3. BIVARIATE STATISTICS

TOOLS

Frequently there is interest in comparing two or more measurements made for the same object or site. Among the most common classical alternatives, we have:

- Scatterplot
- Correlation coefficient
- Regression
- Quantile-quantile plot
- Probability-probability plot

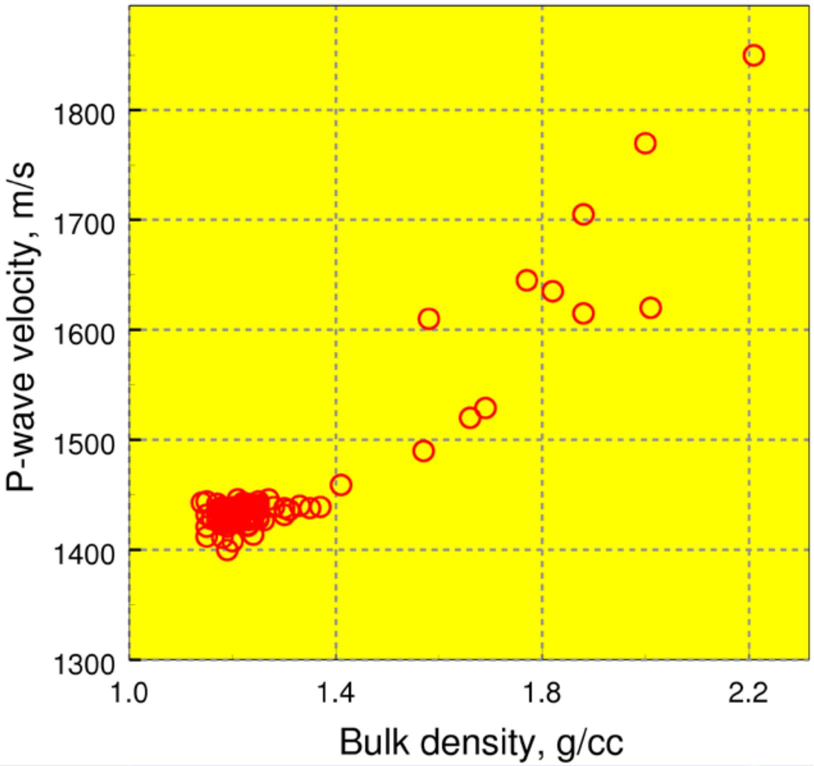
Some of these concepts can be generalized to deal with more than two variables.

SCATTERPLOT

A bivariate scatterplot is a Cartesian posting in which the abscissa and the ordinate are any two variables consistently measured for a series of objects.

Scatterplots are prepared for exploring or revealing form, direction, and strength of association between two attributes.

Mecklenburg Bay sea floor, Germany



COVARIANCE

The covariance is a measure of joint variation.

Given two random variables X and Y with means μ_X and μ_Y, their covariance is the expected value:

$$\mathrm{Cov}_{X,Y} = \mathrm{E}\left[\left(X - \mu_X\right)\left(Y - \mu_Y\right)\right]$$

The covariance estimator when using a sample of point measurements is:

$$\hat{\mathrm{Cov}}_{X,Y} = \frac{1}{n-1}\sum_{i=1}^{n} x_i \cdot y_i - \frac{1}{n \cdot (n-1)}\sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} y_i$$

