AMARILLO BY MORNING: DATA VISUALIZATION IN GEOSTATISTICS

William V. Harper[1] and Isobel Clark[2]
[1]Otterbein College, United States of America
[2]Alloa Business Centre, United Kingdom
wharper@otterbein.edu

"Amarillo by morning, Amarillo's where I'll be" comes from a country song released by George Strait in the 1980's. Around this time Bill moved to the Amarillo Texas area hoping to bury high-level nuclear waste. Like a fine country song, this presentation paints a haunting visual image of the Wolfcamp aquifer underlying the waste site. If a breach occurred, how many mornings until the nuclear waste arrived at Amarillo? Using a variety of visualization tools blended with geostatistical methods, an overview of universal kriging is given at an introductory level. This project was the source of Bill's first geostatistical analysis and was sadly killed in 1988 by the US government. On the plus side, he met his lovely Texas bride Paula there. Meanwhile he and co-author Isobel have continued their geostatistical journey teaching classes which range from 14 year-old Slovakian High School kids to PhD candidates and beyond.

GEOSTATISTICAL DATA VISUALIZATION

Geostatistics was not initially developed by the statistical community but instead has its roots in mining and geology. Unlike classical statistics in which observations are assumed to be independent, data in 2-D or 3-D are spatially correlated with values close to each other in distance frequently exhibiting similar values. Each data value has a location in space. The major goal is to estimate values at locations that have not been sampled. Geostatistics offers many helpful visualization tools to aid in the analysis of such spatial data. Due to the proceedings page limit this paper focuses on a few visualization tools. A more complete set of visualization tools are found in the associated PowerPoint file presented at the ICOTS meeting and stored on the web at http://faculty.otterbein.edu/WHarper/ICOTS2010GeostatisticsVisualization.ppt.

The U.S. Department of Energy studied possible high-level nuclear waste sites in salt, basalt, and tuff. One of the salt sites studied was in the panhandle of northern Texas close to the New Mexico and Oklahoma borders. The Wolfcamp aquifer underlies the potential repository and provides possible travel paths for leached radionuclides to travel. Figure 1 shows the general setting for the 85 potentiometric (groundwater pressures) values from this aquifer (Harper & Furr, 1986). With higher values generally in the lower left (southwest) and lower values in the upper right (northwest) the groundwater gradient would cause water to flow in a northeasterly direction from the repository in Deaf Smith County toward Amarillo in lower Potter county.

The geostatistical technique used is universal kriging (Cressie, 1993; Clark & Harper, 2009) that produces minimum variance linear unbiased estimates. Many iterative steps are involved including distribution analysis, possible data transformation, semi-variogram modeling of spatial variability including trend analysis, cross validation of the proposed model, and finally the kriging that produces grids for mapping of both the expected potentiometric values at un-sampled locations along with a corresponding standard error grid.

A typical geologic set of data like the Wolfcamp data consists of irregularly spaced measurements collected over a given geologic region. For this analysis the Wolfcamp aquifer is assumed to be a two-dimensional plane with uniform thickness. As mentioned above geologic measurements generally have similar values when the distance between them is small. As this distance increases, the variability often increases at least up to a certain point known as the range. The longitude and latitude values given in Figure 1 are converted to miles for this analysis.

Perhaps the most important aspect of a geostatistical analysis is the assessment of the spatial variability used in the kriging prediction that provides both expected values and standard errors at un-sampled locations so that travel paths and times can be computed in a subsequent study not covered here. For data not on a regular grid it is important to evaluate distances between nearest neighbors so that reasonable data collection distance bins may be established to perform semi-variogram modeling of spatial variability. Figure 2 provides an interesting graphic depiction of

In C. Reading (Ed.), *Data and context in statistics education: Towards an evidence-based society. Proceedings of the Eighth International Conference on Teaching Statistics (ICOTS8, July, 2010), Ljubljana, Slovenia*. Voorburg, The Netherlands: International Statistical Institute. www.stat.auckland.ac.nz/~iase/publications.php [© 2010 ISI/IASE]

where the nearest located neighboring value is located for any individual data value. The distances between the points are summarized in a histogram to determine the semi-variogram data bin sizes.

