2001

# Propagation of Errors in Spatial Analysis

Peter P. Siska

I-Kuai Hung
*Arthur Temple College of Forestry and Agriculture, Stephen F. Austin State University*, hungi@sfasu.edu



## Recommended Citation
Presented at the 24th Applied Geography Conference, Fort Worth, Texas 2001

# PROPAGATION OF ERRORS IN SPATIAL ANALYSIS

Peter P. Siska
I - Kuai Hung
College of Forestry
Stephen F. Austin University
Nacogdoches, TX 75692-6901

## 1. INTRODUCTION

In most spatially oriented projects, the conversion of data from analog to digital form used to be an extremely time-consuming process. At present, industrial and research institutions continue to accumulate large quantities of data that are easily accessible to users worldwide, and consequently less time is spent for data input. In addition, the introduction of Internet2 rapidly increased the transfer of spatial data through the electronic highway and opened new avenues for collaboration among research institutions and scientists. It is apparent that this trend will continue in the future. New regional and national centers for spatial data are being established with the objective of providing data to natural resource institutions and developing a high-resolution database of regional significance. Therefore the questions of spatial data accuracy and quality are of utmost importance. The purpose of this paper is to discuss the propagation of errors, outline the major trends and problems that are encountered during spatial data analysis, and demonstrate the propagation of errors during raster data conversion in a GIS environment. The results of this study will contribute to an understanding of errors emanating from the conversion of irregularly spaced points to regular grids using different interpolation methods.

## 2. BACKGROUND AND METHODOLOGY

A critical examination of errors should be an important part of spatial analysis. In general, an error is a deviation from reality. The reality, however, is represented by measurements that already have a certain amount of noise associated with them. Hence the truth or reality often depends on a number of factors that may not be fully determined. Therefore, an error is defined as a difference between the observed and the fitted values and represents a sample from the population of distorted versions of the same truth, just as the Gaussian distribution is used to represent different observations of the same scalar measurement (2). The term ‘propagation of errors’ has been used in association with spatial functions that are available in GIS. Particularly during the raster overlay operations, the magnitude of error increases quite rapidly depending on the function used. For example, assume two raster files are used in an overlay operation and suppose that the error in each grid cell of both files is approximately ten percent, then using the additive function in GIS produces a final map also a ten percent error. However, when the multiplicative function is used in the spatial operation process then the resulting error is inflated up to 20 percent for each grid cell of the final map. The magnitude of errors naturally increases with an addition of every new layer entering the overlay process. A

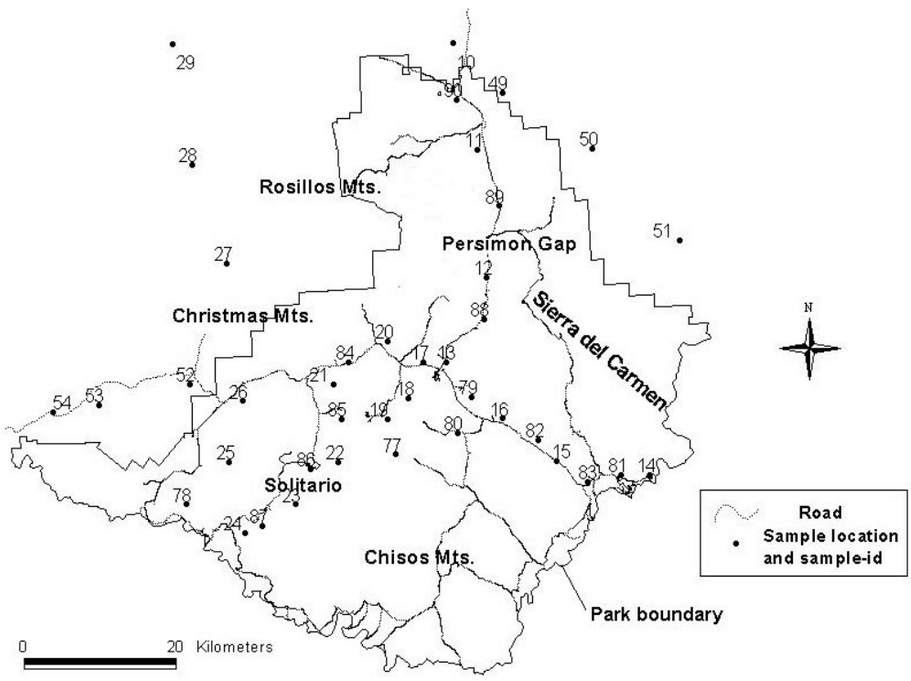
few papers reflected the need to analyze error propagation during the overlay operation (2, 6, 7). The error analysis results in developing models for predicting the errors that tend to propagate independently on each cell during the raster overlay operation (3). In an ideal situation, spatial metadata should provide users with information about a variety of errors that are inherent in the data sets. Such information, however, is rarely available.

