Characteristics of the positional errors of historical maps

Tomaž Podobnikar
Scientific Research Centre of the Slovenian Academy for Sciences and Arts,
Novi trg 2, SI-1000 Ljubljana, Slovenia and
Institute of Photogrammetry and Remote Sensing, Vienna University of Technology,
Gußhausstr. 27-29, A-1040 Vienna, Austria

INTRODUCTION

Data are traditionally evaluated using well-defined data models. But results may be variable if the data are not homogenised or standardised. Comparison of actual data with reference data presents a data model which supposes no error. The problem with evaluating positional quality, especially of maps, is a lack of temporal quality, therefore the reference points for evaluation should be chosen very carefully. The researched data are historical maps, where the data model and the character of quality parameters are more complex, and include subjective factors and uncertainties. The study account for space and time; consider the historians, geographers, and surveyors; and support the reconstruction of a physical and human landscape (Plewe 2002). That acquired information can stimulate the imagination and increase the understanding of possible quality parameters of data and error distribution models (Khorram et al., 1999).

Statistical techniques have traditionally been employed for quality control and error assessment, according to common standards. Visual methods are less accepted; nevertheless they can be important tools for both preliminary and subsequent evaluation (Wood and Fisher, 1993). These methods are excellent for obtaining first impressions and other insights about data sets such as historical maps. Unfortunately, they require an operator with expert knowledge to improve objectivity. This leads to a possible hypothesis of one or more error distribution models that are then simulated with Monte Carlo (MC) statistical methods. Two principles for the evaluation of projected error models and types are applied evaluating land use acquired from historical maps: boundary and surface simulation. Both principles address random, locally systematic, and systematic error distributions. The boundary error is simulated with vector lines and polygons. The surface error is simulated by producing error surfaces that shift every grid point or square-shaped polygon. Some outputs of this research were used for the Triglav national park case study region in Slovenia. There we tried to determine the quality of a range of historical maps, used for land use analysis.

POSITIONAL ERROR SIMULATION

MC simulations as statistical methods are generally used in cases where physical processes are random or the theoretical mathematics in the different hypothesis tests is weak (Openshaw et al., 1991, Brown and Duh, 2004). The MC methods of positional error analysis follow the statistical theories of error distribution and propagation (Burrough and McDonnell, 1998). Errors can be simulated to evaluate differing qualities of classification of borders between land use classes. The simulations may also follow spatial analyses that provide combinations of land use, slope, aspect, tourism resources, etc. A complex probability model can be specified for every source of error if the understanding of spatial variability components is high. Multiple and equally probable outcomes from the model are often needed because a single simulation of the random function/fields is really just one of a large number of representations of the specified probability model (Cressie, 1993, Haining, 2003).

One outcome of the random function is unconditional simulation. That yields outcomes which are consistent with the probability model but in which simulated data values do not correspond with known data values. In the case of a normal distribution, the mean value is 0 and the standard deviation is 1. Simulations that respect data are conditional. In this case they come from a normal probability model where mean and distribution are specified but at the locations where data values are known the outcomes match these. Transformation of the unconditional uniform distribution of a discrete random variable via cumulative to normal distribution can be applied by the Box-Muller method. This is a suitable approach for simulation of error in spatial data. Variables used for MC simulations of error are usually spatially autocorrelated with respect to the nature of the error. A general procedure for the MC simulation algorithm is then:

- generate a set of random numbers;
- transform n random numbers to an appropriate unconditional distribution;
- respect current data and the associated error model to compute a conditional distribution;
- repeat the previous steps N times;
- analyse and evaluate a distribution from the N outputs.

DATA: LAND USE FROM HISTORICAL MAPS

For this case study, the Josephine Ist Military Topography (JMT28.8; figure 1), named for Austrian Joseph II, was used.

