A New Approach to Assess the Positional Accuracy of Maps Generated by GIS Overlay Operations

LIU Wenbao XIA Zongguo DAI Honglei

1 Introduction

Issues on the quality of maps derived from overlay operations have been widely discussed over the past thirty years\textsuperscript{[1-3]}. Some techniques for handling alternative versions of the same cartographic feature (SCF) on different data layers were proposed by Peucker (1976), White (1978), and Dougenik (1980) (see References\textsuperscript{[4-6]}). These techniques have been used to reduce the incidence and number of spurious polygons on the composite maps. However, most of the existing techniques deal primarily with the accuracy, errors and uncertainties of overlaying original maps of the same scale. When a digitized map at 1:2 000 scale is overlaid with another map at 1:5 000 scale during GIS analysis, it is difficult to know the precise level of accuracy of the derived product. Even though the final map can be plotted at a very specific scale, we have rarely paid attention to whether the accuracy of the derived map actually meets the accuracy standards defined for maps at that particular scale\textsuperscript{[7-9]}.

2 General techniques

2.1 Assessing the accuracy of a derived composite map

The quality of source data and derived products is of paramount importance to the reliability of decisions based on the integrated analysis of multiple spatial variables. Some effective methods have been developed to test the accuracy of the original data layers commonly used in GIS analysis. The most widely used method is statistical sampling. First, a set of
sample points are carefully selected from the map products using some sampling schemes such as random, systematic, or cluster sampling. The position and attribute at each of the sample points are verified by field survey or using measurements from sources of known and higher accuracy such as large-scale aerial photographs or maps. Then the results are used to calculate some statistical measures such as the root, mean, square, mean variance, and standard deviation. Obviously, these techniques can also be used in assessing the positional accuracy of maps derived from integrating multiple data layers of various scales. However, these methods are generally costly and time consuming. The results of accuracy assessment also vary with the number of sample points and the adopted sampling scheme. In this paper, we propose a new approach for evaluating the positional accuracy of spatial objects in composite maps derived from overlay operations. The approach involves the following six steps:

Step 1: Classifying geographic features on the derived composite maps. The process is illustrated in Fig. 1.

![Fig. 1 Taxonomy of geographic features shown in composite maps generated by overlay operations](image)

Step 2: Choosing spatial objects for accuracy assessment. It can be seen from Fig. 1 that the accuracy of four different types of spatial objects needs to be assessed: (1) $E_2$: points, lines and polygons only shown on the map of one particular scale; (2) $E_{11}$: the same point features; (3) $F_1$: the line and polygon features defined by the same points; and (4) $F_2$: the lines and polygons defined by non-identical points representing natural features. Obviously, there is also an accuracy issue with attributes of spatial objects.

Step 3: Accuracy assessment for $C_2$. The techniques previously proposed by Burrough (1986), Chrisman (1987), and Goodchild (1995) can be used for assessing the accuracy of attributes of the spatial objects shown on composite maps$^{[1,10,11]}$.

Step 4: Accuracy assessment for $E_2$. The accuracy of geographic features of that type is identical to the accuracy of the original map layer displaying those features.

Step 5: Accuracy assessment for $E_{11}$ and $F_1$. The method presented in this paper is designed specifically for handling these cases. Our approach attempts to estimate the variance of digital representations of points on the original data layers and derived composite maps.

Step 6: Accuracy assessment for $F_2$. In this case, a more general approach is to fit a new line on the final map by the least squares or maximum likelihood methods with different weights for spatial objects shown on source maps of different scales$^{[10]}$. The variance of fittest points of the new line in the final output can also be easily obtained.

At Steps 5 and 6, the approach proposed here
can also be used to eliminate the spurious polygons.