A few examples of errors associated with the data are: sampling error, experimental error, and model error. The objective of this paper is to analyze a specific type of errors, particularly the ones that emanate from the interpolation process in GIS. There are a number of interpolation methods used in spatial analysis, and new studies are needed to evaluate the distribution and magnitude of errors that originate from different interpolation methods. While a few studies have attempted to evaluate errors from kriging interpolation (1), this paper addresses the errors from three interpolation methods that are frequently used in geosciences. These methods are:

a) Inverse distance weighted
b) Splines
c) Kriging (Ordinary)

The performance of each method was tested and the results were analyzed and compared using the fundamental statistical parameters. The results of this research contribute to the understanding of data quality in natural resource management studies, GIS, cartography and geography.

FIGURE 1. SAMPLE DISTRIBUTION




## 3. DATA

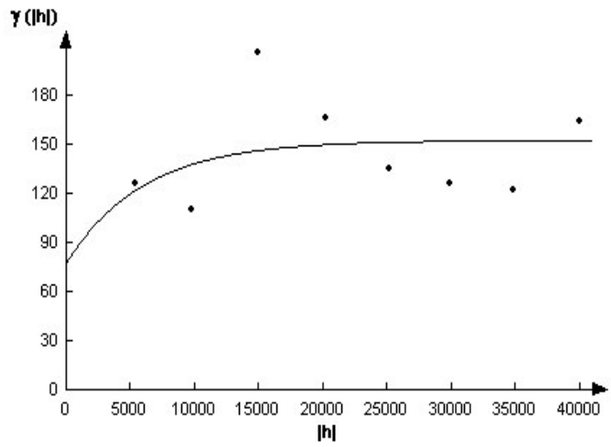
Unexpectedly, the errors begin creeping into the database before any measurements are performed on studied phenomena. The nature of earth phenomena is extremely complex and our knowledge of natural laws often reflects only a fraction of reality. It is impossible to select a perfect sampling strategy and analyze data without any errors. Even though the sampling method might be perfectly unbiased, natural conditions do not always allow taking a sample and performing measurements on a specific random location. In addition,

it is also impossible to take an infinite number of samples from every geographic location. This study depicts a real world situation in which spatial analysis has a significant bias due to the sampling strategy that follows the park road system (Figure 1). The pollen data used are data collected in Big Bend National Park, representing frequencies of the composite pollen grains. Composite pollen comes from a large family of plants generally called weeds. They comprise pollen from dandelions, sunflowers, ragweeds and sagebrush, all of which produce large quantities of pollen. Since the safety of the field crew was an important issue, the accessibility to the main road system dominated the selection of location for field sampling. Hence spatial analysis has to take these difficulties into account. In addition, the selection of appropriate tools is very important for accuracy of analysis. Therefore the errors from three interpolation methods were analyzed to determine which interpolation would be optimal for mapping in a situation with inherently biased sampling.

## 4. ANALYSIS

In the process of mapping natural phenomena the point data sets are frequently converted to the raster form using an interpolation method. Figure 2. depicts the exponential variogram model for composite pollen data: $\gamma(h)$ =78.08 + 153.5 Exponential h/18,008, which best fitted the pollen data. This figure indicates that there is a significant amount of error present in the data before an interpolation is performed. The high nugget effect (78.08) indicates that the experimental error and the error due to the short scale variation are more than 50%. Under ideal circumstances the nugget effect should be zero. The nugget effect is the value of gamma (variogram) when the distance (h) is zero. However, due to the errors mentioned earlier the variogram curve intercepts the vertical coordinate above the origin as it is indicated in Figure 2.

