Error detection through consistency checking

Peng Gong* Lan Mu#
*Center for Assessment & Monitoring of Forest & Environmental Resources
Department of Environmental Science, Policy, and Management
#Geographic Information Science Center and Departments of Landscape Architecture and Environmental Planning
151 Hilgard Hall, University of California, Berkeley, Berkeley, CA 94720-3110
gong@nature.berkeley.edu mulan@gisc.berkeley.edu

Abstract Following a brief discussion on various aspects of data quality, possible methods are examined for the detection of errors in a spatial database. Using examples, we introduce the consistency checking method based on spatial relationships among neighboring objects and attribute relationships among map layers from different sources. Using logical relationships among spatial neighborhoods and among attribute data from different sources, it is desirable to build an error detection mechanism in a spatial database. This mechanism can be automated and has the potential to be one of the powerful tools for error detection and correction suggestion in a spatial database.

Introduction

Data quality can be assessed through data accuracy (or error), precision, uncertainty, compatibility, consistency, completeness, accessibility, and timeliness as recorded in the lineage data (Chen and Gong, 1998). Spatial error refers to the difference between the true value and the recorded value of non-spatial and non-temporal data in a database. Attribute error is more complicated than other types of spatial errors. It is related to scale of measurements. At one scale of measurement, the difference may be regarded as error while not at another scale. For example, an elevation of 497 m recorded in the database with its true value being 492 m will be considered erroneous at the ratio and interval scales but accurate in a general category such as an elevation class between 450 and 500 m which is at the nominal scale. However, sometimes the true value is not known, error can not be evaluated. Under such circumstances, uncertainty is used. Statistically, we use the average from multiple measurements to estimate the true value and the standard deviation of the multiple values as an indicator of the level of uncertainty. Therefore, in order to know the uncertainty of a value, multiple measurements are necessary. For example, a coastal line – the boundary between ocean and land, is uncertain as it changes constantly with time due to such factors as tides and ocean waves. There are more causes of data uncertainty than that the truth is not measurable or there is no truth at all. The conceptual fuzziness of an attribute or a category, which represents the level of data generalization, could also cause data uncertainty. For example, one can not tell what is the true density of a polygon in a database when its category is "high density residential." Similarly, one can not tell exactly which tree species are contained in a class of "evergreen broadleaf forest" due to its high level of abstraction.

Attribute error has been studied for many years. Particularly in remote sensing image classification, a relatively complete procedure exists for classification error analysis (Jensen, 1996). Chen and Gong (1998) divided the classification error analysis into 5 steps:

(1) determine the sampling method for ground truth data collection; Methods include systematic sampling, random sampling, stratified sampling and systematic unaligned sampling, etc.
(2) determine the sample size;
(3) determine the attribute of sample location; this is usually done by field survey or the use of more accurate data sources such as aerial photo interpretation.
(4) compare sample data with classification data and establish the contingency matrix;
(5) calculate various types of errors from the contingency matrix.

This can be applied to the determination of any type of attribute error at the nominal measurement scale.

Precision refers to the closeness of measurements obtained from the same object using the same measurement method. It is related to the level of details contained in the measurement. It can be assessed

by the standard deviation of a number of measurements made from the same object (Gong et al., 1995). Compatibility refers to how easy it is when data collected for other purposes can be used in a particular application. It also refers to how easy it is when data from different sources or collected from different locations are used for the same application. Generally, more specific data have better compatibility than more general data because more specific data can be generalized to general data but not vice versa. For example, there may be two forest maps for two neighboring regions but prepared with different methods, or using different data sources. If the content of one map can be made comparable with the other, then the two maps are compatible. If only the classification system needs to be adjusted to make one map to be compatible with the other, we call that one map can be "cross-walked" to the other. Consistency refers to the level of agreement when a certain phenomenon is represented in the database. For example, if the same river looks different on two types of maps, the level of consistency between the two maps is poor. If the same terrain feature from two map layers are represented by different number of contour lines and/or different levels of smoothness of the contours, the consistency between the two maps is poor. If one map is made from data collected at one time and a second map for the neighboring region is made from data collected at a different time, then the two maps may be temporally inconsistent.