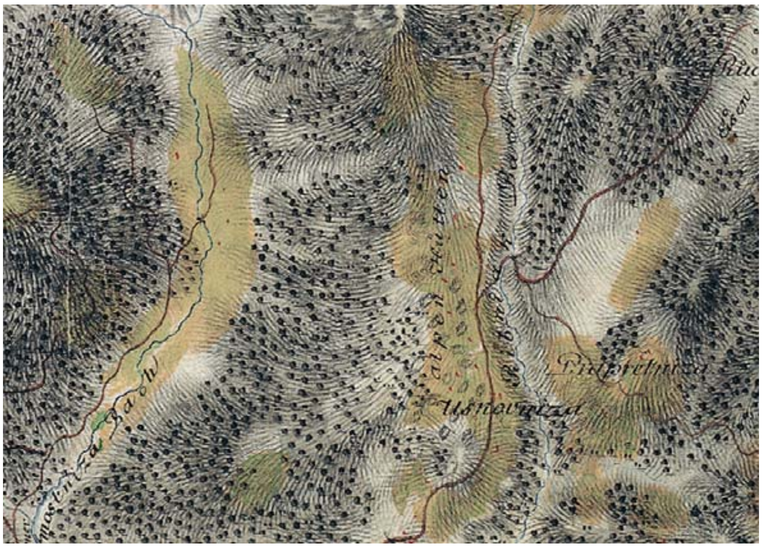

Figure 1. The insertion of the Josephine Ist Military Topography (JMT28.8).

There were some reasonable drawbacks to the historical maps JMT28.8. The old maps were found in the archives and only paper prints were available for scanning to a digital format. Therefore substantial deformations of the paper could be expected that were not easily distinguished from deformations of triangulations and other sources. Maps were made with different mapping techniques, varying object catalogues (map legends) and a lack of projection and transformation parameters. The origin of the original reference system was not available, so we used reference points.

The next problem is illustrated by typically sparsely surveyed mountain. Many details were not measured or were mapped just by eyeballing, and therefore succumbed to gross errors. Nevertheless, identical points on the JMT28.8, linked with contemporary orthophotographs, were carefully chosen considering historical knowledge of measurement techniques and possible environmental changes. Geometrically, the best points were the trigonometrical ones. These were mostly churches or towers. The older maps are, the less confidence can be placed in identical points, due to natural or anthropogenic changes in land use. We found out that was for JMT28.8 better to match identical points roughly in order to more systematically cover an entire mapped area and avoid locally large distortions.

The land use classifications of JMT28.8 were rearranged into 14 classes based on the contemporary land use data set, resembling the Corine land cover nomenclature. The majority of classes were distinguishable, though in some cases, just barely. For better categorisation, additional information was used: the geomorphology of relief elevation, aspect, proximity of urban areas, the increasing elevation of the upper vegetation level during the last decades, and the general order of vegetation zones. Land use data was acquired by a backwards editing method – map-by-map, starting with the newest vector-based land use data.

The difficulties described above greatly affected the thematical and positional accuracy of the georeferenced maps and derived land use data. They were evaluated on the digitised land use data sets despite the difficulty in some cases of distinguishing them (Lester and Chrisman, 1991). Another problem was the extrapolation of the general positional error of entire map surfaces to the boundary error in classified polygons, representing polygons of land use. This could be also considered as a thematic or attribute problem. Briefly, we can say that most of the error comes from the less precisely mapped land use features in the original maps and positionally/thematically incorrect interpretations of the boundaries between the classes. All mostly empirical tests were applied and are described in Podobnikar and Kokalj (2007).

We considered that after a sensitive georeferencing of the maps, the error distribution model was related only to positional accuracy, evaluated with RMSE (root mean square error), evaluated using different methods of transformation through georeferencing, and AME (averaged maximum error), calculated as differences between manually selected identical points on georeferenced maps and nominal data sets (contemporary orthophotos and maps), and then averaged. As areas of the maps were chosen with obviously lower expected quality, the AME was considered to be locally maximal. This measure controls RMSE numerically as well as visually by marking distances and directions (table 1). By statistical and visual evaluation of historical maps and an exploration of the methodology and principles of historical mapping, we deepened our knowledge and understanding of the overall quality of the maps and the positional error distributions of acquired land use data sets.

	JMT28.8
RMSE	16.7–19.1 mm
AME	19.8 mm
Maximal	36.5 mm

Table 1. Positional accuracy evaluation of historical data set JMT28.8.

SIMULATION PROCEDURES AND RESULTS

Due to possible errors, JMT28.8, the simulations of the land use data were performed with the following assumptions:

- random effects: instruments, human factors and others that are indefinable with respect to precision;
- locally systematic effects over larger areas that could be also classified as spatially smooth or random on a smaller scale: triangulation errors, errors in measurements of long distances;
- systematic/gross effects over the whole case study area: projections, measurements.