FIGURE 2.
EXPONENTIAL VARIOGRAM FOR COMPOSITE POLLEN DATA




In addition, increasing the nugget effect inflates the interpolation prediction error (kriging variance) and hence increases uncertainty of estimated values. As such, the quality of the raster data is devalued. Therefore the spatial analysis prior to interpolation possesses two major sources of errors: a) errors associated with the sampling strategy and b) measurement errors. The decision about the appropriate interpolation method must be made to predict the values at an unsampled location and convert the point coverage to grid format. Spatial data are inherently auto-correlated i.e. the similarity and dissimilarity between location are inversely correlated with distance. The inverse distance weighted (IDW) method therefore appears to be appropriate for converting the point coverage to

the raster format. The algorithm is of the form: $\hat{z}(x_0) = \frac{\sum_{i=1}^{N} z(x_i) d_{ij}^{-r}}{\sum_{i=1}^{N} d_{ij}^{-r}}$; where 'z' is the sample values, 'd' is the distance between the sample values and the location for which we wish to have an estimate, and r is an exponent associated with linear, quadratic, and cubic, functions of the distance. Besides the distance relationship that is depicted by IDW, kriging, on the other hand, uses also structural relationship among samples that was modeled by variogram function. variogram model. The general kriging model is of the form:

$z(x_0) = m(x) + \gamma(h) + \varepsilon$; where m(x) is deterministic function, $\gamma(h)$ is spatially correlated relationship depicted by variogram and $\varepsilon$ is random error.

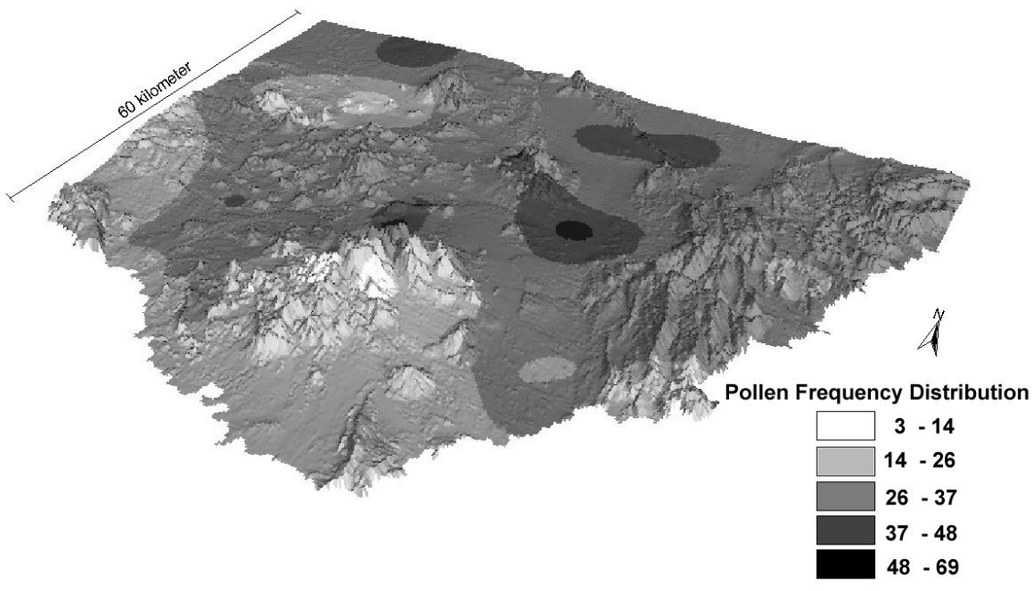
The last frequently used method of interpolation is splines. The splines are piece-wise functions that are fitted accurately to small number of data. The bicubic splines are usually used for surface interpolation. Typical effect of splines is smoothing. They produce smooth surfaces therefore the loss of accuracy to the smoothing effect was evaluated in this project.