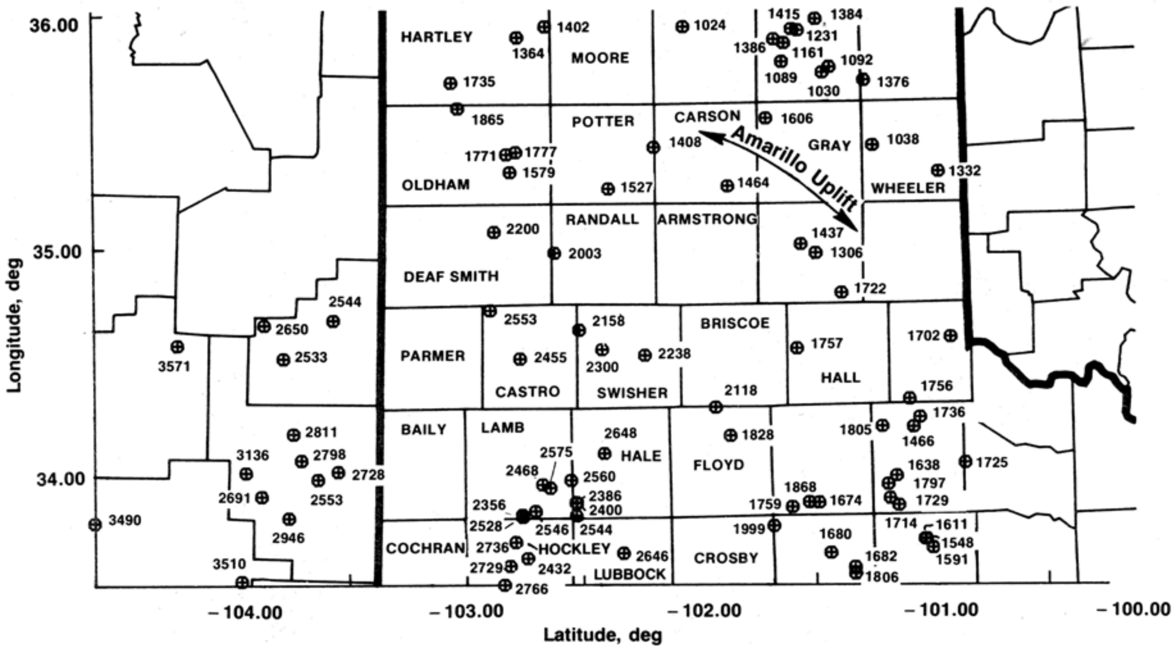


Figure 1. Wolfcamp Potentiometric data in Texas and New Mexico

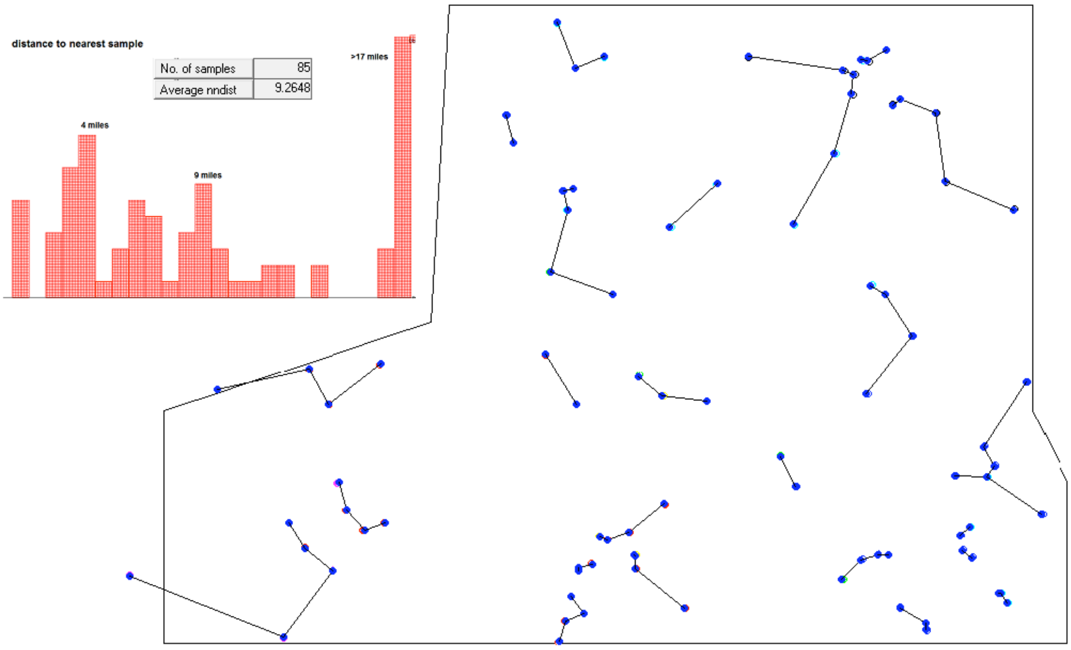


Figure 2. Nearest Neighbor identification for the Wolfcamp Potentiometric data

The theoretical semi-variogram is estimated by the empirical semi-variogram $\gamma^*(h)$ defined as $\gamma^*(h) = \frac{1}{2N_h}\sum_h \left(x_i - x_j\right)^2$ where x_i, x_j are data values and N_h is the number of data values roughly a distance of h miles apart perhaps in a specified direction. The process creates a cloud of individual terms of the semi-variogram (found in the PowerPoint presentation) summarized in Figure 3 into directional semi-variograms to provide a visual way to investigate anisotropy and/or trend. Anisotropy implies that the spatial variability changes based on direction.

Figure 3 shows the spatial variability on the vertical axis increases more rapidly in the southwest to northeast direction as the distance *h* between the data values increases on the horizontal axis. For this application each bin covers a distance of 5 miles, e.g., the first bin collects the data for the directional empirical semi-variograms for points that are 0 to 5 miles apart. Each of the four major directions of the compass are shown with arrows pointing in the relevant direction and are also color coded (shown in PowerPoint) to allow the user to more easily follow a given direction. The size of the arrow is indicative of how many data pairs N_h were found for that direction at a given distance apart. Figure 4 shows the corresponding semi-variogram directional shading plot that aids the visual assessment. The parabolic upper right tail of the northeastern directional semi-variogram is indicative of possible trend that must be investigated before assessing anisotropy.

