A Proposal for Standard Test Cases for Checking Geostatistical Software

Deepak Bhandari

Centre for Computational Geostatistics (CCG)
Department of Civil and Environmental Engineering
University of Alberta

There is a need for reference results and standards for geostatistical results. The CCG has begun moving in this direction with the establishment of standard results and implementation of recursive testing capability in CCG-supported software. Some principles and its results are shown in the proposal and all related data and results are given in soft format. We explain how these data and results can be used.

Introduction

Geological feasibility and economic evaluations often call geostatistical models. Nowadays these geostatistical models are created with the help of computers. There are many softwares available that use geostatistical principles for geostatistical analysis. Some times people know the algorithm and use it in their own way in the software to produce the results but not able to compare the produced results with any reference and some times user person get the results but do not know the exact algorithm used by the software. So there is a big gap between software developers and software users. At the same time these softwares also change over time and become very complex.

The aim of this work is to establish reference data and geostatistical results that can be used to check various geostatistical softwares and versions. These data and results can be used by both software developers and software users.

Background

A standard is a widely accepted and approved example of something (e.g. a weight, measure or description); something to which another thing confirms or by which the accuracy or quality of another thing is judged or measured. It serves as a basis for comparison, it is single reference point against which other similar things can be checked or measured for accuracy compliance [5].

The standards and it use exist from ancient times. It is proved by relics from ancient civilizations such as Babylon and early Egypt (around 7000 years ago). The earliest known standards were physical standards for weights and measure. Nowadays there are several standards available (ASTM standards, ISO standards etc.) in different areas. These standards are documented in a particular structure to make them unambiguous and for easy understanding. For example the major sections of an ASTM standard are: Scope, Reference Documents, Terminology, Significance and Use, Apparatus, Reagents, Procedure, Calculation and Results, Report, Precision and Bias, Keywords, Annexes and Appendixes, References, Footnotes [1]. Geostatistics is based on application of statistics and principles in geology and its related field. The use of geostatistical principles is spreading in different fields like mining, petroleum, and environment as more computational speed and software are developed. There are no standards accessible to check the use of geostatistical principles used by different software. There might be individual reference for the individual software company for their own use but as a standard there should be one umbrella of standard geostatistical results under which implementation of geostatistical principles in different softwares can be checked.

There are several technical terms and definitions used in the field of Geostatistics. We will define and explain the keywords and important terminology used in this report as we go on.

Methodology

There are number of geostatistical principles used for analysis of single or multiple data sets e.g. normal score transformation, declustering, variogram modeling, kriging etc. These principle's use, depend on the kind of analysis one wants to do with the data. The major sections of these standards:

Principle: Includes the geostatistical principle.

Scope: Includes the information relating to the purpose of standard, and if the standard has any known limitation.

Results: Includes different cases for the geostatistical principal and its results.

Keywords: Includes the technical terms or words used in the principle and cases along with its reference.

All data and output results are put in soft format. The data files and results are flat ASCII files with a header and data in free format [4]. The header consists of (1) a title line, (2) the number of variables (an integer on the second line), (3) a one-line description of each variable, (4) all the variables per line.

There is one or more number of cases in each principle to establish the results. Each case is defined with name. This name for individual case is the combination of topic number followed by case number in that topic i.e. "case 4.2.1-01" means topic number is 4.2.1 and case is 01.

In the process of establishing these geostatistical results sometimes we need different parameters to be used for different cases. These parameters are explained in the individual case in the report.

Data

We use "cluster.dat" as standard data set to establish the results. At the same time we also use some small and simple data set as "1.dat" and "data.dat". We need small data sometimes to produce simple results easily by manual calculations and to show the results (data) in pictures for easy and clear understanding of principles and its results. All data sets used are as follows:

- cluster.dat: 140 numbers of sample data with both primary and secondary information. The location coordinates of each datum has also been given in the form of X and Y coordinates, where X direction represents East and Y direction represents North. The statistics of the primary and secondary variable is shown in the corresponding histogram plots (Figure 3-1).
- 1.dat: 4 numbers of data (Figure 3-2).
- data.dat: 47 numbers of sample data (Figure 3-3). The location coordinates of each datum has also been given in the form of X and Y coordinates, where X direction represents East and Y direction represents North.

Some Results

Normal Score Transformation

Principle: Many spatial techniques require normal score transformed data as input. Original data are transformed to a standard normal or Gaussian distribution.

$$y_i = G^{-1}(F(z_i)) \quad , \quad i = 1.........n$$

From practical acceptance and transformation to normal space point of views, we consider $F(z_i)$ as follows:

$$F(z_i) = \sum_{k=1}^{i} P_k - \frac{P_i}{2} \quad , \quad i = 1.........n$$

where P_i is the probability of i^{th} value.

