STATISTICAL CARTOGRAPHY
WHAT IS IT?

Waldo Tobler
Professor of Geography
University of California
Santa Barbara, CA 93106

There is a long historical association of statistics and cartography, especially as relates to the theory of the adjustment of observations. Almost all of this history can be evoked by simply mentioning the name of Carl F. Gauss, an inventor of the method of least squares. In this tradition redundant measurements are used to estimate the amount of error contained in empirical observations, and "optimal" estimates are obtained by minimizing the mean square of this error, relative to some model of the phenomena being investigated. The classical theory is applied to the adjustment of surveys but the more modern work, under the name of collocation, also has applicability to interpolation problems, as encountered in the preparation of an isopleth map. These theoretical techniques are widely used in geodesy, but unfortunately are only rarely taught to cartographers or statisticians.

There is also a tradition in which cartography takes the form of graphical illustration of statistical data. Today this is often referred to as "thematic" cartography, sometimes "statistical" cartography. The early roots lie in the work of Playfair, Minard, Quenelet and similar individuals, and are detailed by Funkhouser and several reports of the International Statistical Association, a tradition which continues to this day. By the mid-1800's choropleth and isopleth maps had been invented, and data were being assembled by rectangular grid cells - a technique lately "rediscovered" in relation to computer assisted cartography. Today thematic cartography is an active area of experimentation, research, and (more recently) psychological testing. One need which I have emphasized for many years is to incorporate our uncertainty into such maps by drawing them in defocused form, fuzzy in proportion to the variance of the data. On a modern CRT one might do this by alternating displays within the refresh cycle. We are now also moving into a period in which color graphics are becoming very inexpensive and it is apparent that experimentation with the tensorial nature of this medium is underway. As our banks of data increase in size the importance of visual summaries can be expected to increase in importance.

In elementary statistics the first descriptive measures learned concern the central tendencies. There are of course the two-dimensional variants of these: the center of gravity, the bivariate median, the point of minimum aggregate travel; and the dispersion measures (the standard distance, bivariate ellipses, Mendeleev's cartography and its extensions by Bachi, and so on). It is even possible to go further than is usually done, to bivariate regression, for example - treating the picture of a child's face as a linear (or non-linear) function of the picture of the face of its parents, or regressing a geographical map on its historical precedents. The β-coefficient in such a regression is of course now a tensor rather than a scalar value. We can also apply transformations to the geographical variables. In urban studies logarithmic distances from the center of the city are often used, and cartograms are a form of bivariate uniformization designed to stretch the space so that the effect of some variable, such as the size of the areal units, is eliminated. Unfortunately most useful two dimensional transformation are not separable and require the solution of partial differential equations. Thus they are not easily effected.

The analysis of time series data is often studied only in the more advanced statistical courses. This is mostly because the observations can no longer be considered independent; there is an essential order to the phenomena which the simpler measures (mean, variance, etc.) do not capture. But the geographical case is even more complicated. It would be considered absurd to arrange times series data in alphabetical order, by month say, before the analysis. Yet one commonly finds geographical data analyzed in alphabetical order. The spatial dependencies are thus ignored, and most statistical tests are invalid in such a situation. It is rare that a statistician applies a geographical "runs" test, or considers geographical adjacency relations when doing a test for the 'significance' of some observation, even when the phenomena are as important as the incidence of leukemia. It is necessary to model the spatial spread effects in these data in order to obtain valid inferences. There are also resolution effects which cannot be ignored. When using spatially aggregated data, by county say, we are obtaining a blurred picture of the phenomena being studied. I am continually amazed at how many professional statisticians are unaware of the sampling theorem and its implications for areal measurements. Concommitantly it is obvious that the many techniques which have been developed for the enhancement of spatial data (e.g., edge detection) should be applicable to the bivariate geographical arrangement of health related phenomena. For the comparison of two geographical arrangements (e.g. pollution and cancer) bivariate cross spectral analysis recommends itself, especially as this can now be done using the fast Fourier transform. The interesting and colorful bivariate cross-maps recently introduced by the Bureau of the Census do not seem to me to have quite the sensitivity needed for such important studies. At a more advanced level it is usually appropriate to consider dynamic effects and to go on to time-space series work. Thus this short essay has only touched on a few facets of the relation between cartography and statistics.

Bibliography

Bennett, R. (1979) Spatial Time Series. Cambridge, University Press.

Cliff, A., et al. (1975) Elements of Spatial Structure. Cambridge, University Press.

Davis, J. and M. McCullagh. (1975) Display and Analysis of Spatial Data. New York, J. Wiley.

Funkhouser, H. (1937) "Historical Development of the Graphical Representation of Statistical Data". OSIRIS, vol. 3, pt. 1, pp. 269-405.

Getis, A. and B. Boots. (1978) Models of Spatial Processes. Cambridge, University Press.

Hagerstrand, T. (1967) Innovation Diffusion as a Spatial Process. (Pred Translation), Chicago, University Press.

Helmert, F.R. (1924) Die Ausgleichungsrechnung nach der Methode der kleinsten Quadrate. 3rd ed., Teubner, Berlin.

Kaula, W. (1967) "Theory of Statistical Analysis of Data Distributed Over a Sphere", Reviews of Geophysics, vol. 1, pp. 83-107.

Larimore, W. (1977) "Statistical Inference on Stationary Random Fields", Proc. IEEE, 65, 6, pp. 961-970.

Mantel, N. (1967) "The Detection of Disease Clustering and a Generalized Regression Approach", Cancer Research, 27, 2, pp. 209-220.

Matheron, G. (1971) The Theory of Regionalized Variables and its Applications. Fontainbleau, Ecole Superieure des Mines.

Moritz, H. (1970) Eine Allgemeine Theorie der Verarbeitung von Schwermessungen nach kleinsten Quadraten. Heft Nr. 67A, Munich, Deutsche Geodatische Kommission.

Neft, D. (1966) Statistical Analysis for Areal Distributions. Philadelphia, Regional Science Assn.

Robinson, A. (1971) "The Genealogy of the Isopleth", Cartographic Journal, pp. 49-53.

Rosenfeld, A. and A. Kak. (1976) Digital Picture Processing. New York, Academic Press.

Sibert, J. (1975) Spatial Autocorrelation and the Optimal Prediction of Assessed Values. Ann Arbor, Department of Geography, University of Michigan.

Tobler, W. (1969) "Geographical Filters and their Inverses", Geographical Analysis, I, 3, pp. 234-253.

Tobler, W. (1973) "A Continuous Transformation Useful for Districting", Annals New York Academy of Sciences, 219, pp. 215-220.

Tobler, W. (1975) "Linear Operators Applied to Areal Data", in J. Davis and M. McCullagh, Display and Analysis of Spatial Data. New York, J. Wiley.

Tobler, W. (1975) "Mathematical Map Models", Proceedings, International Symposium on Computer Aided Cartography, Reston, ACSM, pp. 66-73.

Tobler, W. (1978) "Comparing Figures by Regression", Computer Graphics (ACM Siggraph), 12, 3, pp. 193-195.