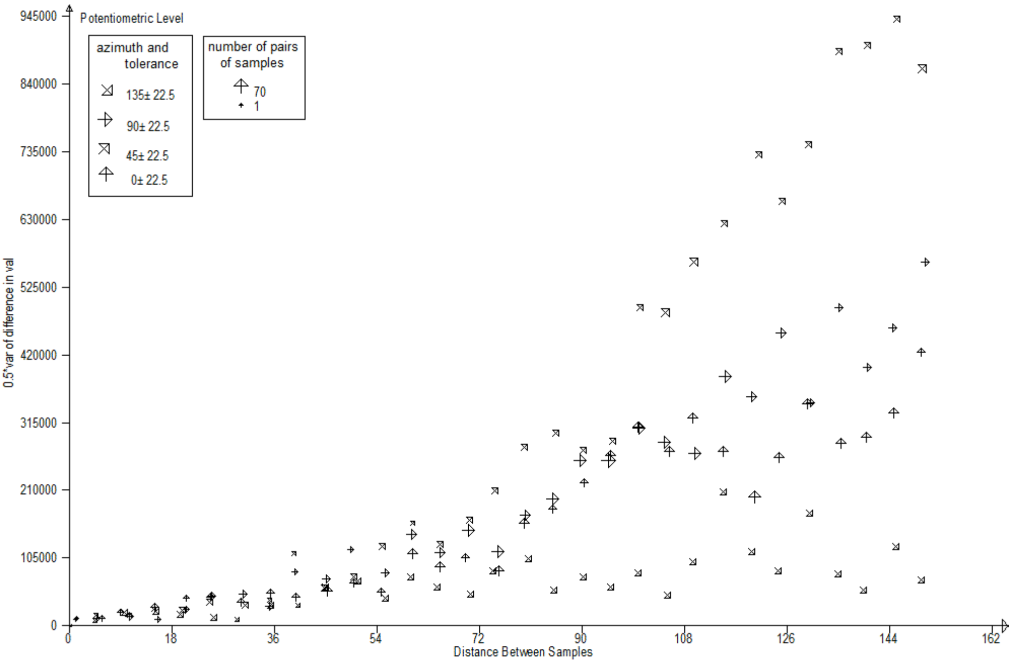


Figure 3. Directional Semi-Variograms for the Wolfcamp data

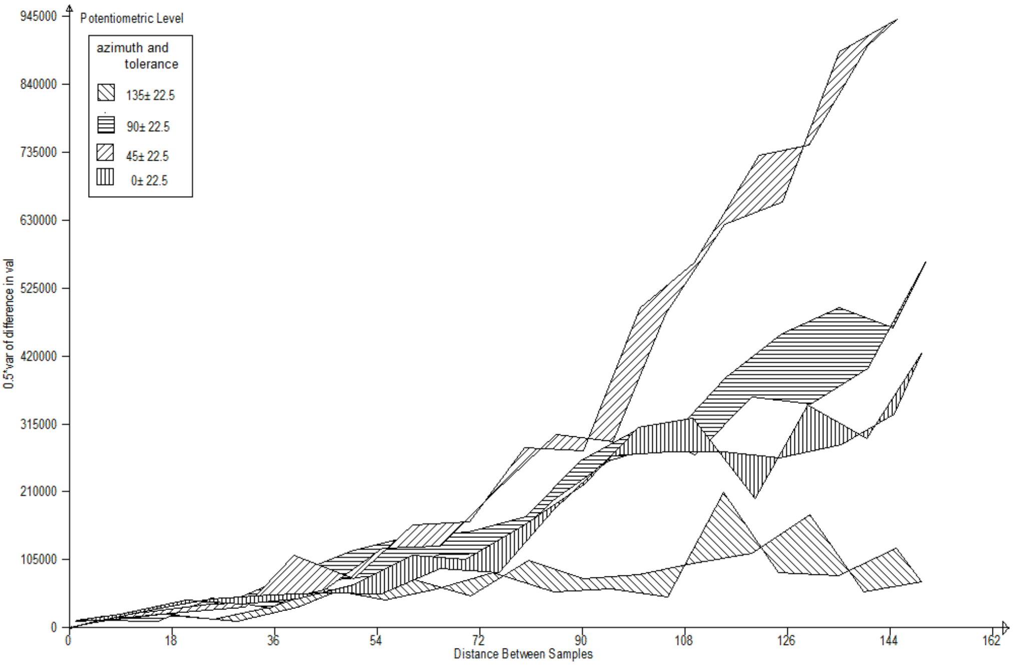


Figure 4. Directional Shaded Plot Semi-Variograms for the Wolfcamp data

A global trend assessment is performed using traditional least squares regression to assess what local level trend might be advisable in universal kriging. While a quadratic global trend is statistically significant at the 95% confidence level, it only increases the variability explained by the regression from 89% for a linear fit to 91% for the quadratic. Universal kriging combines a local trend fitting in a specified search radius with the kriging weighted individual values in that search area. Using the residuals from the global trend assessment the directional semi-variograms fit to the linear residual trend is shown in the shaded plot (PowerPoint also has the directional semi-variogram similar to Figure 3 for the linear residuals) in Figure 5. This shaded semi-variogram plot shows similar variability in all four major directions when the trend has been accounted for. Thus there is no visual evidence of anisotropy and the data is then considered to be isotropic.