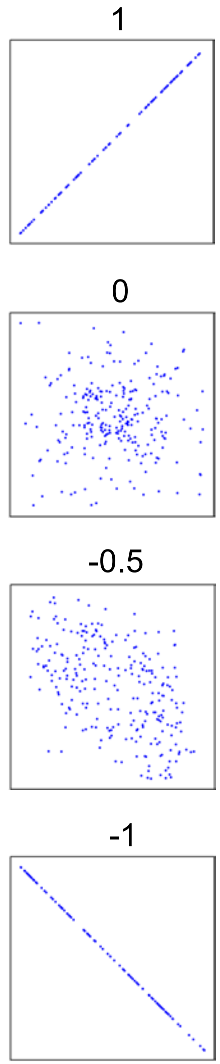
USGS

CORRELATION COEFFICIENT

This coefficient is the number most commonly used to summarize bivariate comparisons. If σ_X and σ_Y are the standard deviations for two variables, their correlation coefficient, ρ, is given by:

$$\rho = \frac{\text{Cov}_{X,Y}}{\sigma_X \cdot \sigma_Y}$$

- It only makes sense to employ ρ for assessing linear associations.
- ρ varies continuously from -1 to 1:
  1, perfect direct linear correlation
  0, no linear correlation
  -1, perfectly inverse correlation

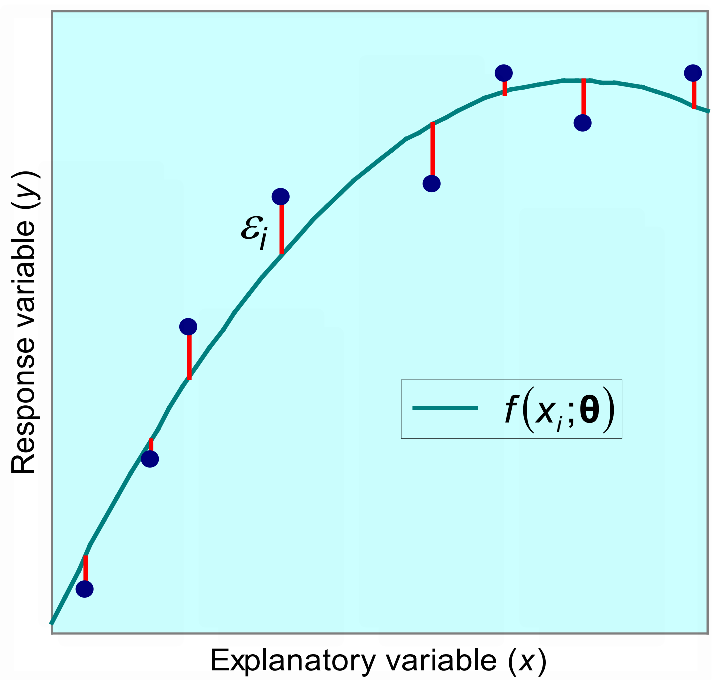


REGRESSION

Regression is a method for establishing analytical dependency of one or more variables on another mainly to determine the degree of their dependency, to estimate values not included in the sample, or to summarize the sample.

- The variable in the abscissa is called the regressor, independent, or explanatory variable.
- The variable in the ordinate is the regressed, dependent, or response variable.
- In many studies, which variable goes into which axis is an arbitrary decision, but the result is different. Causality or the physics of the process may help in resolving the indetermination.

REGRESSION MODEL



- The model is:

$$y_i = f(x_i;\boldsymbol{\theta}) + \varepsilon_i$$

- $f(x_i;\boldsymbol{\theta})$ is any continuous function of x that is judiciously selected by the user. $\boldsymbol{\theta}$ are unknown parameters.
- Term ε is a random variable accounting for the error.
- Having selected $f(x_i;\boldsymbol{\theta})$, parameters $\boldsymbol{\theta}$ are calculated by minimizing total error, for which there are several methods.

Avoid applying the model outside the extreme values of the explanatory variable.