2.2 Matrix formed by SPF from multi-scale maps

In order to properly develop positional error propagation model for SCF in map overlay operations more efficiently, we begin with a definition of digital coordinate matrix \( L \) for the same point feature (SPF). Now let \( Z_j = (X_j, Y_j), j = 1, 2, \ldots, n \) be a sequence of SPF within the area of interest, where \( n \) is the number of the SPF. Let \( z_{ij} = (x_{ij}, y_{ij}), i = 1, 2, \ldots, m; j = 1, 2, \ldots, n \) denote the digital point data series of \( Z_j = (X_j, Y_j), j = 1, 2, \ldots, n \) on \( m \) data layers of different scales in a spatial database, where \( m \) is the number of data layers. Then

\[
L = \begin{bmatrix}
x_{11} & y_{11} & x_{12} & y_{12} & \cdots & x_{1n} & y_{1n} \\
x_{21} & y_{21} & x_{22} & y_{22} & \cdots & x_{2n} & y_{2n} \\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
x_{m1} & y_{m1} & x_{m2} & y_{m2} & \cdots & x_{mn} & y_{mn}
\end{bmatrix}
\]

is defined as the digital coordinate matrix of SPF. In this matrix, \( x_{ij}, y_{ij} \) represent the coordinates of \( j \)th SPF, \( X_j, Y_j \) on \( i \)th data layer.

2.3 The model of positional error propagation of SPF in map overlay operations

2.3.1 First case: overlay of two maps \((m = 2)\)

For \( m = 2 \), we can obtain the following simple digital coordinate matrix from Eq. (1),

\[
L_2 = \begin{bmatrix}
x_{11}y_{11}x_{12}y_{12}\cdots x_{1n}y_{1n} \\
x_{21}y_{21}x_{22}y_{22}\cdots x_{2n}y_{2n} \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\
x_{m1}y_{m1}x_{m2}y_{m2}\cdots x_{mn}y_{mn}
\end{bmatrix}
\]

Now suppose the elements in \( L_2, X_j, Y_j (j = 1, 2, \cdots, \)\( n \) ), are independent random variables with varying positional accuracy, and let

\[
P_2 = \begin{bmatrix}
p_{x_{11}}p_{y_{11}}p_{x_{12}}p_{y_{12}}\cdots p_{x_{1n}}p_{y_{1n}} \\
p_{x_{21}}p_{y_{21}}p_{x_{22}}p_{y_{22}}\cdots p_{x_{2n}}p_{y_{2n}} \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\
p_{x_{m1}}p_{y_{m1}}p_{x_{m2}}p_{y_{m2}}\cdots p_{x_{mn}}p_{y_{mn}}
\end{bmatrix}
\]

be a weight matrix of \( L_2 \). From Eq. (2), we can write

\[
\begin{align*}
\Delta x_{ij} &= x_{ij} - x_{ij} \\
\Delta y_{ij} &= y_{ij} - y_{ij} \\
\end{align*}
\]

where \( \Delta x_{ij}, \Delta y_{ij} \) are the differences between the representations of the SPF in two separate layers. Since \((X_{2j} - X_{1j}) \) and \((Y_{2j} - Y_{1j}) \) \((j = 1, 2, \cdots, n)\) equal to zero for accurate representations of the SPF, by expanding the right side of Eq. (4), we have

\[
\begin{align*}
\Delta x_{2j} &= x_{2j} - x_{1j} = (x_{2j} - x_{1j}) - (X_{2j} - X_{1j}) \\
\Delta y_{2j} &= y_{2j} - y_{1j} = (y_{2j} - y_{1j}) - (Y_{2j} - Y_{1j}) \\
\end{align*}
\]

By applying propagation law of weight reciprocal \([12]\) on \( \Delta x_{2j} \) and \( \Delta y_{2j} \) in Eq. (5), we can obtain

\[
\begin{align*}
\rho_{\Delta x_{2j}} &= \frac{\rho_{x_{1j}}\rho_{x_{2j}}}{\rho_{x_{1j}} + \rho_{x_{2j}}} \\
\rho_{\Delta y_{2j}} &= \frac{\rho_{y_{1j}}\rho_{y_{2j}}}{\rho_{y_{1j}} + \rho_{y_{2j}}} \\
\end{align*}
\]

(6)