Scope: Transforming any kind of distributed data other than normally distributed in to standard normal/Gaussian distribution.

Results: Case 4.1-01: The normal score transformation with 4 numbers of data (“1.dat”) is shown in Figure 4-1. The result in soft format are given in case 4.1-01. Case 4.1-02: The normal score transformation results of “cluster.dat” are given in soft format in case 4.1-02.

Keywords: Standard normal distribution (Ref.: 2.1.1)

Cell Declustering

Principle: Declustering technique assign each datum a weight $(w_i, i = 1,..., n)$ on its closeness to surrounding data.

1. Divide the volume of interest in to a grid of cells $l = 1,......., L$
2. Count the occupied cells L_0, $L_0 \le L$, and the number of data in each occupied cell $n_{l_0}, l_0 = 1,...., L_0$, where $\sum_{l_0=1}^{L_0} n_{l_0} = n =$ the number of data.
3. Weight each data according to the number of data falling in the same cell, for example, for datum i falling in cell l', $l' \in [0, L_0]$ the cell declustering weight [3]:

$$w_i = \frac{n}{n_{l'} . L_0} \quad , \quad i = 1,........, n$$

Scope: Cell declustering technique assign each datum a weight $(w_i, i = 1,..., n)$ on its closeness to surrounding data. Cell declustering is more robust when the boarders of the area of interest are not well defined.

Results: Case 4.2.1-01: The cell declustering results of “cluster.dat” (Figure 4-2) are given in soft format in case 4.2.1-01.

Keywords: Declustering (Ref.: 2.1.2)

Polygonal Declustering

Principle:

1. Define the boarders of the area of interest.
2. Draw polygons for each sample data according to area of influence, considering the surrounding data.
3. Calculate the weights for each datum as follows [3]:

$$w_i = \frac{n.A_i}{\sum_{i=1}^{n} A_i} \qquad , \qquad i = 1,............, n$$

where, n is total number of sample data in the area of interest.

Scope: Ploygonal declustering technique assign each datum a weight $(w_i, i = 1,..., n)$ on its closeness to surrounding data. Polygonal declustering works well when the boarders of the area of interest are well defined. In polygonal declustering the weights are taken proportional to the areas.

Results: Case 4.2.2-01: The Polygonal declustering results of “cluster.dat” are given in soft format in case 4.2.2-01.

Keywords: Declustering (Ref.: 2.1.2)

Variogram Calculation

Principle: The variogram for lag h is defined as the average squared difference of values separated by h distance.

$$2\gamma(h) = \frac{1}{N(h)} \sum_{N(h)} [z(u) - z(u+h)]^2$$

where $N(h)$ is the number of pairs for lag distance h.

Scope: To establish spatial structure or variability.

Results: Case 4.3-01: In this case we use data from “data.dat” and calculate semivariogram (Figure 4-4) for four numbers of lags in directions of both 0 degree and 90 degree azimuth. Case 4.3-02: In this case we use normal score transformed primary data from “cluster.dat” and calculate semivariogram for 10 number of lags, with a lag distance of 4 and lag tolerance of 2. The azimuth tolerance taken is 90 degree (omni-directional) and band width is taken as 10. The calculated semivariogram (Figure 4-5) results are given in soft format in case 4.3-02.

Keywords: Lag (Ref.: 2.1.3), lag tolerance (Ref.: 2.1.4), band width (Ref.: 2.1.5).

Implementation of the Idea with ‘Pangeos’ Software

Setting up standard results and cases in the software in the form that can be useful in future to check the software algorithms can be of much interest to the software developers and industry. Keeping this in mind some trial cases to check geostatistical algorithms in ‘Pangeos’ software has been developed.

Here in the ‘Pangeos’ some geostatistical cases have been setup along with its standard results. Now with the help of a script developed in its *script editor* tool, one can call the individual case, which generate the results of that case and compare it with its standard results. While running, it checks each value of the individual variable against the setup standard reference values, whether they are identical or not. The comparison output is stored in a file with its date and time of execution of the script. From these setup cases any new version of this software can be checked to see whether the results generated with that particular version are identical with the standard results or not.

Discussion and Future Work

This short note documents the start of an important endeavour: establishing reference standard results that can be used systematically for checking algorithms and software.

The standards are not intended to be the results of GSLIB. In fact, as much as possible, the reference results will be generated analytically or with multiple software tools backed up by independent verification.

As the algorithms become more complex, such as simulation algorithms, it is difficult to establish all the results mathematically so the help of programming and software becomes necessary.