Assumptions for the second and the third categories of error were not so easily and uniquely simulated as those for the first category were. Systematic or even gross errors may not be distributed randomly, but their nature is sometimes similar to stochastic, and at the same time they are almost impossible to remove. In spite of the inexact and uncertain classification of error types, we attempted to simulate all three groups using statistical MC simulation as the most practical but still reasonable solution.

Technically, the MC error simulation of error models and types was applied using two different principles that were denoted as the "boundary" and "surface" principle. The proposed methods of simulation involved several common steps that were considered in both principles. The first step was to create random values applying standardised unconditional normal distributions. Those distributions were multiplied with the RMSE values of the JMT28.8 (Table 1). The RMSE was transformed to standard deviation values σ that are suitable for simulations of stochastic processes with the following simplification: RMSE = σ. We considered that the actual degree (value) of RMSE is possibly higher than σ, but the character of error was modelled consistent with previous tests. All simulations were handled with ESRI's macro language AML, but processed with GNU C.

Boundary principle

The first approach under the boundary principle followed the rules of boundary simulations (Burrough and McDonnell, 1998) applying vector lines. RMSEs measured on the JMT28.8 were assigned to the land use boundaries acquired from the JMT28.8 as σ. The average σ was 28,800 · 17.9 mm = 515 m (table 1). Even though this is high, we were not able to reduce this error, as more reference points could not be located.

Furthermore, we assumed that the shape of a particular simulated object remains similar to the acquired original. Applying that condition, neighbourhood nodes of the boundary lines should be correlated. Rather than programming the correlation between the nodes within the simulation of one data set, a simpler approach was to apply the correlation coefficient over the entire study area. The correlation coefficient CORR was applied using equation:

$$x_{\mathrm{chang}} = x_{\mathrm{orig}} + \sigma(x_1 \cdot \mathrm{CORR} + x_2\,(1 - \mathrm{CORR})),$$
$$y_{\mathrm{chang}} = y_{\mathrm{orig}} + \sigma(y_1 \cdot \mathrm{CORR} + y_2\,(1 - \mathrm{CORR})),\ 0 \le \mathrm{CORR} \le 1$$

The chosen value of CORR depends on correlation between generalisation of the vector lines and σ, or, on the other hand, on the computed ratio of absolute to relative accuracy. The numerical evaluation was confirmed by the example presented in figure 2 where the appropriate CORR is between 0.5 and 0.75. With regard to the unconditional normal distributions, x_1, y_1, were calculated once for entire data set, and x_2, y_2 separately for every node point.

The second approach under the boundary principle was analogous to the first, except that this simulation was applied to the boundaries between polygons and not to the lines themselves. Preserving the original topology on simulated polygons after positions of the nodes were changed, as much as possible requires a much more complex numerical solution. We solved the problem successfully and quite robustly, even where low correlation coefficients had been used. With CORR between 0.50 and 0.95, we also simulated some systematic effects. The simulations were repeated for 100 times and the singular data sets were rasterised to unique resolution, and then all summed up:

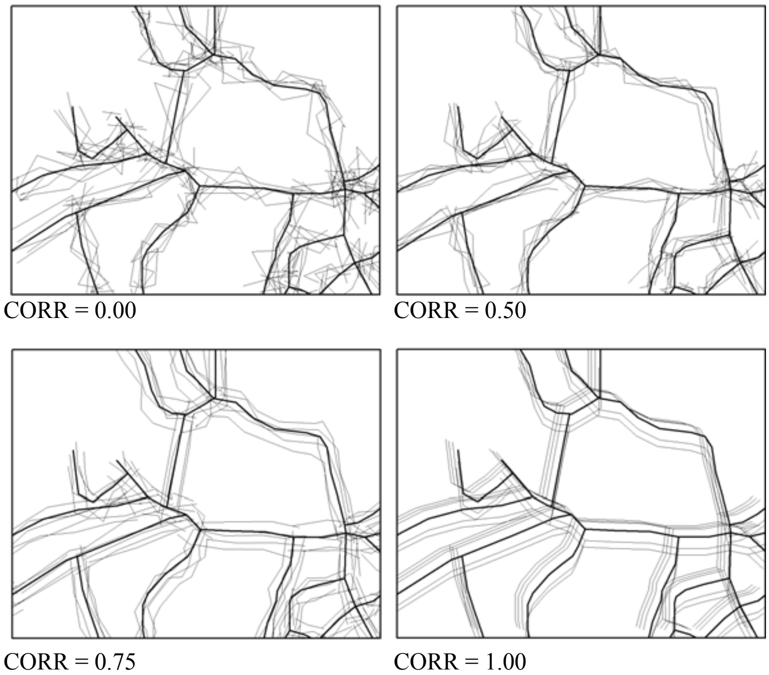

Figure 2. MC with 5 simulations of the vector lines boundaries using different correlation coefficients between neighbour nodes.

