Influence of Correlation on the Quality of Area Computation

Gerhard Navratil, Claudia Achatschitz
TU Vienna, Institute of Geoinformation and Cartography

Abstract. The computation of areas is a standard task for computational geometry. Geographic information systems (GIS) usually use a straightforward approach to represent geographic features: Coordinates define locations of points, points are connected by lines which then form polygons, and closed polygons represent pieces of land. Thus, a GIS uses coordinates to compute areas for pieces of land. The coordinates only have limited accuracy, which is also valid for the computations based on these coordinates. Accuracy investigations usually assume that all coordinates used have been determined independent of each other. Unfortunately this is not the case. We want to investigate in this paper how dependencies can be taken into consideration. The results show that data quality measures increase if we model correlations.

1. Introduction

There is an increasing demand on quality indicators for data in geographic information systems. Users usually want to know if a data set can be used for a specific task. This indicator is called usability (Nielsen 1993). Unfortunately many different parameters influence usability. Statistic measures are often used for attributes as well as geometry to specify quality. Well known statistic measures are standard deviation or error probability (Stoyan 1993).

Data in geographic information systems (GIS) are always based on observations because we want to model the world and observations are the only method to get data about the real world (Frank 2001). In some, very rare cases we directly store the observations, for example if observing air pollution at a specific point. In most cases, however, we only store derived values. We use angles and distances to determine coordinates, we aggregate or classify data, or we compute a value from a variety of parameters because we cannot observe the value we are interested in (e.g., value of a specific piece of land).

All types of observations have limited quality. Modern distance measurement equipment such as used in surveying instruments, for example, has quality parameters specified in the manuals. There are usually two parameters defining a linear connection between the observed distance and the standard deviation of the resulting value:

$$s_s(d) = s_0 + d \cdot s_1 \qquad (1)$$

The limited quality of the original observations influences the quality of the derived parameters. We usually assume that a small number of quality parameters can describe the quality of a set of derived parameters.

A standard task for producing derived parameters in GIS is the computation of areas for pieces of land. GIS typically use one of two types of representations: raster representation or vector representation (Molenaar 1995). The computation of areas for raster representation is simple counting of all elements representing the piece of land. The process is more difficult if performed using a vector representation. The most common method to compute areas assumes that the boundary points are connected by straight lines. This allows using the formulae of C.F. Gauß. The two versions are the trapezoid formula

$$2F = \sum (y_{i+1} - y_i)(x_i + x_{i+1}) = \sum (x_{i+1} - x_i)(y_i + y_{i+1}) \qquad (2)$$

and the triangular formula

$$2F = \sum y_i (x_{i-1} - x_{i+1}) = \sum x_i (y_{i-1} - y_{i+1}) \quad \text{(Kahmen 1993).} \qquad (3)$$

There has already been discussion by authors on the effects of standard deviation when applying these formulae (Schwenkel 1990; Ghilani 2000; Niemeier 2002). The usual approach starts with independent (uncorrelated) coordinates and determines the quality of area based on that assumption. The idea of independent coordinates, however, does not lead to the most economic decisions because at least the

coordinates of a point are dependent since they have been computed together using the same observations.

We want to discuss how to model dependencies between coordinates. Using a test data set we also show the differences of the influence of the assumption of independent and dependent coordinates. Finally we also discuss the consequences of our results for future development.

2. Mathematical Background

Observations are processes influences by a wide range of random effects and produce slightly different results if repeated. Such processes are called randomized processes in statistics and the resulting values are randomized values. A set of values (e.g., the range of numbers an AD-converter can produce) limits the possible result of the process. Each repetition of the process provides a realization of the randomized value. Infinite repetition of the process would provide probabilities for each possible resulting value. The pattern of the probabilities is called distribution of the process. A specific kind of distribution is the normal distribution. The process follows normal distribution if it is only affected by random influences. Normal distribution is defined by mean value μ and standard deviation σ or variance σ^2 (Stoyan 1993).

Applying a function to randomized values produces a result which is also a randomized value. Each value then influences the standard deviation of the result based on its own standard deviation and the functional connection between value and result. The simple method assumes that the parameters of the function are independent of each other. A more sophisticated method also models the influence of dependencies.