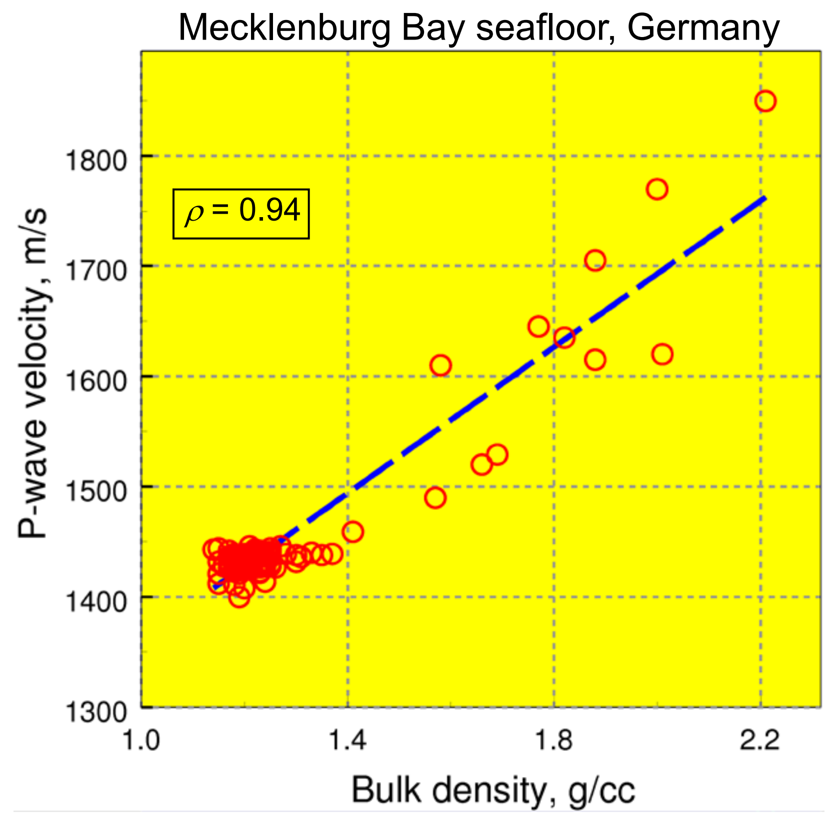
LINEAR REGRESSION

The simplest regression case is one in which the parameters are estimated by minimizing the mean square error, r^2, associated with a linear model,

$$r^2 = \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2$$

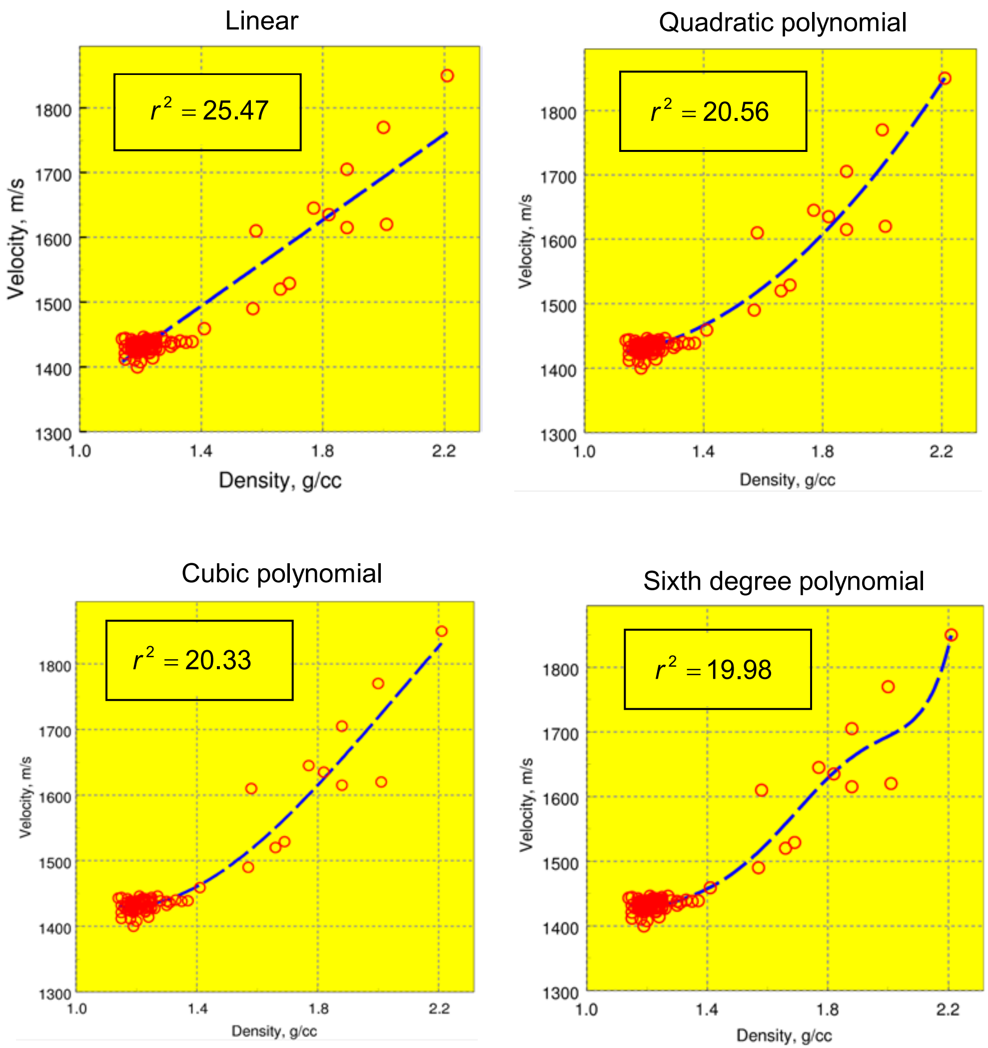
In this special situation, ρ^2 accounts for the proportion of variation accounted for by the regression.