References

[1] American Society for Testing and Materials. Retrieved May 21st, 2006, from http://www.astm.org/cgi-bin/SoftCart.exe/studentmember/s101_thestandard.html?L+mystore+zhtb7425+1148254236

[2] C. V. Deutsch and A. G. Journel. *GSLIB: Geostatistical Software Library: and User’s* Guide*, Second Edition*, pages 50-55. Oxford University Press, New York ,1998.

[3] C. V. Deutsch. *Geostatistical Reservoir Modeling*, pages 50-56. Oxford University Press, New York, 2002.

[4] C. V. Deutsch and Chad T. Neufeld. *CCG Software Catalogue Vol. 1*, pages vii-4. Centre for Computational Geostatistics, 2005.

[5] Monash University Library. Retrieved May 21st, 2006, from http://www.lib.monash.edu.au/vl/standards/stand01.htm

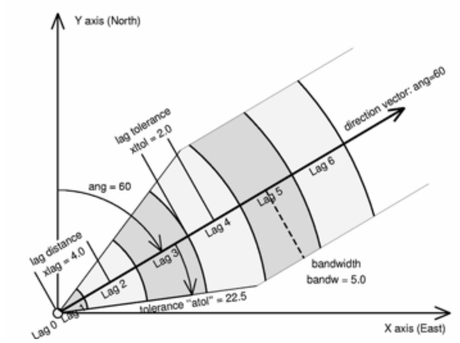


Figure 1: Illustration of lag distance, lag tolerance and band width. *(Source: Deutsch and Journel, 1998)*