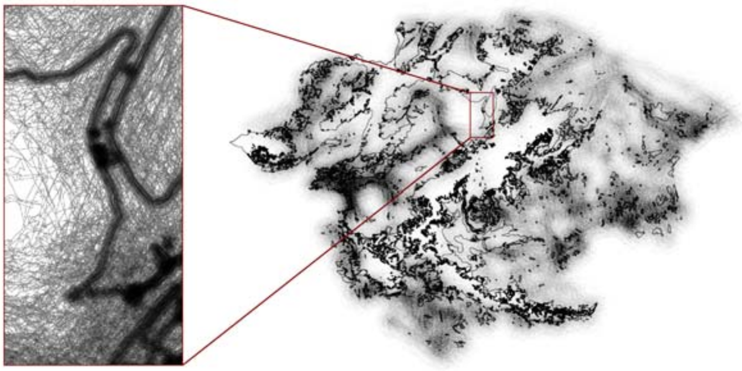

Figure 3. Land use data acquired from the JMT28.8 simulated 100 times with the MC method using the land use boundaries for the entire Triglav national park (area approx. 50 by 50 km).

- lines: by presenting five simulated data around original lines (figure 2);
- lines: fuzzy boundary as a "shadow" effect using all 100 simulated lines (figure 3); and
- polygons: fuzzy boundary by presenting 100 simulated data using different colours.

Surface principle

With continuous surfaces, the error distributions were simulated for entire areas of land use acquired from the JMT28.8 maps. The idea arose from the determination that important parts of error distribution are locally systematic effects, especially in the mountainous areas of the historical maps. These can be suitably represented by an autocorrelated random surface. Locally/regionally shifted homogenous areas were simulated; for example, one valley or one settlement is shifted in context on the map. Such homogenous areas are difficult to classify, but they may be simulated. The same foundations were used to apply spatially rough effects related to errors on short distances. This error distribution is quite similar to the boundary principle described above, but instead of lines representing boundaries, small areas were simulated as smooth and rough continuously varying error distribution surfaces.

Two artificially autocorrelated random surfaces were generated following a standardised unconditional normal distribution. Moran I was measured to achieve the required degree of autocorrelation (Haining 2003). This procedure was performed for the smooth and rough random surfaces bearing in mind their different spatial resolutions. The procedure is based on the exchanging of randomly selected points on the random surface that covers the study area and autocorrelation is controlled with the Moran I. This procedure may require many iterations.

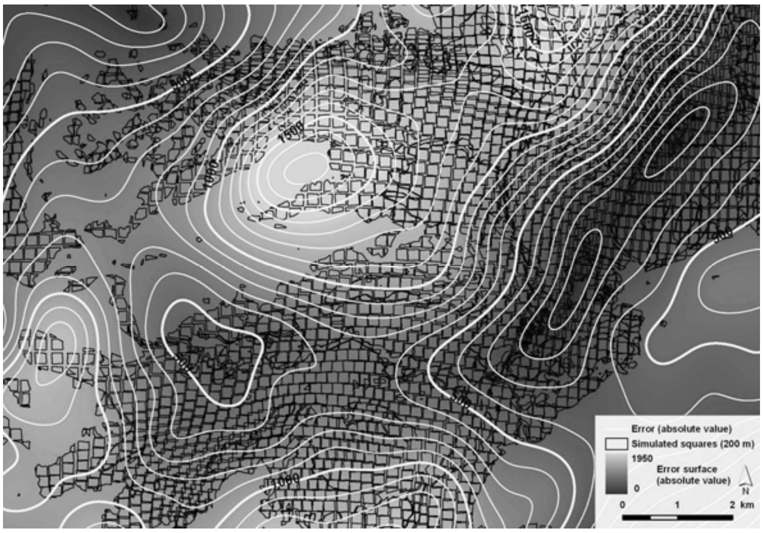

Figure 4. One of the six MC simulations of selected land uses from the JMT28.8. A random distribution surface represents absolute values between coordinates shifted in the *x* and *y* directions. Brighter areas mean larger shifts. Contour lines support shaded surface. The autocorrelated arrangement of square polygons with 200 m sides presents possible distortions of land use data due to shifts.