In the example, $\rho = 0.94$. Hence, in this case, the linear regression explains 88% $\left(100 \cdot \rho^2\right)$ of the variation.



USGS

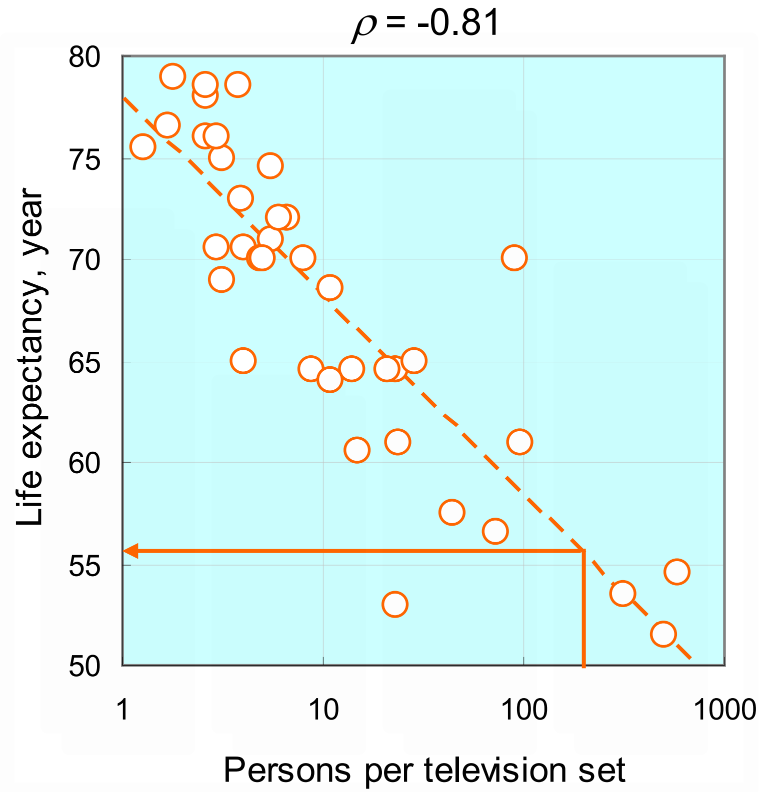
NONLINEAR REGRESSION



- In theory, the higher the polynomial degree, the better the fit.
- In practice, the higher the polynomial, the less robust the solution.
- Overfitting may capture noise and not systematic variation.