There are primarily four types of errors in a GIS database: positional, temporal, attribute, and logical. Logical error refers to the inconsistency of relationship among different features presented in a database. It is usually manifested through other types of errors. Thus, logical relationships of mapped features can be checked for error detection. Positional error has been widely investigated for its determination (Gong et al, 1995; Stanislawski et al., 1996; Kiiveri, 1997; Veregin, 2000), modeling (Zheng and Gong, 1997; Shi and Liu, 2000). Essentially, positional error is the error contained in the coordinate values of points, lines and volumes. Thus, it is one type of numeric errors. Numeric error is relatively a simple type of spatial data error. Currently, few GIS systems are truly incorporating the temporal axis as an index that supports explicit query in time. When time is not explicitly used as an index like geographical coordinates, it is treated as an attribute just as elevation is treated in a 2D GIS. Thus, in most existing GISs time and elevation are treated as attributes.

Error propagation and uncertainty detection has attracted research attention for the past decade. The following table lists some of the research papers done in this field. Most of the papers deal with single variable, and among them, a lot mentioned modelling the spatial autocorrelation to estimate data uncertainty. Consistency check between variables from different sources has been introduced (Scott, 1994) which is the emphasis of this paper. Typical approaches among various research are error modeling, simulation, calculation and visualization, etc (Table 1).

Table 1. Review of some of the recent error studies.

Paper	Error Type	Problem and solution	Major Approach
Ehlschlaeger, 1996	positional inconsistency Single variable elevation	Visualization approach to view the elevation surface change by applying a nonlinear interpolation model to develop animations.	Visualization
Griffith, et al. 1994	logical inconsistency Single variable	The standard error difference between area mean and population mean caused bias in estimating population mean. Using census tract data at Syracuse, New York, added the underlying spatial autocorrelation in estimating the standard error to get population mean	Modelling

Heuvelink, 1995	logical inconsistency Single variable	Error propagation from different spatial variation model fittings Compared Discrete Model of Spatial Variation (DMSV), Continuous (CMSV), and Mixed (MMSV) models with Netherlands high groundwater level data, and suggested adopting MMSV when undetermined.	Modelling
Heuvelink, 1998	Single variable	Discussed the errors of many models used in soil science coming from not only the input and also the model itself. It discussed the error propagation process in data interpolation and aggregation as well.	Simulation
Hunter, Goodchild, 1997	logical inconsistency Single variable	Slope and aspect uncertainties from realized models Added spatial autoregressive random field as a disturbance to elevation, and propose a worst-case scenario by choosing a rho value within the domain of 0 and 0.25. "Uncertainty" includes "error"	Modelling
Kiiveri, 1995	Positional inconsistency, polygon Single variable	Took a look at the inconsistency though the polygon boundary change, length, perimeters and areas calculations after the overlay operation	Calculation , Simulation
Mowrer, 1996	Positional inconsistency Single variable	Applied Monte Carlo technique of sequential Gaussian simulation to estimate old-growth subalpine forests. Suggest using the technique with the technology of GPS and GIS to improve decision making	Simulation
Phillips, 1995	attribute inconsistency Multiple variables	Use of simulation modelling to measure uncertainties. Model potential evapotranspiration as a function of temperature, humidity, and wind.	Simulation
Phillips, 1999		Theoretical discussion of a major challenge in physical geography: the detection of the signals of complex deterministic dynamics in real landscapes and data. It introduced the nonlinear dynamical system (NDA) theory, reviewed most recent literatures, and compared approaches relevant to deterministic uncertainty.	
Scott, 1994	logical inconsistency Multiple variables	Introduced Exploratory Data Analysis (EDA) tool to help quality assessment and data integrity in GIS by using statistical techniques. It suggested four components in this issue (p384): (1)distribution checks of both categorical and ratio data; (2)logical consistency checks of the relationships between attribute data values and between attribute classes; (3) proximity checks of the spatial distribution of data attributes; and (4) plot and map reviews of the spatial distribution of geographical features and their associated attributes	Modeling, Calculation
Shi, 1999	Positional inconsistency, line	Developed G-band model to handle positional error of line segments. With end points normal distribution assumption, the model applied stochastic process to discuss the uncertainties of end points as well as points on the segments.	Modelling

Stanislawski, 1996	Positional inconsistency single variable digitized points	Estimated positional accuracy by dividing errors into absolute error and relative error, while the absolute one represents horizontal cartographic data accuracy, and the relative one represents variability in spatial relationships.	Modelling

In the rest of the paper, we use example to discuss how errors can be detected through consistency checking in spatial databases. Discussions will be made with a suggestion on the development of an spatial data inconsistency checking mechanism in spatial databases.

