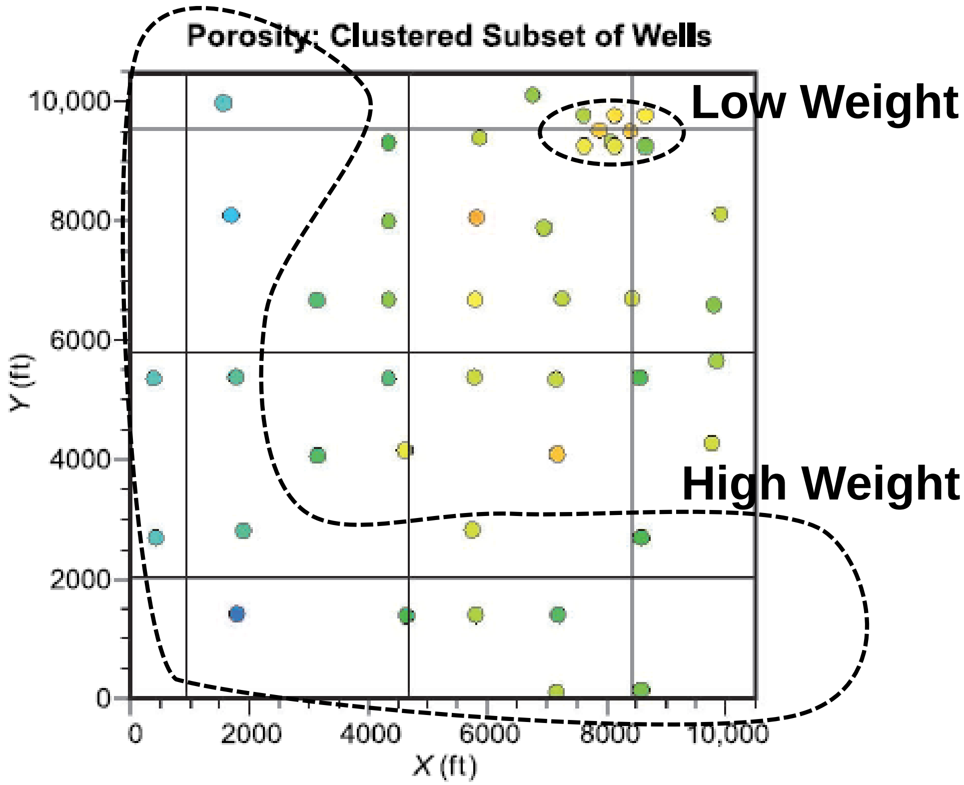
(Geo)statistics
Sampling Bias

Addressing Bias:

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

1. Break model up into subsets.
    - Avoid densely sampled hi quality reservoir inflating average over the entire reservoir

2. Declustering weights
    - Weight based on local sampling density
    - We will do this later



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

(Geo)statistics Cognitive Bias

In any statistical 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.

PGE 337 Lecture 1: Statistics

Lecture outline . . .

- Statistical Methods
- Sampling Methods

Next time

Probability,
Frequentist and
Bayesian Concepts

Introduction

General Concepts

Statistics

Probability

Univariate

Bivariate

Time Series Analysis

Spatial Analysis

Machine Learning

Uncertainty Analysis