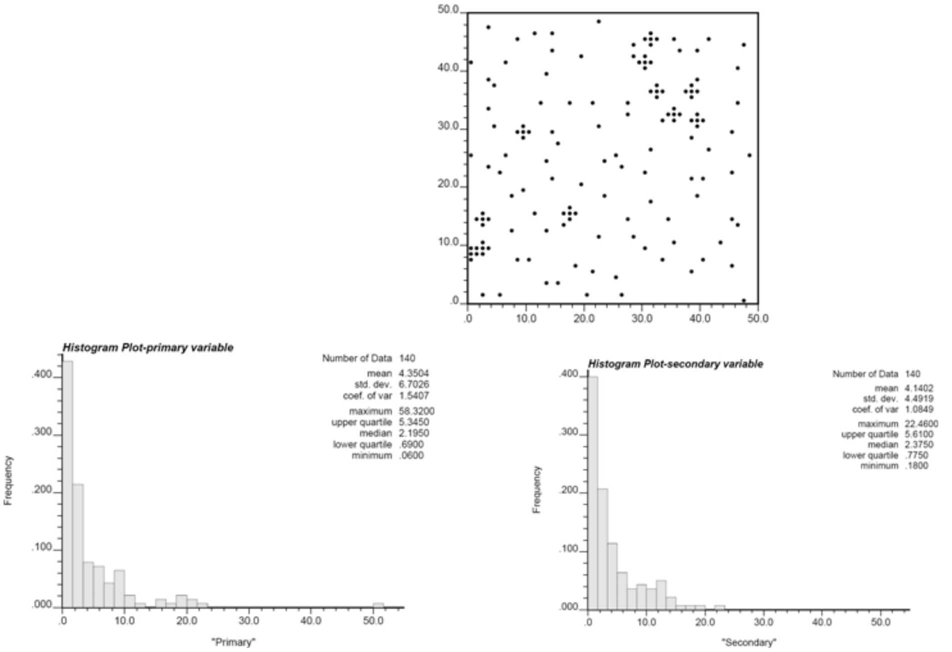


Figure 2: Location map and histogram plot of data – cluster.dat

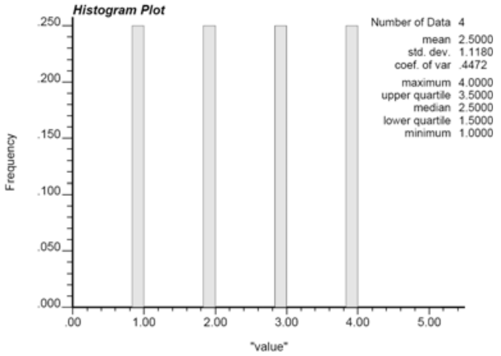


Figure 3: Histogram plot of data – 1.dat

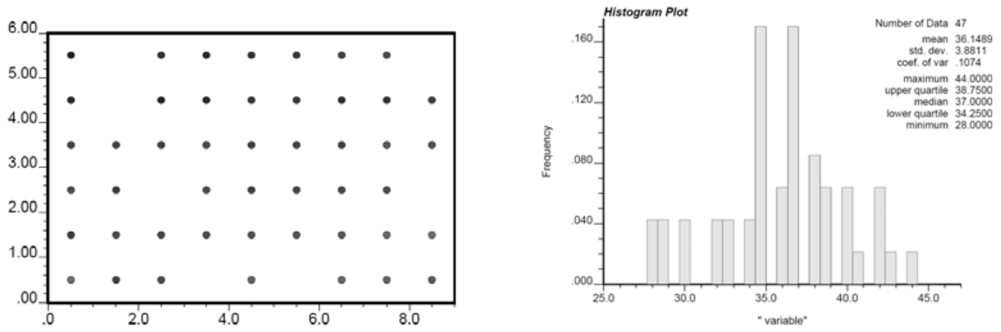


Figure 4: Location map and histogram plot of data – data.dat

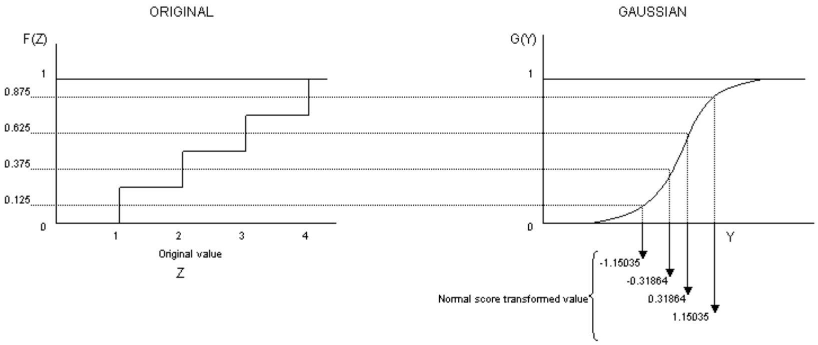
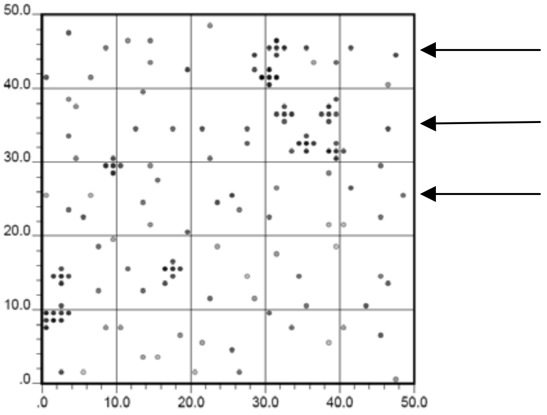


Figure 5: Normal score transformation – case 4.1-01



weight = 140/(3 × 25) = 1.8667

weight = 140/(2 × 25) = 2.8000

weight = 140/(5× 25) = 1.1200

Figure 6: Cell Declustering – case 4.2.1-01

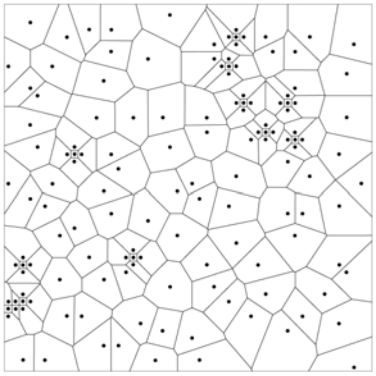


Figure 7: Polygonal Declustering – case 4.2.2-01

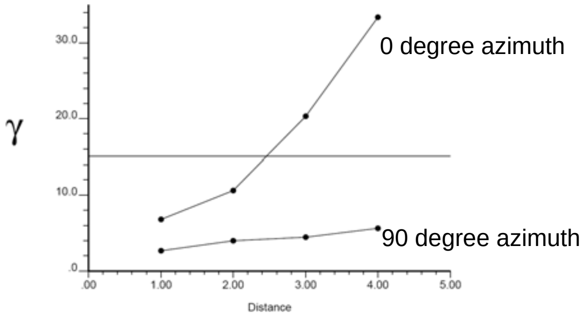


Figure 8: Calculated semivariogram with four lags – case 4.3-01

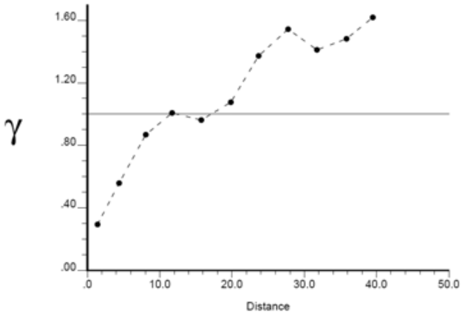


Figure 9: Calculated semivariogram – case 4.3-02