IMPLICATIONS

Countries with a population of more than 20 million in 1990



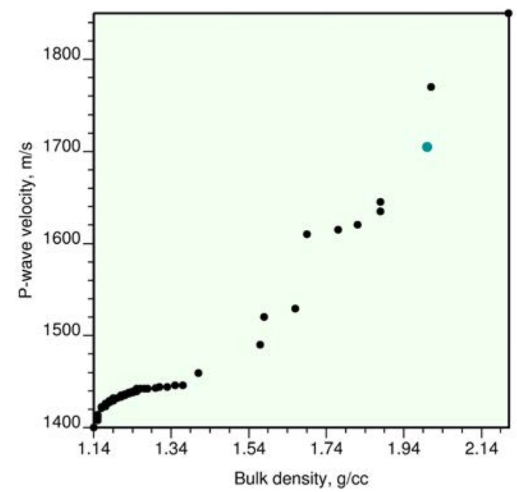
High to good correlation:

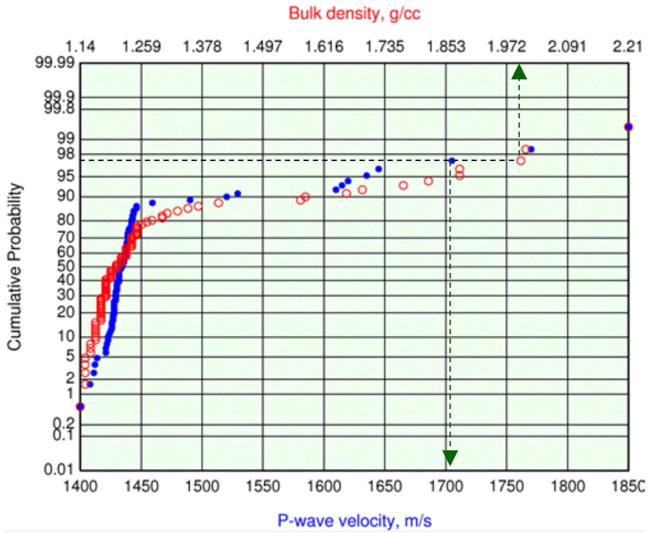
- allows prediction of one variable when only the other is known, but the inference may be inaccurate, particularly if the correlation coefficient is low;
- means the variables are related, but the association may be caused by a common link to a third lurking variable, making the relationship meaningless;
- does not necessarily imply cause and effect.

USGS

QUANTILE-QUANTILE PLOT

- A quantile-quantile or Q-Q plot is a scatterplot based on ranked data.
- The pairing is independent of the object or site where the observations were taken. The first pair of coordinates has the minimum value for each attribute, the second pair is made of the second smallest readings, and so on until finishing with the two maximum values. Interpolations are necessary for different size samples.
- Q-Q plots are sensitive to a shift and scaling of the distributions.