Simple models assume that the parameters of the function are independent of each other. This will not be true for a variety of applications. Dependencies emerge from random or systematic influences which affect different parameters in a similar way. Distance observations with electro-magnetic waves, for example, are influenced by a wide variety of parameters (like temperature, pressure, or air turbulence). Observations taken in rapid succession from the same instrument position will be influenced in a similar way and thus statistically the distribution of the observations will be similar. Linear, random dependencies are called correlations (Reißmann 1976, p. 175).

2.1 Simple Error Propagation

The error propagation law (Reißmann 1976, p. 28) models the standard deviation of the function result, if the parameters of the function are randomized values and follow normal distribution. The error propagation law defines the standard deviation for a given function σ_f and its parameters l_1 to l_n with the standard deviations σ_1 to σ_n as follows.

$$F = F(L_1, L_2, \ldots, L_n)$$

$$s_F = \sqrt{\left(\frac{\partial F}{\partial L_1} s_1\right)^2 + \left(\frac{\partial F}{\partial L_2} s_2\right)^2 + \ldots + \left(\frac{\partial F}{\partial L_n} s_n\right)^2} \qquad (4)$$

2.2 Error Propagation for Correlated Parameters

A generalization of the error propagation law, the covariance propagation law, can deal with correlations. The formula is (Reißmann 1976, p. 177)

$$s_F^2 = \mathbf{f}^T \boldsymbol{\Sigma}_{\mathbf{xx}} \mathbf{f} \; . \qquad (5)$$

The matrix Σ_{xx} holds holding variances and co-variances for the parameters of the function and the vector f contains the first derivatives of the function. The connection between the correlation ρ_{xy}, that variances σ_x and σ_y, and co-variance σ_{xy} is defined as

$$r_{xy} = \frac{s_{xy}}{s_x s_y}. \tag{6}$$

The variances of the parameters are in the principal diagonal and the co-variances between a parameter l_a and l_b can be found in row a, column b. Usually we assume that the co-variance between l_a and l_b is the same as between l_b and l_a. Therefore, Σ_{xx} is usually symmetric matrix.

3. Error Propagation for Area Computation

The computation of the area usually uses either formula (2) or formula (3). Since the difference between (2) and (3) is the sorting order only, it does not matter which formula we use. In this paper we assume the use of the trapezoid formula (2).

3.1 Simple Error Propagation for Area Computation

The error propagation law uses the partial derivatives of the function. We differentiate formula (2) with respect to the parameters of the function to use it with formula (4). The parameters for the area computation are the coordinates of the boundary points. The partial derivatives are

$$\begin{aligned} \frac{\partial F}{\partial x_i} &= \frac{1}{2}(y_i + y_{i+1}) - (y_{i-1} + y_i) = \frac{1}{2}(y_{i+1} - y_{i-1}), \\ \frac{\partial F}{\partial y_i} &= \frac{1}{2}(x_i - x_{i+1}) + (x_{i-1} - x_i) = \frac{1}{2}(x_{i-1} - x_{i+1}). \end{aligned} \tag{7}$$

The combination of (5) with (4) then gives

$$s_F = \sqrt{\sum_i (y_{i+1} - y_{i-1})^2 s_{x_i}^2 + (x_{i-1} - x_{i+1})^2 s_{y_i}^2}\ . \tag{8}$$

The assumption that the standard deviation of all coordinates is equal, $\sigma_x=\sigma_y=\sigma$,?leads to the known formula for the standard deviation of the area (Schwenkel 1990; Niemeier 2002):

$$s_F = \frac{s}{2}\sqrt{\sum_i (y_{i+1} - y_{i-1})^2 + (x_{i-1} - x_{i+1})^2} \tag{9}$$

3.2 Correlation between Coordinates

Correlations between coordinates require using formula (5) instead of formula (4) for the computation of the standard deviation. The partial derivatives (7) define the elements for the vector f. The principal diagonal of Σ_{xx} contains the variances of the coordinates. These values are often known. The values of the correlation, however, are most likely unknown.