Error detection through consistency checking

We can divide inconsistency into spatial inconsistency, temporal inconsistency, attribute inconsistency and inconsistency among any combination of space, time and attribute. Spatial inconsistency is a process that cartographers must deal with on an operational basis. Map generalization is a major task of cartographers. It includes spatial displacement (a process of spatial error introduction), spatial simplification through selection, aggregation and smoothing, and attribute abstraction through classification. This process itself introduces a huge amount of error particularly on small scale maps. Traditional spatial analysis based on maps is restricted by map scale, as maps from different scales can not be overlaid with each other for multi-layer (variable) analysis. In a GIS system, maps of the same spatial location can be enlarged or reduced to map with each other regardless their original scales. Spatial inconsistency could occur under this circumstance.

1. An example of spatial inconsistency

Figure 1 shows a reservoir and a highway overlap with each other as a result of overlaying a drainage map with a transportation map. The highway extends at the same side of the reservoir and the stream with no reason for it to run over it. Provided that the maps have the same scale and their projections and other factors that control the geometrical properties of the two maps are consistent, it is most likely that the error is caused by some displacement of the reservoir as highways are usually surveyed with high precision. This situation could change if we over a large scale drainage map with a small scale road map. Under such circumstances, the error is most likely due to the generalization of the roads on the road map. Therefore, the ways of correcting the error, or at least removing the inconsistency if we believe the error is not correctable, vary with the actual situation. This requires the knowledge of the scale and accuracy report of each map if there exists any. The relevance of metadata of spatial databases is obvious here.

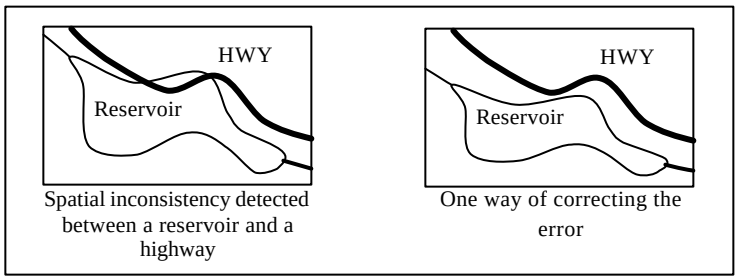

Figure 1. Spatial inconsistency found in map overlay.

2. Logical error detection of individual objects

Certain objects in a map database have logical relationships with other objects. For example, a parking lot should exit to a road. If a parking lot is by itself without any entrance or exit, then there is a logical error. Consider a bridge, it could either be across a stream, river or another road and its two sides should be connected to roads. These are the knowledge that can be coded to automatically check if there is any logical errors associated with each bridge. Such logical error detection associated with bridges is

particularly useful in detecting errors of other attributes that are connected to bridges. Figure 2 illustrates the situation for a parking lot and a bridge. This method can be applied to any type of object whose relationship with other objects can be logically expressed.

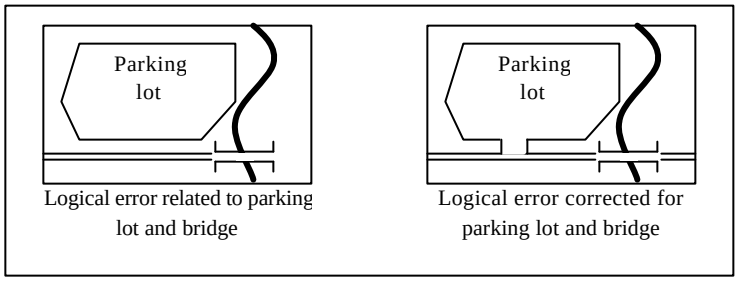

Figure 2. Logical inconsistencies associated with individual objects

3. Attribute error identification through logical consistency checking among different map layers

In a spatial database, data are often organized in different map layers. Each map layer may be obtained from different sources. Attribute error on one map layer may not be detected without being compared with attribute data from other map layers. For example, a forest fire history map contains the distribution of burnt areas with an attribute of time of fire occurrence (e.g., Figure 3a). Are there any mistakes in the fire history records? Some such errors may be detected when the fire history map is overlaid onto an up-to-date forest cover map (e.g., Figure 3b). Fire history records can be checked according to the current stage of forest restoration. In the example as shown in Figure 3, the two fire occurrence times should obviously be exchanged because they are not consistent with the age of the vegetation.