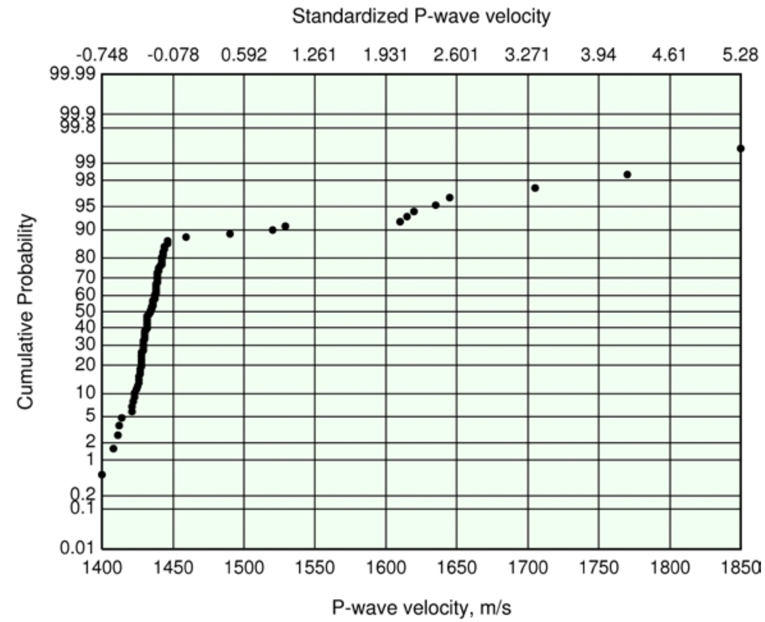




STANDARDIZED VARIATE

If X is a random variable with mean μ and standard deviation σ, the standardized variate, Z, is the transformation:

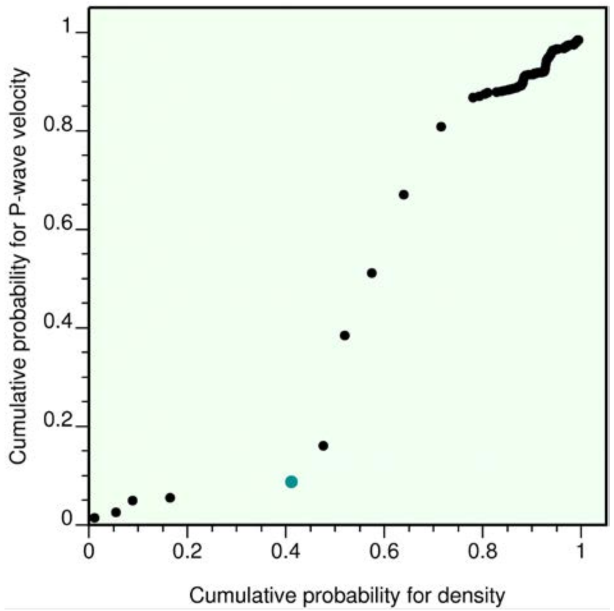
$$Z = \frac{X - \mu}{\sigma}$$

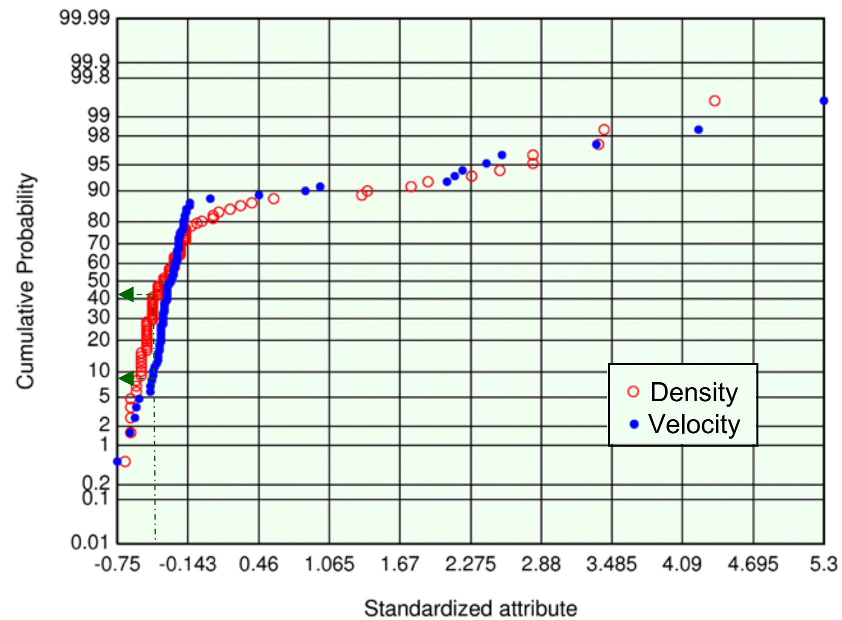


A standardized variate always has a mean of zero and a variance of one.