## 5. RESULTS

The cross-validation process yielded a set of error data that represented the difference between the observed and predicted values using the three interpolation methods listed above. The frequency distribution of errors indicated that the kriging errors were skewed twice more than the errors that emanated from the spline and IDW interpolation method. The observed values of composite pollen were nearly normally distributed with 0.38 skew value (IDW errors had 0.37 and spline errors 0.33). Hence, the investigation indicates that kriging in GIS does not preserve the original distribution as do two other interpolation methods. More information is listed in Table 1.

Another important parameter elucidating the performance of interpolation procedure is the error variance. In general, increasing the variance of errors decreases confidence in predicted values and consequently increases uncertainty about the content of the raster data file. In addition, this uncertainty will affect subsequent analysis in GIS particularly in an overlay operation.

FIGURE 3.
IDW INTERPOLATION OF COMPOSITE POLLEN



The results of this work indicate that the spline interpolation yielded the highest error variance whereas the kriging and IDW error variance were significantly lower, kriging being the lowest (Table 1).

TABLE 1. SUMMARY

| Method | RMSE | MAVE | VE | VAVE |
|---|---|---|---|---|
| IDW | 12.38 | 9.42 | 156.3 | 66.14 |
| Kriging | 12.37 | 9.11 | 157.7 | 71.8 |
| Splines | 16.53 | 12.19 | 280.3 | 127.74 |

MAVE - mean absolute value of errors, VE – Variance of errors, VAVE – Variance of absolute value of errors.

The magnitude of errors is usually a significant parameter frequently used in evaluating the predictive performance of spatial processes. By definition, the sum of errors is equal to zero, so the absolute value of errors can be used to evaluate the accuracy of the interpolation process. The results showed (Table 1.) that kriging interpolation in GIS indicated the least absolute error 9.11 (the smallest variance of errors was mentioned earlier). The inverse distance method in linear form, however, indicated a value close to that of kriging's (9.42). In addition, IDW interpolation showed the smallest variance of absolute errors. Spline interpolation, on the other hand, indicated the highest absolute error - 12.19. However, it should be recognized that the distribution of errors was very similar. All three interpolation methods under or overestimated observed values in a similar pattern, and the strongest relationship depicted by correlation coefficients was between the errors arising from kriging and IDW. The Root mean square error (RMSE) is frequently used in evaluating errors in remote sensing, GIS, and mapping. RMSE is defined as the square root of an average squared difference between the observed and predicted values: $RMSE = \sqrt{SSE_i^2 / n}$ ; where SSE is sum of errors (observed – estimated values) and n – is the number of pairs (errors). Analogous to the absolute magnitude of errors, the RMSE was smallest in kriging interpolation (12.37). Similarly, IDW indicated a small value of 12.38, and spline interpolation inflated errors the most at 16.53. Careful observation of the variogram for each predicted point in the data set during cross-validation procedure indicated, for the most part, high levels of randomness in the spatial distribution of composite pollen values. However, in the northwest area the nugget effect was smaller and the range greater, indicating spatial dependency and therefore kriging performed slightly better than the IDW method. The additional set of statistical parameters was selected to evaluate performance power of each interpolation method. The results are summarized in Table 2. Covariance values were calculated to determine the strength of relationship between the selected parameters of interpolation. As the table indicates the spline interpolation showed the strongest relationship between the observed and predicted values. On the other hand the spline interpolation indicated also the strongest relationship between the estimated (predicted) values and the errors. The results are summarized in Table 2.