Thus, according to the definition of variance \([12]\), the estimation \( \delta \) of the variance factor can be derived using the true errors of selected samples as follows:

\[
\delta_{20}^2 = \frac{1}{2n} \sum_{j=1}^{n} (\rho_{x_{1j}}\Delta x_{2j}^2 + \rho_{y_{1j}}\Delta y_{2j}^2) \\
\]

(7)

Then the variance of digital data in matrix \( L_2 \) can be written as

\[
\begin{align*}
\delta_{x_{2j}}^2 &= \delta_{20}^2 \rho_{x_{1j}}^{-1} \\
\delta_{y_{2j}}^2 &= \delta_{20}^2 \rho_{y_{1j}}^{-1} \\
\end{align*}
\]

(8)

As a result, the estimated value and variance of SPF on the derived composite map can be obtained as follows:

\[
\begin{align*}
\hat{x}_{2j} &= \left( \sum_{i=1}^{2} \rho_{x_{ij}}\hat{x}_{ij} \right) / \sum_{i=1}^{2} \rho_{x_{ij}} \\
\hat{y}_{2j} &= \left( \sum_{i=1}^{2} \rho_{y_{ij}}\hat{y}_{ij} \right) / \sum_{i=1}^{2} \rho_{y_{ij}} \\
\end{align*}
\]

(9)

\[
\begin{align*}
\delta_{x_{2j}}^2 &= \delta_{20}^2 / \sum_{i=1}^{2} \rho_{x_{ij}} \\
\delta_{y_{2j}}^2 &= \delta_{20}^2 / \sum_{i=1}^{2} \rho_{y_{ij}} \\
\end{align*}
\]

(10)

A special case of interest is that the accuracy of the digital representations of the original data points in each map layer is the same. Let the weight of the digital representation of SPF on the first layer be \( \rho_1 \), and on the second layer be \( \rho_2 \). In this case, Eq. (6) will be simplified to

\[
\rho_{\Delta x_{2j}} = \rho_{\Delta y_{2j}} = \rho_{\Delta x_{1j}} = \frac{\rho_1 \rho_2}{\rho_1 + \rho_2} \\
\]

(11)

Eq. (7) will become

\[
\delta_{20}^2 = \frac{1}{2n}\rho_{\Delta x_{2j}}^2 \sum_{j=1}^{n} d_{2j}^2 \\
\]

(12)

where \( d_{2j}^2 = \Delta x_{2j}^2 + \Delta y_{2j}^2 \) is the squared Euclidean distance between digital representations of \( j \)th
SPF in two separate layers. Consequently, Eq. (8) can be changed to
\[
\sigma_{x_i}^2 = \sigma_{y_j}^2 = \frac{1}{2n} \sum_{j=1}^{n} d_{x_i}^2
\]
\[i = 1, 2, j = 1, 2, \ldots, n\] (13)

Similarly, Eqs. (10) and (11) will be simplified to
\[
[x_{2j} = \left( \sum_{i=1}^{n} \frac{p_i x_{ij}}{\sum_{i=1}^{n} p_i} \right) \sum_{j=1}^{n} \frac{d_{x_i}^2}{n}, \quad j = 1, 2, \ldots, n\]
\[
y_{2j} = \left( \sum_{i=1}^{n} \frac{p_i y_{ij}}{\sum_{i=1}^{n} p_i} \right) \sum_{j=1}^{n} \frac{d_{y_j}^2}{n}, \quad j = 1, 2, \ldots, n\] (14)

The estimation \(\sigma_{x_i}^2\) of the variance factor for three-layer overlay takes the following form:
\[
\sigma_{x_i}^2 = \frac{1}{4n} \sum_{j=1}^{n} d_{x_i}^2
\]
\[j = 1, 2, \ldots, n\] (16)

At the same time, Eq. (13) will become
\[
\sigma_{x_i}^2 = \sigma_{y_j}^2 = \frac{1}{4n} \sum_{j=1}^{n} d_{x_i}^2
\]
\[i = 1, 2; j = 1, 2, \ldots, n\] (17)