A standardized Gaussian distribution is a called a standard normal distribution. It is often denoted by $N(0,1)$.

PROBABILITY-PROBABILITY (P-P) PLOT





A P-P plot is another scatterplot prepared by extracting information from the cumulative distributions of two variates.

- If the variates are in different units, preliminary standardization is necessary.
- For given thresholds, the axes show the cumulative probabilities for the two distributions being compared.

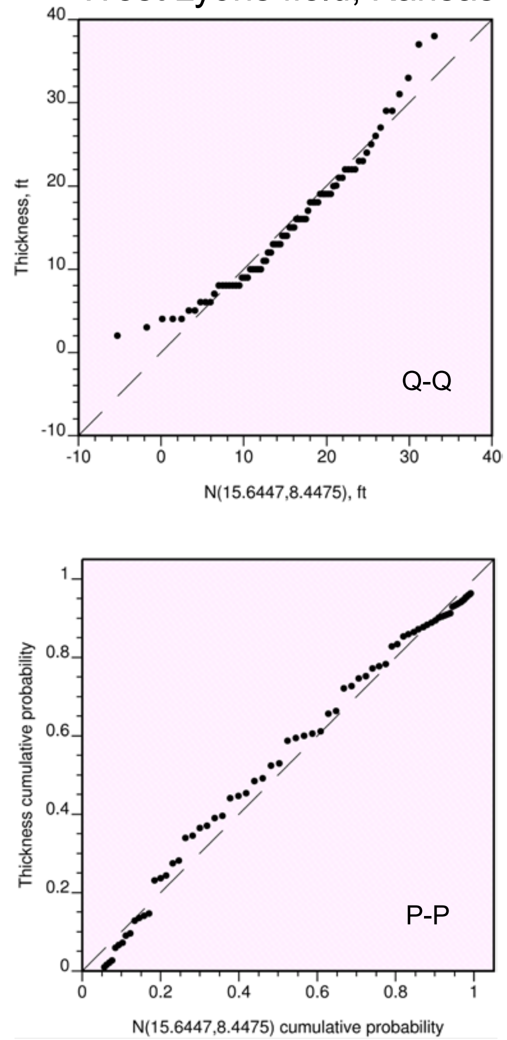
USGS

Q-Q AND P-P PLOTS

Main use of the Q-Q and P-P plots is as a quick way to decide the degree of similarity between two distributions.

- If the distributions are the same, the points align along the main diagonal.
- There is no statistic or level of significance for evaluating the results.
- P-P plots are insensitive to shifting and scaling, and the vertical scale is in probability units.
- The Q-Q plot in the example illustrates its potential for calling the user's attention to the fact that normal distribution with the same parameters can take negative values.

West Lyons field, Kansas



4. SGeMS SOFTWARE

GENERAL DESCRIPTION

The Stanford Geostatistical Modeling Software (SGeMS) is a general-purpose, user-friendly, state-of-the-art geostatistical software package.