TABLE 2.
COVARIANCE VALUES FOR SELECTED PARAMETERS

| Covariance values | IDW | Kriging | Splines |
|---|---|---|---|
| Observed vs Predicted | 7.88 | 4.46 | 12.84 |
| Observed vs Errors | 128.25 | 131.67 | 123.29 |
| Observed vs. Absolute value of Errors | 16.39 | 25.4 | 35.84 |
| Estimated vs. Errors | -25.05 | -20.96 | -149.79 |
| Estimated vs Absolute Value of Errors | 2.2 | -7.7 | 8.9 |

In order to determine the significance of the differences between interpolation methods analysis of variance (ANOVA) was performed. The results shown in Tables 3 and 4. indicated that, at alpha level 0.05 and also 0.1, the differences between the mean absolute errors were not significant since the test failed to reject the zero hypothesis that $\mu_1 = \mu_2 = \mu_3 = 0$ .The Tukey grouping (Table 4) confirmed that there was no significant difference between the mean absolute errors that emanated from all the three interpolation methods used. Hence, statistically all methods performed similarly with no significant differences. In a more rigorous sense, however, when the alpha level was greater than 0.1, the kriging and IDW method performed significantly better than the spline interpolation in GIS.

TABLE 3. RESULTS FROM ANOVA

| Source | Df | SS | MS | F-value | Pr > F |
|---|---|---|---|---|---|
| Model | 2 | 374390.01 | 187195.01 | 1.66 | 0.1948 |
| Error | 114 | 12860322.18 | 112809.84 | | |

## 6. CONCLUSION

Understanding errors and their propagation during data manipulation and processing is becoming one of the major issues in spatial analysis. Raster data formats allow a continuous representation of reality, and interpolation methods are frequently used to predict values in unsampled locations. During this process, the point data set or vector coverages are transformed into a raster format that supports analysis of spatial patterns that are necessary for understanding the spatial relationships in geography and natural resources (Figure 3). Such procedures are highly complex and their accuracy depends on the selection of proper interpolation method.

This study applied three interpolation methods: kriging, IDW, and splines in a GIS environment to extremely biased data sets. The difficulties were associated with a biased sampling strategy, lack of data, and lack of spatial dependence (i.e. a high level of randomness). The results indicated that the inverse distance weighting method competed well with two mathematically more sophisticated methods in almost all critical parameters such as mean absolute error, RMSE, and variance of errors. Due to the complex mathematical nature of kriging and spline methods the IDW is recommended for interpolation of highly biased data for mapping purposes. Its implementation is simple, easy to understand and, as it was shown in this paper, is as accurate as other frequently used interpolation methods. The study suggests that the interpolation process and consequently the conversion of point data to raster format can be highly erroneous even though the visual output in GIS is attractive (Figure 3). This was corroborated by the fact that the relationship between observed and predicted values was extremely low and high between observed values and interpolation errors.

TABLE 4. TUKEY'S STUDENTIZED RANGE TEST

| Tukey Group | Mean | N | Class |
|---|---|---|---|
| A | 273.15 | 39 | Spline |
| A | 153.32 | 39 | IDW |
| A | 152.98 | 39 | kriging |

## 7. REFERENCES


1. Bancroft, A. B., and G. Hobbs. 1986. Distribution of Kriging Error and Stationarity of the Variogram in a Coal Property. Mathematical Geology 8(7): 635-651

2. Goodchild, M., F. Guoqing, and Y. Shiren. 1992. Development and Test of an Error Model for Categorical Data. International Journal of Geographic Information Systems 6(2): 87-104.

3. Heuvelink, G. B. M., and P. Burrough 1989. Propagation of Errors in Spatial Modeling with GIS. International Journal of Geographical Information Systems 3(4): 303-322.

4. Siska, P. P., and R. C. Maggio. 1997. The Role of Relief Dissectivity in Distribution of Kriging Errors from Digital Elevation Modeling. In: Papers and Proceedings of the Applied Geography Conferences, Vol. 20, ed. F. A. Schoolmaster, 186-194.

5. Siska, P. P., and I. K. Hung. 2000. Data Quality in Applied Spatial Analysis. In: Papers and Proceedings of the Applied Geography Conferences, Vol. 23. ed. F. A. Schoolmaster, 199-205.

6. Veregin, H. 1994. Error Modeling for the Map Overlay Operation. In: The Accuracy of Spatial Databases. Ed. M. F. Goodchild, 3-18. London: Taylor and Francis.

7. Veregin, H. 1995. developing and Testing of an Error Propagation Model for GIS Overlay Operations. International Journal of Geographical Information Systems 9(6): 595-619.