

A method for addressing variable data quality and clustered data

Keefer, D.A. and D.R. Larson, Illinois State Geological Survey, 615 E. Peabody Drive, Champaign, IL 61820

Large databases, such as the central database at the Illinois State Geological Survey, can have advantages and disadvantages when it comes to creating 2-D and 3-D maps and models of geologic systems. These large data sets can provide valuable insight about changes in materials over short distances. These large data sets, however, typically have areas of very clustered data and can provide too much information for inclusion and display in geologic maps and models. The data also are of variable quality; at one extreme, data are precisely located and accurately described, while data at the other extreme have inaccurate locations and incorrect lithologic descriptions. Verifying the correct location for individual records can be very time consuming and evaluating the accuracy of the lithologic descriptions requires inferences based primarily on the comparability of each record with neighboring data. However, without declustering the data and evaluating the accuracy of the data, it can be very difficult to make geologic models that contain the features that the modeler wants expressed.

Computer interpolation (contouring) software packages typically create surface and volume models that consist of uniformly spaced grid nodes, where a single value (e.g. elevation of a stratigraphic contact) is assigned to each node. The value assigned to each node is typically a weighted average of neighboring data points. For the resulting geologic model to capture the variability observed in the data, the spacings between grid nodes must be smaller than the spacing between data values. In areas of clustered data where the data spacing is smaller than the grid node spacing, grid node values will be determined by a larger number of data points than in areas where the data spacing is larger than the node spacing. When large numbers of data points are used to calculate individual node values, the modeler may need to use more soft data values or make many adjustments to grid files in order to produce surface models with realistic geomorphic and stratigraphic features. In areas with clustered data and large local variability of values, more soft data will be needed to produce models with realistic features than in areas where either local variability is small or data are not clustered.

Methods for addressing the clustering of data depend on how the quality of the data are treated. The data can be viewed as having a relatively consistent problem with errors, where errors are assumed to be unidentifiable but fairly similar between wells. If this scenario describes the data errors, geologic models created using a grid spacing that is larger than the minimum data spacing (i.e. models based on clustered data) should predict the average behavior of the deposits. The data can also be viewed as being variable in data quality, with some records being more accurate than others. If this scenario describes the data errors, then there can be some benefit in trying to create geologic models using only a subset of the data (i.e. declustered data sets); this subset ideally includes only the most accurate and relevant records. The spatial distribution of this subset can be carefully selected so the data have a relatively unclustered distribution, with a minimum data spacing that is larger than the grid node spacing of resulting computer maps and models. Models created with these declustered data sets can be easier to modify, creating surface maps and volume models that are reasonable according to interpretations of the geologist. The difficulty with attaining this idealized data set from a large database can be in developing a method to evaluate the accuracy and relevancy of any given data point.

Using a data set from East-Central Illinois, which contains data of variable quality that is spatially clustered at a scale smaller than the selected grid node spacing, a ranking system is presented for evaluating the accuracy and relevance of geologic data. This ranking system explores the use of five different characteristics of geologic data, including: lithologic value; hydrologic value; spatial importance; driller reliability; and, land-surface elevation reliability. The "Lithologic Value" of a datum is

determined using five parameters to evaluate the probable reliability and importance of the lithologic data from each record. These parameters include: lithologic description detail; total depth; purpose of borehole; presence and type of geophysical logs; and, presence and type of samples. The “Hydrologic Value” of a datum is determined using five parameters to evaluate the relative completeness of each well record in describing the hydrology of the materials that the well is constructed in. The “Spatial Importance” of a datum is a measure of the area (or volume for 3-D modeling efforts) that each point represents. This characteristic looks at the clustering around each data point and at the total depth of the hole. Data points that have many close neighbors will have less importance than points with fewer close neighbors. The “Driller Reliability” of a datum is self explanatory and is based on the recognition that some drillers provide more accurate well logs than others. For this factor, the Lithologic Value of well logs are summarized for each driller and compared to the Lithologic Value of well logs from other local drillers.

To use this ranking system, each datum is rated for each of the five characteristics. Then, based on the relevance of each factor to the geologic modeling priorities, the entire data set can be sorted by these ratings. These ratings can be used for exploratory data analysis and for declustering of large data sets. The five component ratings can also be used for more focused explorations of the database. The use of these ratings for exploratory data analysis is demonstrated for a data set in East-Central Illinois. The ratings are used to evaluate the local variability in elevation of specific stratigraphic contacts. The ratings are also used to look at larger-scale variations in these same stratigraphic contacts. Analysis of the ratings of the five component characteristics will also be presented. Finally, these ratings will be used to compare and contrast surface models made from clustered and declustered data sets.

This ranking approach is simple, and ensures that the specific nuances of the data set and the distribution of materials are used to determine the relevancy of the data. The use of this type of ranking-based declustering approach for geologic modeling can allow the modeler to more easily control the features of geologic models. The identification and selection of a more accurate subset of declustered data will reduce the variability of the data sets and resultant geologic models. This type of ranking system also can be useful in early stages of a mapping project. Depending on the spatial distribution and the interpreted relevance of the data, only a subset of records might be selected for locational verification and use in modeling. In projects with large data sets this can result in a significant savings of time and money.