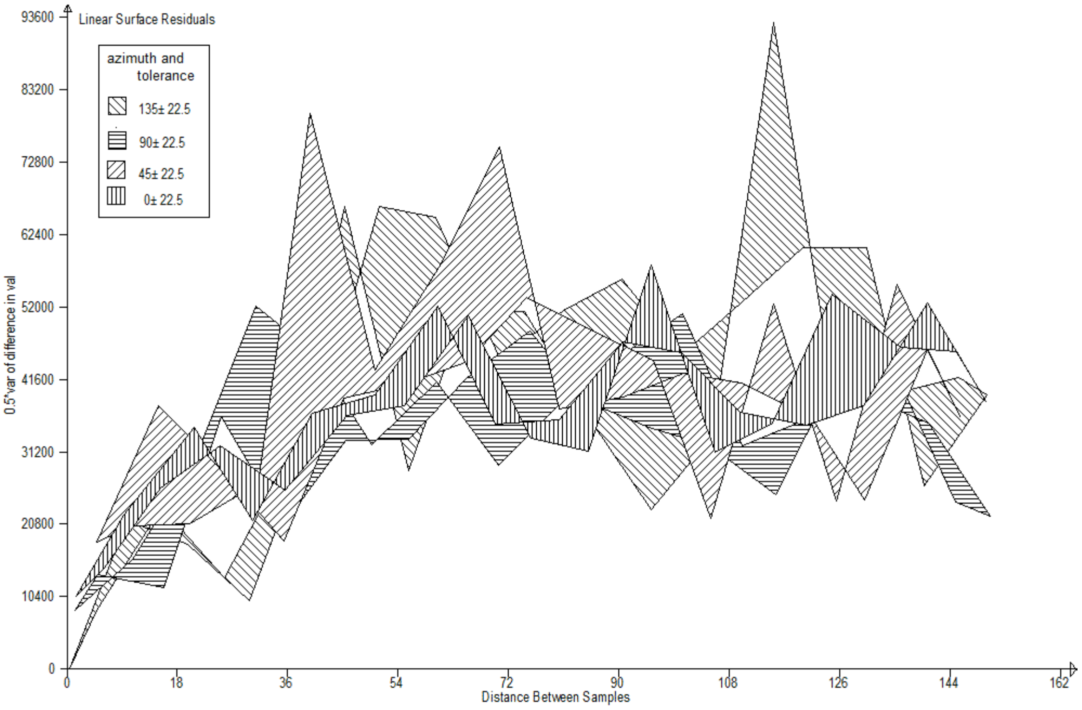
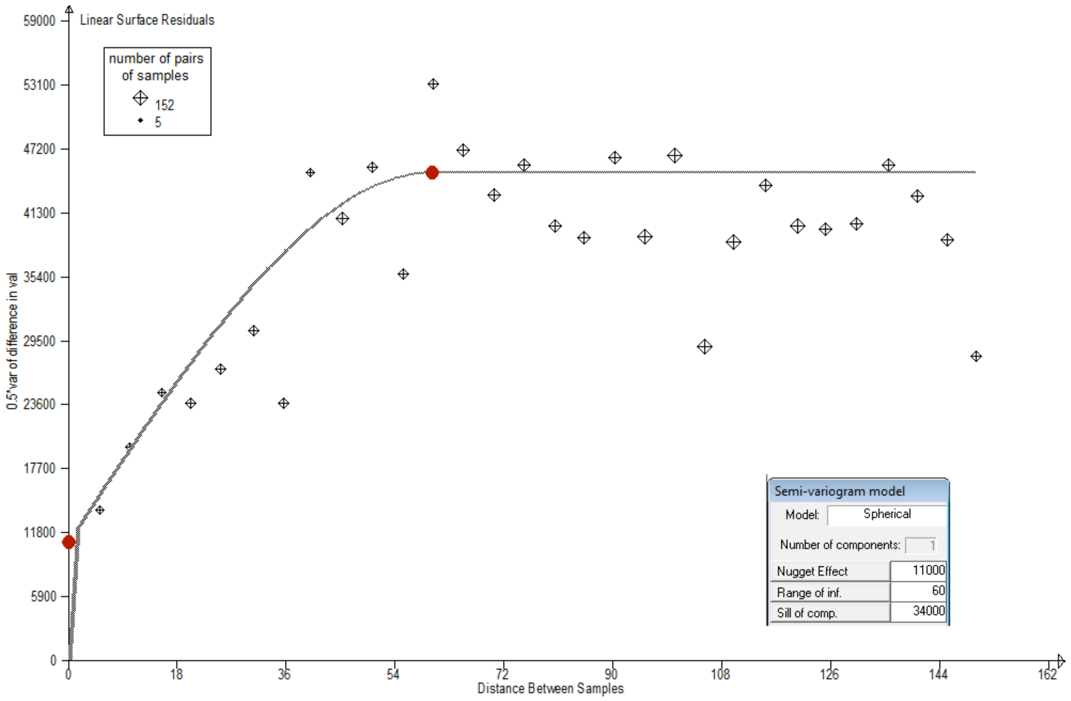


Figure 5. Directional Shaded Plot Semi-Variograms on Linear Regression Residuals

Since the data are felt to be isotropic an omni-directional semi-variogram is used to model theoretical semi-variograms as seen in Figure 6. There are various theoretical semi-variogram to try as may be seen near the top of Figure 6. The Wolfcamp linear residuals are well fit by a spherical semi-variogram with a nugget of 11,000 ft^2, a sill of 34,000 ft^2, and a range of influence of 60 miles. The nugget is an estimate of the spatial variance as the limit between adjacent data values approaches zero. As the distance between sample points increases along the horizontal axis, the spatial variance grows for a spherical semi-variogram until the observations appear to be independent from one another at the range of 60 miles. From 60 miles on the spatial variability levels off at 45,000 ft^2 which is the sum of the nugget and sill.