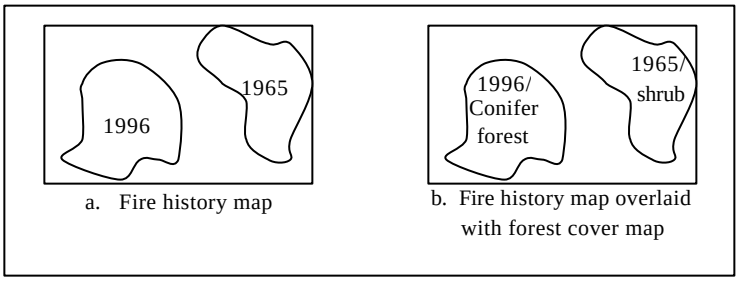

Figure 3. An example of logical inconsistency between two map layers. An old 1965 fire is now still covered by shrub but a relatively new burnt area in 1996 is covered by conifer forest. If the forest map is considered as correct, the age of the fires should be switched between the two fires.

The simple example illustrated in Figure 3 indicates that consistency checking between maps from different sources may be a useful tool in attribute error detection. In the following, we examine some real map layers of central California.

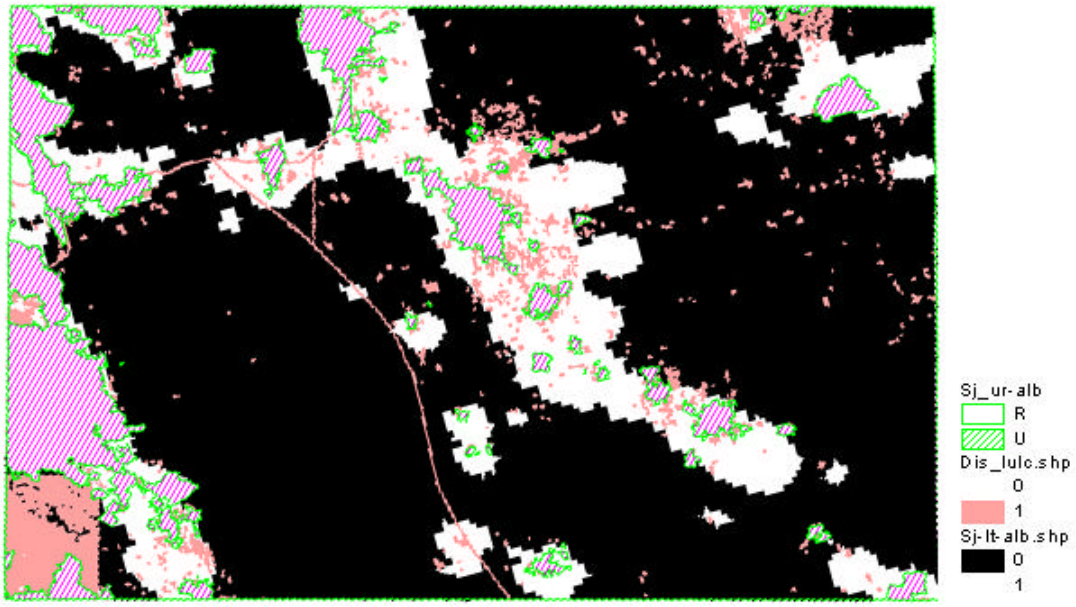

Figure 4. Urban land use from East San Francisco Bay across great central valley to part of Sierra Nevada. Data are obtained from three sources: Census Population Density data in pattern bounded in green, USGS Land Use and Land Cover data in brown, and urban area in white determined from the city light data from the Defense Meteorological Satellite Program (DMSP).

Knowing how the maps are made helps us to detect attribute errors. From Figure 4, it is obvious that the urban area determined from the DMSP city light data is largely exaggerated. This is partly caused by the poor spatial resolution of the city light data (1 km resampled from the original 600 m) and the less accurate city light intensity thresholding algorithm applied in urban area detection. It can be considered as the extreme end of over-commission of urban land in the mapped area. Almost any area not included in the urban area determined from the city light data is not likely to be urban. The urban area from the other two map sources are relatively consistent except at the lower left corner where there is a large tract of urban land only claimed by the USGS source. Therefore, before any other verification we are almost certain that it is an attribute error over that tract of land. The particular error in the USGS data layer was verified by road density and Landsat TM imagery.