After the smooth and rough surfaces were computed, they were combined to produce a more complex error surface, considering empirical results of discovery and a determination of their portion.

The value of the portion comes from the relationship between many imprecise measurements between selected local/regional areas, and has a higher precision inside of those areas. Specifically, the chosen portion of the smooth surface was much higher than that of the rough one, but could be different (mountains vs. plains).

As with the boundary principle, the approximated $\sigma = 515$ m was annotated to the standardised unconditional complex error surface. This required attributing shifts in the x and y directions due to complex error surfaces. The shifts were applied using two approaches. The first was tessellation of the land use areas into uniform square areas (vector polygons) with 200 m sides based on attributes of land use categories. Each tessellated square was shifted independently in the x and y directions according to the values of the complex error surface. This approach was additionally combined with the second approach to the boundary principle described in the previous section, so the borders of all the square polygons were simulated (figure 4). Thus we effectively simulated error distribution due to all three presumed effects: random, locally systematic and systematic/gross. Twelve complex simulated error surfaces were produced for shifts in the x and y directions, thus the error was simulated for six times.

The second approach to the surface principle is similar to the first approach, but instead of squares as vector polygons, denser grid points were produced to simulate a grid-based land use error distribution surface.

CONCLUSION

Explanation and objectification of more complex but significant phenomena hidden within spatial data is a great challenge for the expansion of this study. We proved that a more effective research method is to study particular problems on the actual applications using real datasets, rather than try to generalise a theory again. A successful study that was presented using simulation Monte Carlo methods should be supported by other aspects, even by in situ observations, admitting the capabilities of researchers for objective interpretations of the problems. The procedures for consistent analyses can be significantly simplified, making the concept clearer and the process more effective.

ACKNOWLEDGMENT

The study was supported by the Ministry of Higher Education, Science and Technology of Slovenia, Interreg IIIB, SISTEMaPARC project, TMIS.plus.II, funded by the Austrian Research Promotion Agency within the ASAP programme, Norbert Pfeifer, Zoran Stančič, Theo Tijssen, and other anonymous persons.

BIBLIOGRAPHY

Brown, D.G., Duh, J.-D., Spatial simulation for translating from land use to land cover. International Journal of Geographical Information Science, 18, 35–60, 2004.

Burrough, P., McDonnell, R., 1998 Principles of Geographical Information Systems—Spatial Information Systems and Geostatistics. Oxford University Press.

Cressie, N.A.C., 1993 Statistics for Spatial Data. John Wiley & Sons.

Haining, R., 2003 Spatial Data Analysis: Theory and Practice. Cambridge University Press.

Khorram, S., Biging, G.S, Chrisman, N.R., Colby, D.R., Congalton, R.G., Dobson, J.E., Ferguson, R.L., Goodchild, M.F., Jensen, J.R., Mace, T.H. 1999, Accuracy Assessment of Remote Sensing- Derived Change Detection. ASPRS.

Lester, M., Chrisman, N., Not All Slivers Are Skinny: A comparison of Two Methods for Detecting Positional Error in Categorical Maps. In GIS/LIS ‘91 Proceedings, 28 October–1 November 1991, Atlanta, GA. ASPRS: 2: 648–658, 1991.

Openshaw, S., Charlton, M., Carver, S., Error propagation: a Monte Carlo simulation. In Handling Geographical Information: Methodology and Potential Applications, I. Masser and M. Blackmore (Eds.). Longman: 78–101, 1991.

Plewe, B., The Nature of Uncertainty in Historical Geographic information. Transactions in GIS, 6, 431–456, 2002.

Podobnikar, T., Kokalj, Ž., Quality of Historical Land Use Data in Triglav NP. In S. Wagenecht, E. Csaplovics (eds.). Spatial Information Systems for Transnational Environmental Management of Protected Areas and Regions in the Central European Space. Band 2, RHOMBOS-Verlag, Berlin: 2007 (in press).

Wood, J.D., Fisher, P.F., Assessing interpolation accuracy in elevation models. IEEE Computer Graphics and Applications, 13, 48–56, 1993.