Geostatistics involves iterative modeling to arrive at a model that fits the spatial data. In addition to visual assessments, cross-validation is used to partially verify that the model does not have obvious flaws in fitting the data. Each data value is removed one at a time and estimated by the surrounding values. The cross-validation residuals are normalized by their kriging standard error and these standardized residuals ideally have an average of 0 and a standard deviation of 1. For the above model assumptions using a 60 mile search radius that determines which values are used to estimate each removed value the average and standard deviation are 0.00 and 1.04, respectively. While this does not guarantee the above model choices are all ideal, it does provide a degree of comfort. Cross-validation graphics (shown in PowerPoint) visually identify where the larger deviations of predicted versus actual values occur providing another visualization tool for investigation of spatial data including the search for possibly anomalous data.



Model: Spherical
Number of components: 1
Nugget Effect 11000
Range of inf. 60
Sill of comp. 34000

Figure 6. Omni-direction Semi-variogram on Wolfcamp Linear Residuals

After the modeler is content with the cross-validation and the visual assessments, universal kriging is performed. This was done on a 10 mile grid and the results passed to a contour plotting package such as Surfer. Figures 7 and 8 are the resulting predicted potentiometric surface and the corresponding standard error map developed from the universal kriging grids. Thus one can assess the uncertainty of the prediction. Additionally the standard error map can be used to help decide where additional data might be collected to reduce travel time uncertainty among other things.

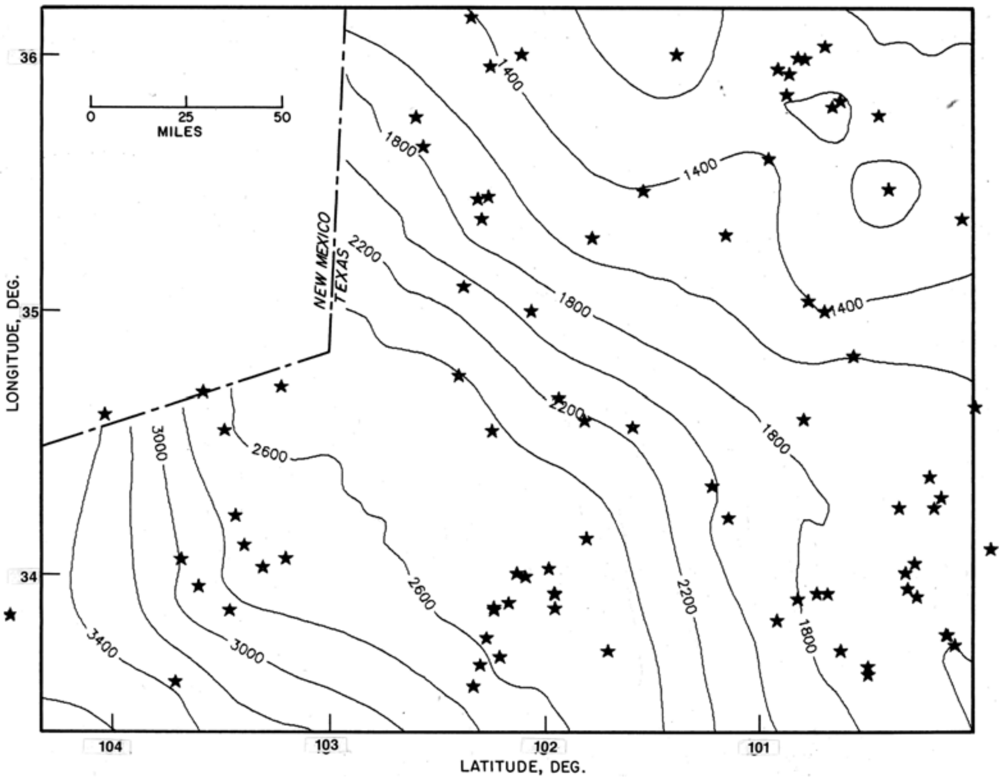


Figure 7. Universal Kriging Potentiometric Surface Estimation for the Wolfcamp Aquifer