Discussions

From the illustrations in the above section, it can be seen that inconsistency is a useful indicator of spatial data errors. Inconsistency may exist on a single map layer or among different map layers. Inconsistency can be detected automatically. This requires a good knowledge of various characteristics of spatial data. Inconsistency checking should be made in at least four aspects: self checking of data completeness such as various spatial, attribute components of an object represented in the database; spatial consistency among neighbors of objects; multivariable (multi-attribute) consistency through comparison; and spatial consistency among multiple variables. It is expected that the level of complexity in consistency checking increases in a similar order. Some of the corrections for spatial errors as reflected by inconsistency can be

done automatically while it is more appropriate to correct errors or reduce uncertainties through an interactive process.

Like a spelling checker in a word processing software, an inconsistency checker is envisioned that is developed for each database. It can be fired to run in the batch mode or at the background once new data are added into the database. Some detected inconsistencies are corrected according some rules and are highlighted while some others are left uncorrected. All inconsistencies should be recorded to alert data analysts for final correction decision. Some special visualization tools can be used for the purpose of inconsistency warning. A mechanism should be built for database manager and data users to track changes made to data and to allow for reverse processing should automatic correction is considered done inappropriately.

Acknowledgements

This research is partially supported by a Kearney Foundation Grant to Ron Amundson and Peng Gong at the University of California.

References

Chen, J. and P. Gong, 1998. *Practical GIS: Building and Maintaining a Successful GIS*, Science Press, Beijing, p.186.

Ehlschlaeger, C.R., A.M. Shortridge, M.F. Goodchild, 1997. Visualizing spatial data uncertainty using simulation, *Computers and Geosciences*, 23(4):387-395.

Gong, P., X. Zheng, J. Chen, 1995. Boundary uncertainties in digital maps: an experiment on digitization errors, *Geographic Information Sciences*, 1(2):65-72.

Gong, P., R. Pu, and J. Chen, 1996. Mapping ecological land systems and classification uncertainty from digital elevation and forest cover data using neural networks, *Photogrammetric Engineering and Remote Sensing*, 62(11):1249-1260.

Griffith, D.A., R. Haining, G. Arbia, 1994. Heterogeneity of attribute sampling error in spatial data sets. *Geographical Analysis*, 26(4):300-326.

Heuvelink, G.B.M., 1996. Identification of field attribute error under different models of spatial variation. *International Journal of Geographical Information Systems*, 10(8):921-935.

Hunter, G.J., M.F. Goodchild, 1997. Modeling the uncertainty of slope and aspect estimates derived from spatial databases, *Geographical Analysis*, 29(1):35-49.

Jensen, J.R., 1996. *Digital Image Processing, a Remote Sensing Perspective*, Prentice Hall, Upper Saddle River, New Jersey, 316p.

Kiiveri, H.T., 1997. Assessing, representing and transmitting positional uncertainty in maps, *International Journal of Geographical Information Science*, 11(1):33-52.

Mowrer, H.T., 1997. Propagating uncertainty through spatial processes for old-growth subalpine forests using sequential Gaussian simulation in GIS. *Ecological Modeling*, 98(1):73-86.

Philips, D.L., D.G. Marks, 1996. Spatial uncertainty analysis – propagation of interpolation errors in spatially distributed models. *Ecological Modeling*, 91(1):213-229.

Scott, L.M., 1994. Identification of GIS attribute error using exploratory data analysis. *Professional Geographer*, 46(3):378-386.

Shi, W., W.B. Liu, 2000. A stochastic process-based model for the positional error of line segments in GIS, *International Journal of Geographical Information Science*, 14(1):51-66.

Stanislawski L.V., B.A. Dewitt, R.L. Shretha, 1996. Estimating positional accuracy of data layers within a GIS through error propogation. *Photogrammetric Engineering and Remote Sensing*, 62(4):429-433.

Veregin, H., 2000. Quantifying positional error induced by line simplification. *International Journal of Geographical Information Science*, 14(2):113-130.

Zheng, X., P. Gong, 1997. Linear feature modeling with curve fitting: parametric polynomial techniques, *Geographic Information Sciences*, 3(1):7-19.