The problem is therefore finding suitable correlation values. Three theoretical models including correlations seem to be possible:

1. Single Point Model: The two coordinates of a point are correlated but there is no correlation between coordinates of different points. Coordinates of a point emerge from a common process. The model assumes that the correlation between the coordinates of different points is too small to affect the result.
2. Flat Model: All coordinates of all points are correlated but the correlation between the two coordinates of a point is higher than the correlation between coordinates of different points. All points emerge from the same process (e.g., digitizing) and thus are correlated equally. Since the coordinates of a point are determined in a unique sub-process, they have a higher correlation than coordinates of different points.
3. Distance Dependant Model: All coordinates are correlated depending on the distance between the points. This model is probably the most promising model since it uses neighborhood relations to

stipulate values for the correlation. Kraus and Ludwig (1998) used this model for their study on the intersection of vector data sets.

Mathematically correct would be using the co-variance matrix as resulting from an adjustment computation. The geodetic datum (the connection between the relative geometry defined by observations and the absolute geometry specified by coordinates) used for the computation influences the co-variance matrix. The S-transformation introduced by Baarda (1981) provides the transition between different geodetic datum definitions. A method providing access to co-variance matrices are measurement-based systems (Buyong, Kuhn et al. 1991; Goodchild 1999). Unfortunately we do not have the original observations in most cases. Therefore we have to use theoretical models.

3.3 Error Propagation for Area Computation with Correlated Coordinates

The single point model is the simplest model. The co-variance matrix for this model is

$$\Sigma_{xx} = \begin{pmatrix} s_{x1}^2 & s_{x1y1} & 0 & 0 & \cdots \\ s_{x1y1} & s_{y1}^2 & 0 & 0 & \cdots \\ 0 & 0 & s_{x2}^2 & s_{x2y2} & \cdots \\ 0 & 0 & s_{x2y2} & s_{y2}^2 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}. \quad \textbf{(10)}$$

The combination of (5) and (10) leads to

$$s_F^2 = \frac{1}{2}\sum_i (y_{i+1} - y_{i-1})^2 s_x^2 + (x_{i-1} - x_{i+1})^2 s_y^2 + 2(y_{i+1} - y_{i-1})(x_{i-1} - x_{i+1}) s_{xy} . \quad \textbf{(11)}$$

All other cases can only be expressed in matrix form. The differences are only visible in the co-variance matrices. The co-variance matrix for the flat model is

$$\Sigma_{xx} = \begin{pmatrix} s_{x1}^2 & r_1 s_{x1} s_{y1} & r_2 s_{x1} s_{x2} & r_2 s_{x1} s_{y2} & \cdots \\ r_1 s_{x1} s_{y1} & s_{y1}^2 & r_2 s_{y1} s_{x2} & r_2 s_{y1} s_{y2} & \cdots \\ r_2 s_{x1} s_{x2} & r_2 s_{y1} s_{x2} & s_{x2}^2 & s_{x2y2} & \cdots \\ r_2 s_{x1} s_{y2} & r 2 s_{y1} s_{y2} & s_{x2y2} & s_{y2}^2 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix} \quad \textbf{(12)}$$

with the correlations ρ_1 between the coordinates of a point and ρ_2 between the coordinates of different points.

The distance dependant model requires a co-variance function as proposed by Gauß or Hirvonen.

$$\text{Gauß: } C(d) = C(0) \cdot e^{-k^2 d^2} \quad \textbf{(13)}$$

$$\text{Hirvonen: } C(d) = \frac{C(0)}{\left(1 + A^2 d^2\right)^p} \quad \textbf{(14)}$$

In both formulae the parameter d is the distance between the points and C(0) is the maximum value the co-variance can have. The parameters A and p (Hirvonen) respectively k (Gauß) determine the behavior of the co-variance with increasing distance.

4. Test Results

A simple test can prove that the standard deviation of an area depends on the shape. A computation of areas and standard deviations for rectangles with different height-width ratios but the same size proves this point (see Table 1).