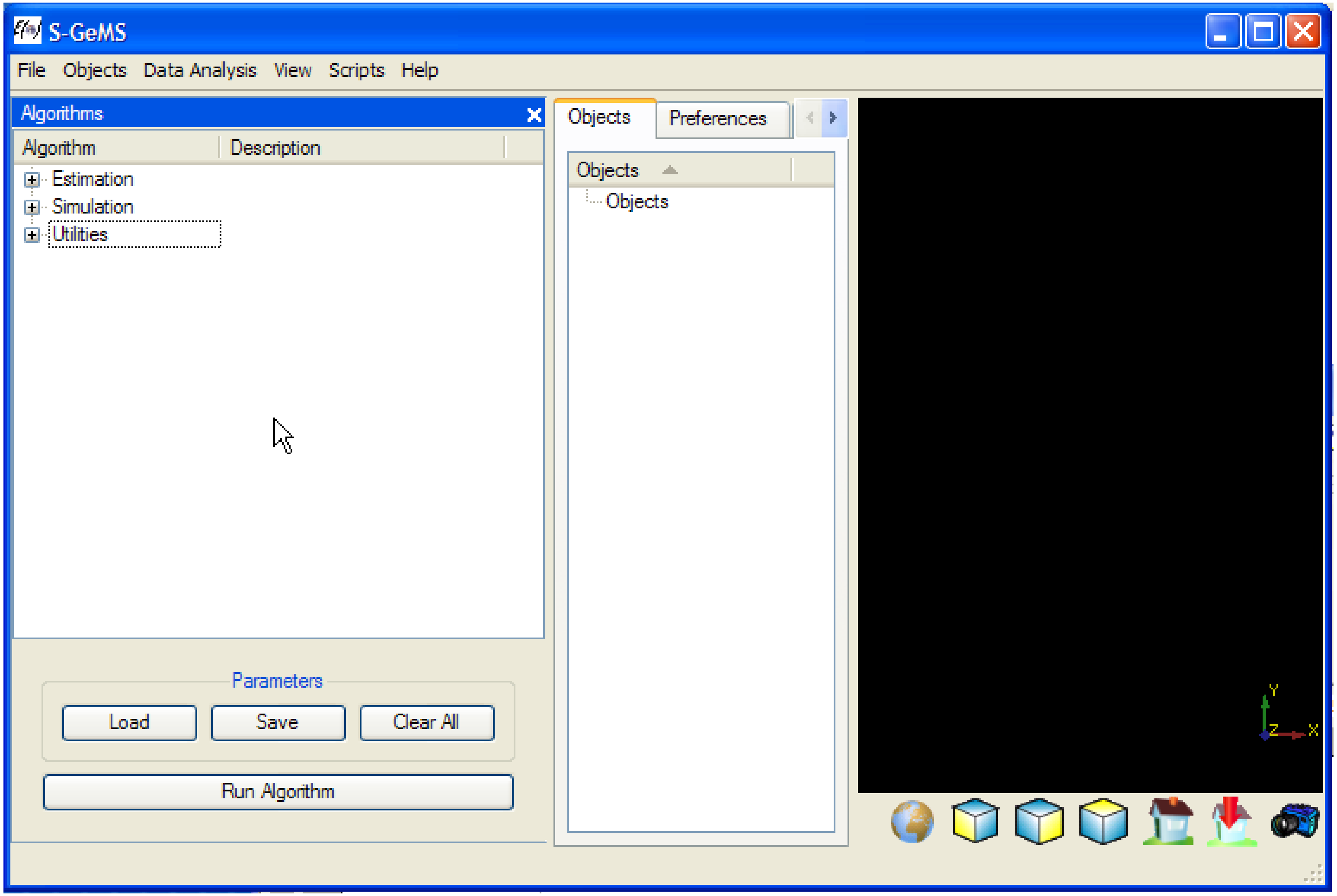
The software code is in the public domain, downloadable from http://sgems.sourceforge.net/ (accessed in October 2017).

The code is in C++ and runs interactively under Windows.

A manual published by Cambridge University Press is listed in the selected bibliography.

SGeMS is an outgrowth of the Stanford Geostatistical Library (GSLIB), with which it shares several features. For details about GSLIB, please consult the book listed in the bibliography.

MAIN MENU



USGS

OBJECTS

Objects are files with numerical information that can be of two types:

A. Data, when each record specifies location.
B. Grid, when locations are in a regular rectangular (usually square) or parallelepipedal (usually cubic) pattern. Location is implicit and specified in a separate file. A grid file contains attribute values only. Typically they are the result of some calculation, but they can be used as input for subsequent processing.

Evidently, no process is possible without prior loading of an object, which is done by:

- clicking on <Objects> in the upper left corner of main menu, or
- dragging the file icon into the Object area of the main menu.

USGS