As a result, Eqs. (14) and (15) evolve into
\[
[x_{2j} = \frac{1}{2} \sum_{i=1}^{n} x_{ij}, \quad j = 1, 2, \ldots, n\]
\[
y_{2j} = \frac{1}{2} \sum_{i=1}^{n} y_{ij}, \quad j = 1, 2, \ldots, n\] (18)

\[
\sigma_{x_i}^2 = \sigma_{y_j}^2 = \frac{1}{8n} \sum_{j=1}^{n} d_{x_i}^2
\]
\[j = 1, 2, \ldots, n\] (19)

2.3.2 Second case: overlay of multiple map layers (\(m \geq 3\))

Now we pay our attention to the more general case in which the results obtained from the two-layer overlay will be extended to multiple-layer overlay. For the purpose of simplifying the mathematical derivation, it is assumed that the errors of digital representations of SPF for all the map layers are independent. In other words, the correlation of representation errors is not considered.

According to the Arc/Info way of overlay operation, if \(m \geq 3\) in Eq. (1), then we have the following coordinate and weight matrices:
\[
L_3 = \begin{bmatrix}
\hat{x}_{21}, \hat{y}_{21}, \hat{x}_{22}, \hat{y}_{22}, \ldots, \hat{x}_{2n}, \hat{y}_{2n} \\
\hat{x}_{31}, \hat{y}_{31}, \hat{x}_{32}, \hat{y}_{32}, \ldots, \hat{x}_{3n}, \hat{y}_{3n}
\end{bmatrix}
\] (20)

\[
P_3 = \begin{bmatrix}
p_{x_{21}}, p_{y_{21}}, p_{x_{22}}, p_{y_{22}}, \ldots, p_{x_{2n}}, p_{y_{2n}} \\
p_{x_{31}}, p_{y_{31}}, p_{x_{32}}, p_{y_{32}}, \ldots, p_{x_{3n}}, p_{y_{3n}}
\end{bmatrix}
\] (21)

The elements in the first row of these matrices are the new coordinates and weights of the points generated by the overlay operation in the first two map layers, and the elements in the second row are the coordinates and weights of the original data points in the third map layer.

The estimation \(\sigma_{x_i}^2\) of the variance factor for three-layer overlay takes the following form:
\[
\sigma_{x_i}^2 = \frac{1}{2n} \sum_{j=1}^{n} (p_{\Delta x_{ij}} \Delta x_{ij}^2 + p_{\Delta y_{ij}} \Delta y_{ij}^2)
\] (22)

where \(p_{\Delta x_{ij}} = p_{x_{ij}} + p_{x_{ij}}, \quad \Delta x_{ij} = x_{ij} - x_{ij}, \quad \Delta y_{ij} = y_{ij} - y_{ij}\), and the correlation of representation errors is not considered.

Similarly, Eqs. (8), (9) and (10) will be modified to the following forms:
\[
\sigma_{x_i}^2 = \sigma_{x_i}^2 = \frac{1}{2n} \sum_{j=1}^{n} (p_{\Delta x_{ij}} \Delta x_{ij}^2 + p_{\Delta y_{ij}} \Delta y_{ij}^2)
\] (23)

\[
x_{3j} = \left( \frac{p_{x_{21}} x_{2j} + p_{x_{31}} x_{3j}}{p_{x_{21}} + p_{x_{31}}} \right) / \left( \frac{p_{x_{21}} + p_{x_{31}}}{} \right), \quad j = 1, 2, \ldots, n\]
\[
y_{3j} = \left( \frac{p_{y_{21}} y_{2j} + p_{y_{31}} y_{3j}}{p_{y_{21}} + p_{y_{31}}} \right) / \left( \frac{p_{y_{21}} + p_{y_{31}}}{} \right), \quad j = 1, 2, \ldots, n\] (24)

\[
\sigma_{x_i}^2 = \sigma_{x_i}^2 = \frac{1}{2n} \sum_{j=1}^{n} (p_{\Delta x_{ij}} \Delta x_{ij}^2 + p_{\Delta y_{ij}} \Delta y_{ij}^2)
\] (25)

Similarly, the two-layer and three-layer models can be easily extended to \(m\)-layer overlay operations by replacing the subscripts 2 and 3 in Eqs. (23) and (25) with \((m - 1)\) and \(m\) respectively.