Table 1. Standard deviation for the areas of rectangles with standard deviation for the coordinates of 10cm.

b [m]	h [m]	F [m^2]	σ_F [m^2]	σ_F/F [%]
1,00	1,00	1,0	0,141	14,1
1,50	0,67	1,0	0,164	16,4
2,00	0,50	1,0	0,206	20,6
3,00	0,33	1,0	0,302	30,2
4,00	0,25	1,0	0,401	40,1
5,00	0,20	1,0	0,500	50,0
10,00	0,10	1,0	1,000	100,0

A test with a cadastral data set (a small village with the surrounding area of approximately 7km^2 in eastern Austria) from Austria with over 1.200 parcels showed standard deviations up to 103m^2 or 13,3% of the area (Navratil 2003). However, over 900 parcels (75%) have a standard deviation of less that 10m^2. The mean value for the standard deviation is 11,5m^2 and the median is 4,7m^2. The influence of shape on the standard deviation was also clearly visible. Table 2 shows mean and maximum standard deviation for different types of parcels. Street parcels have an unfavorable ration between length and width. Thus streets have the highest value for the mean relative standard deviation. Parcels for buildings, agricultural areas, and forests have a better shape resulting in better values for the mean standard deviation. The major difference between these types of parcels is the area because parcels for buildings are usually smaller than the other ones. Some parcels for buildings are extremely small (less that 50m^2) and have a bad shape, which is the reason for the high maximum of relative standard deviation.

Table 2. Relative standard deviation for different types of parcels from a cadastral data set. No correlations included.

Type of parcel	Mean [%]	Max [%]
Street	1,7	6,0
Building	0,9	9,0
Agriculture/ Forest	0,5	5,0

Introduction of correlation between the coordinates of a point does not change the distribution significantly (see Table 3). The correlation was set to 0,7. The value was selected after considering the analysis of a few results of adjustment computations and the values for correlation between the coordinates of a point found in these computations. The standard deviation drops slightly. The maximum value for the standard deviation, for example drops from 103m^2 to 99m^2. The mean value for the standard deviation drops to 9,5m^2. Now almost 80% of the parcels have a relative standard deviation of less than 1%.

Table 3. Relative standard deviation for different types of parcels from a cadastral data set using the single point model for the correlations.

Type of parcel	Mean [%]	Max [%]
Street	1,5	6,5
Building	0,8	11,9
Agriculture/ Forest	0,4	3,0

The second model includes correlation between coordinates of different points. In the test we used the values 0,7 for coordinates of the same point and 0,5 for coordinates of different points. The result shows a significant reduction of the mean standard deviation from 11,5m^2 in the uncorrelated case to 6,5m^2. Also the maximum value for the standard deviation drops to 75,8m^2. This behavior is as expected because correlation models similar behavior for the coordinates. Thus the coordinates will move in similar ways if we repeat the process of defining the points. The similarity of the behavior will reduce the changes in the geometry of the parcels. The position of the parcels will change but the shape will not be affected as much as in the case of uncorrelated coordinates where each coordinate moves independent of the other coordinates. A derived value (in our case the area), which depends on the shape but not on the position will therefore change less if the coordinates are correlated and thus the standard deviation will drop.

Table 4. Relative standard deviation for different types of parcels from a cadastral data set using the flat model with correlation of 5,0 between all coordinates and 0,7 between coordinates of a point.

Type of parcel	Mean [%]	Max [%]
Street	1,0	4,8
Building	0,6	9,0
Agriculture/ Forest	0,3	1,4

The third model includes a co-variance function. The numbers have been computed with the model of Hirvonen and the parameters A = 1 and p = 1. The maximum value for the standard deviation is now 104 m^2 (the ill-shaped street parcel shown in Figure 1), which is 10,2% and the mean relative standard deviation is 0,8%.

In general the values for the standard deviation are higher than in the second model. The reason is the distribution of the correlations. Correlation in the third model is distance-dependent and therefore distant points have less correlation than points, which are close to each other. This may even lead to correlations close to zero if the distances between the points are big enough. Thus, larger parcels are influenced less by the correlation model than smaller ones.

Table 5. Relative standard deviation for different types of parcels from a cadastral data set using the distance dependant model for the correlations: Model of Hirvonen, A = 1, p = 1.