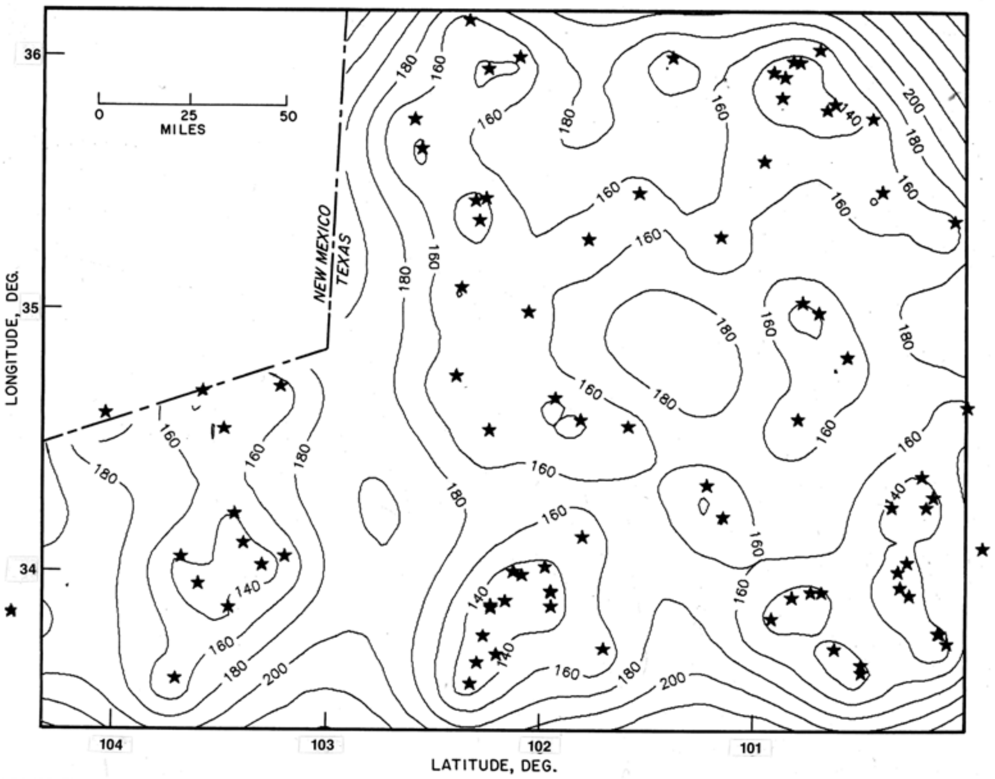


Figure 8. Universal Kriging Standard Error Map for the Wolfcamp Aquifer

Far field flow simulation modeling in the unlikely case of radionuclides leaching into the Wolfcamp aquifer predicted groundwater travel times from 141,000 to roughly 800,000 years (Devary et al., 1984). The most likely flow paths headed toward Amarillo. With the recent change in U.S. policy in 2009, it now appears that the Yucca Mountain tuff site not killed by congress in 1988 may now be dropped and new alternatives considered including possibly re-opening of the old salt and basalt sites. Time will tell what is in store for some morning in Amarillo.

CONCLUSION

Key geostatistical visualization tools are illustrated with data for the Wolfcamp aquifer that underlies an area that once was considered for a potential high-level nuclear waste repository. Without such visual methods, geostatistical modeling would be a mathematical black box with many errors and misconceptions. As with most statistical applications, visualization tools aid both correct modeling and understanding of data. This is especially true when modeling two or three dimensional spatial data sets.

REFERENCES

Clark, I., & Harper, W. V. (2009). *Practical Geostatistics Case Studies 2009.* Columbus, Ohio, USA: Ecosse North America LLC.

Cressie, N. A. C. (1993). *Statistics for Spatial Data*. New York: John Wiley.

Devary, J. L., Harper, W. V., Sykes, J. F., & Wilson, J. L. (1984). Far-Field Flow Uncertainty Analysis for the Palo Duro Basin. *Materials Research Society Symposia Proceedings: Scientific Basis for Nuclear Waste Management VII, 26* (pp. 397-404). Burlington, MA, USA: Elsevier Science Publishing.

Harper, W. V., & Furr, J. M. (1986). Geostatistical Analysis of Potentiometric Data in the Wolfcamp Aquifer of the Palo Duro Basin, Texas, *BMI/ONWI-587*, Columbus, Ohio, USA: Battelle Memorial Institute.