3 Application examples

Since this study is focused on modeling positional error propagation of digital representations of SPF, two examples are provided to show how to use our approach to assess the accuracy of point objects shown in all data layers and representing the same geographic features in reality, i.e., the Eu case in our classification shown in Fig. 1.

3.1 An example for two-layer overlay

In order to examine the efficiency of the approach proposed in this paper, the Monte Carlo simulation
data is used for the test. Now consider two data layers at scales of 1:1,000 and 1:5,000, in which there is a common polygon with 6 vertexes. The Monte Carlo simulation data is shown in Table 1. The theoretical variance factors are $\sigma_1^2 = 0.25 \text{ m}^2$ and $\sigma_2^2 = 1.25 \text{ m}^2$ for the two data layers. The weights for the two data layers can be defined as follows:\[ p_i = \frac{\sigma_i^2}{M_i}, i = 1, 2 \] (26)

where $M_i (i = 1, 2)$ is the denominator of scales of the original analog maps. Let $M_0 = M_1$ then we have $p_1 = 1$, $p_2 = 1/5$. Thus, we can obtain $p_3 = 1/6$ from Eq. (11) and $\sum_{i=1}^{M_2} d_{ij}^2 = 18.074$. From Eq. (12), we get $\sigma_{z_{ij}}^2 = 0.251$. Using Eq. (13), we have $\hat{o}_{ij} = \sigma_{x_{ij}}^2 = 0.251 \text{ m}^2$ and $\hat{o}_{ij}^2 = \sigma_{y_{ij}}^2 = 1.255 \text{ m}^2$ for the two simulated data layers. These values are very close to the theoretical factors selected from Monte Carlo simulation. From Eq. (15), we can obtain $\hat{o}_{ij}^2 = \sigma_{ij}^2 = 0.209 \text{ m}^2$ for the points generated during the overlay operation.

3.2 An example for four-layer overlay ($m = 4$)

Now consider a four-layer overlay with source maps at scales of 1:500, 1:1,000, 1:2,000 and 1:5,000, in which there is a common polygon with 6 vertexes. The simulated digital data of SPF in each layer are shown in Table 2.

Let $M_0 = M_2$, then according to Eq. (26), we have $p_1 = 2$, $p_2 = 1$, $p_3 = 1/2$, $p_4 = 1/5$. The calculated results are also shown in Table 2.

4 Conclusion

The tests involving two-layer and four-layer overlay operations have shown that our model works extremely well with simulated data sets. In addition, a general framework for assessing the accuracy of derived composite maps has been proposed by classifying geographic features on the basis of their origin, the relationship between spatial ob-

| Number of point | Theoretical values/m | Data layer 1/m | Data layer 2/m |
|-----------------|----------------------|---------------|---------------|
|                 | $x$ | $y$ | $x$ | $y$ | $x$ | $y$ | $x$ | $y$ |
| 1               | 1 000/3 | 500 | 1 732.555 | 500.500 | 1 730.937 | 498.873 |
| 2               | 1 500/3 | 1 000 | 2 597.579 | 999.485 | 2 596.958 | 1 001.101 |
| 3               | 1 500/3 | 2 000 | 2 598.587 | 2 000.512 | 2 599.224 | 2 001.154 |
| 4               | 1 000/3 | 2 500 | 1 731.558 | 2 499.513 | 1 733.178 | 2 498.886 |
| 5               | 500 | 2 000 | 866.512 | 2 000.507 | 865.893 | 1 998.855 |
| 6               | 500 | 1 000 | 865.515 | 999.508 | 867.130 | 998.885 |