Type of parcel	Mean [%]	Max [%]
Street	1,5	6,4
Building	0,8	10,2
Agriculture/ Forest	0,4	2,6

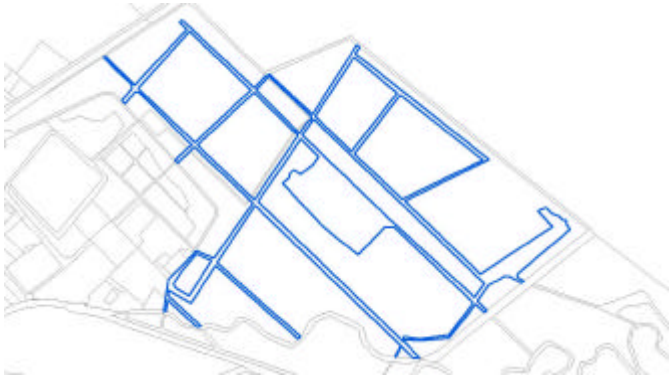

Fig. 1. Ill-shaped street parcel.

5. Conclusions

The analysis of the accuracy of the area computation showed that the result in heavily influenced by correlation. In general the accuracy is better than what results from the assumption of uncorrelated coordinates. Our test data set produced results which were 10 to 40% better than the results with uncorrelated coordinates. This fact should be taken into consideration if specifying quality parameters for points which should be used for area computations. The European Commission, for example, requires area for tasks like the support of agriculture. In such cases the European Commission specifies relative accuracy, which must be guaranteed for the areas. The quality demands for the observations will be higher if we disregard the correlation and this will result in higher costs. Thus inclusion of correlation can save costs.

The selection of the most suitable model is difficult. The second model with fixed correlation between all coordinates of the data set could be the best solution for small data sets from photogrammetric evaluations. In larger areas correlation could be distance-dependent. The problem in the distance-dependent model is the selection of the co-variance function. There are several different functions and each of these functions requires parameters. Which function provides best (most realistic) results must be the topic for further investigation.

Acknowledgements

This work was supported by the project ReviGIS (Revision of the Uncertain Geographic Information) financed by the European Commission.

References

Baarda, W. (1981). S-transformations and criterion matrices. Delft, Netherlands Geodetic Commission.

Buyong, T., W. Kuhn, et al. (1991). "A Conceptual Model of Measurement-Based Multipurpose Cadastral Systems." Journal of the Urban and Regional Information Systems Association (URISA) **3**(2): 35-49.

Frank, A. U. (2001). "Tiers of ontology and consistency constraints in geographic information systems." International Journal of Geographical Information Science **75**(5 (Special Issue on Ontology of Geographic Information)): 667-678.

Ghilani, C. (2000). "Demystifying Area Uncertainty: More or Less." Surveying and Land Information Systems **60**(3): 183 - 189.

Goodchild, M. F. (1999). Measurement-Based GIS. International Symposium on Spatial Data Quality, Hong Kong, Department of Land Surveying and Geo-Informatics, Hong Kong Polytechnic University.

Kahmen, H. (1993). Vermessungskunde. Berlin, de Gruyter.

Kraus, K. and M. Ludwig (1998). "Genauigkeit der Verschneidung geometrischer Geodaten." Zeitschrift für Vermessungswesen **123/3**: 81-87.

Molenaar, M. (1995). Spatial Concepts as Implemented in GIS. Geographic Information Systems - Materials for a Post-Graduate Course: Vol.1 - Spatial Information. A. U. Frank, Dept. of Geoinformation, Technical University Vienna**:** 91-154.

Navratil, G. (2003). Precision of Area Computation. ESRI 2003 - 18. European User Conference, Innsbruck, Austria.

Nielsen, J. (1993). Usability Engineering. Cambridge, Mass., AP Professional.

Niemeier, W. (2002). Ausgleichungsrechnung: Eine Einführung für Studierenden und Praktiker des Vermessungs- und Geoinformationswesens. Berlin, de Gruyter.

Reißmann, G. (1976). Die Ausgleichungsrechnung. Berlin, VEB Verlag für Bauwesen.

Schwenkel, D. (1990). "Genauigkeit digitalisierter Flächen." Allgemeine Vermessungs-Nachrichten **6**: 220-224.

Stoyan, D. (1993). Stochastik für Ingenieure und Naturwissenschaftler. Berlin, Akademie Verlag.