Simulation variance/m²  Estimated variance/m²

| Number of point | Simulation variance/m² | Estimated variance/m² |
|-----------------|-------------------------|-----------------------|
|                 | 0.25 | 0.25 | 1.25 | 1.25 |
|                 | 0.251 | 0.251 | 1.255 | 1.255 |

| Number of point | Estimated variance on composite layer/m² |
|-----------------|----------------------------------------|
|                 | 0.209 | 0.209 |

Table 2. Coordinates of vertexes of a polygon in four map layers of different scales

| Number of point | 1 | 2 | 3 | 4 |
|-----------------|---|---|---|---|
| $x$/m | $y$/m | $x$/m | $y$/m | $x$/m | $y$/m | $x$/m | $y$/m |
| 1 | 1 150.367 | 1 999.633 | 1 150.515 | 1 999.476 | 1 150.735 | 1 999.265 | 1 151.162 | 1 999.838 |
| 2 | 1 149.635 | 2 000.358 | 1 149.489 | 2 000.517 | 1 149.276 | 2 000.727 | 1 148.838 | 2 001.162 |
| 3 | 1 200.358 | 2 049.642 | 1 200.524 | 2 049.485 | 1 200.726 | 2 049.271 | 1 201.157 | 2 049.842 |
| 4 | 1 199.644 | 2 050.365 | 1 199.487 | 2 050.521 | 1 199.276 | 2 050.728 | 1 198.843 | 2 051.158 |
| 5 | 1 250.371 | 2 099.864 | 1 250.519 | 2 099.472 | 1 250.726 | 2 099.261 | 1 251.149 | 2 098.835 |
| 6 | 1 249.628 | 2 100.369 | 1 249.477 | 2 100.513 | 1 249.275 | 2 100.729 | 1 249.851 | 2 101.160 |

Estimated variance/m²  Estimated variance on composite layer/m²

| Number of point | Estimated variance/m² | Estimated variance on composite layer/m² |
|-----------------|-----------------------|----------------------------------------|
|                 | 0.135 | 0.135 | 0.270 | 0.270 | 0.540 | 0.270 | 1.350 | 1.350 |
|                 | 0.497 | 0.513 |
jects and geographic entities, and the relationship between spatial objects in different map layers of variable scales.

References

1. Chrisman N. R. (1987) The accuracy of map overlays: a reassessment. Landscape and Urban Planning, 14(5): 427-439
2. Newcomer J. A., Szaigin J. (1984) Accumulation of thematic map errors in digital overlay analysis. The American Cartographer, 11(1): 58-62
3. Heuvelink G. B. M., Burrough P. A. (1993) Error propagation in the cartographic modeling using Boolean Logic and continuous classification. Int. J. of Geographical Information System, 7: 231-246
4. Peucker T. K. (1976) A theory of the cartographic line. International Yearbook of Cartography, 16: 134-143
5. White D. (1978) A design for polygon overlay. Dutton G. Harvard Papers on Geographic Information Systems, 6. Cambridge, MA: Laboratory for Computer Graphics and Spatial Analysis, Harvard University
6. Dougenik J. (1980) Whirlpool: A geometric processor for polygon coverage data. Auto Carto 4, 304-311
7. Goodchild M. F., Gopal S. (1989) Accuracy of Spatial Databases. New York: Taylor and Francis, XIII
8. Liu W B (1995) A theory of uncertainty of spatial data in GIS. PhD Dissertation. Wuhan: Wuhan Technical University of Surveying and Mapping (in Chinese)
9. Liu W B (1998) Models of data quality and dynamic spatial relations in GIS-T. Postdoctoral Technical Report. Nanjing: Southeast University (in Chinese)
10. Burrough P. A. (1986) Principles of geographical information systems for land resources assessment. Oxford: Clarendon
11. Goodchild M. F. (1995) Attribute accuracy. In: Guptill S. C., Morrison J. L. eds. Elements of Spatial Data Quality. Oxford: Elsevier Scientific, 59-79
12. Mikhail E. M. (1976) Observation and Least Squares. New York: IEP-Dun-Donelly
13. Binder K., Hermann D. W. (1992) Monte Carlo simulation in statistical physics. Berlin: Springer-Verlag