STATISTICAL EVALUATION OF ACCURACY FOR DIGITAL CARTOGRAPHIC DATA BASES

Arnold Greenland & Robert M. Socher
IIT Research Institute
5100 Forbes Boulevard
Lanham, Maryland 20706

MAJ Michael R. Thompson
U.S. Army Engineer Topographic Laboratories
Fort Belvoir, Virginia 22060

BIOGRAPHICAL SKETCHES

Dr. Greenland is a Senior Analyst for the IIT Research Institute. He earned a B.S. in Mathematics from Case Western Reserve University in 1969, an M.A. in Mathematics from the University of Rochester in 1973, and a Ph.D. in Probability Theory from the University of Rochester in 1975. Dr. Greenland has applied statistical methods to a wide range of engineering, economic, and environmental problems. His current interest is in the application of statistical methods to digital geographic data bases.

Mr. Socher is a Senior Programmer/Analyst for the IIT Research Institute. He earned a B.A. in Mathematics from St. John's University, Minnesota in 1967. During his career, Mr. Socher has developed a broad base of knowledge in automated data processing for cartographic/terrain data. For the past seven years, he has served as project manager guiding the design and development of interactive/batch graphics software on the Digital Terrain Analysis Station (DTAS) for the U.S. Army Engineer Topographic Laboratories.

MAJ Thompson received a B.S. degree from the United States Military Academy in 1973 and was commissioned as a military officer in the U.S. Army. He has served in infantry assignments in both Europe and the United States and is a graduate of the Infantry Officers Advanced Course. In 1981, MAJ Thompson earned an M.S. in Photogrammetry and Geodesy from Purdue University and is currently assigned to the U.S. Army Engineer Topographic Laboratories (ETL) as an R&D Coordinator.

ABSTRACT

The major statistical techniques used by cartographers to measure accuracy in digital cartographic data bases -- namely, linear and circular probable errors -- were derived for application to analogue cartographic products. Digital products because of their use, application, and construction do not necessarily fit in the accuracy model of paper products. This paper documents the application of a relatively new measure, the Kappa statistic, for analyzing the accuracy of digital geographic feature data bases. The specific application was the comparison of two digital cartographic data bases provided by the Defense Mapping Agency (DMA).

BACKGROUND

In order to evaluate the Army terrain analysis requirements for digital terrain data, the adequacy of each of two DMA prototype data

sets was assessed in terms of data content and completeness, absolute and relative accuracy (vertical and horizontal), resolution of the elevation and feature data, and the format or structure in which the data are recorded (including the coordinate systems and reference datums). This evaluation was aided by the use of a commercial, interactive graphics system, the Digital Terrain Analysis Station (DTAS), which is designed to automate tactical terrain analysis.

The first prototype data set which was provided by the DMA Hydrographic/Topographic Center (HTC) is limited to those natural and man-made features which are of tactical military significance. The data set consists of six feature topics which include Surface Configuration, Vegetation, Surface Materials (soils), Surface Drainage, Transportation, and Obstacles. The prototype is in a 12.5 meter gridded format and utilizes the Universal Transverse Mercator (UTM) coordinate system. Each grid point consists of 199 bits where the first 16 bits contain the elevation and the other 183 bits contain the associated codes for the features.

The second prototype data set which was produced by DMA Aerospace Center (AC) is the High Resolution Prototype Data Base enhanced for tactical terrain analysis applications with the addition of three new micro-descriptors (Surface Drainage, Transportation, and Vegetation). The data set is in a vector format and utilizes the World Geodetic System (WGS) coordinate system. Each feature record consists of two or three subrecords:

- A primary feature description
- Zero to six optional micro-feature descriptions
- A "delta set list" of relative coordinate pairs for the location information.

Elevation information is obtained from a DMA standard Digital Terrain Elevation Data (DTED) magnetic tape file. The data is in a 1201 x 1201 matrix with a grid spacing of 3 seconds of arc which equates to approximately 60 meters by 90 meters. This data also uses WGS coordinates. Both prototype data sets consist of two areas in the state of Washington, Fort Lewis and Yakima Firing Center.

To evaluate the ability of the two DMA prototype data sets to support terrain analysis requirements, software routines were developed to read and reformat the DMA data into the DTAS data base. With the DMA prototype data sets reformatted for the DTAS, terrain analysis models (products), developed as part of a software development program were executed using both DMA prototype data sets as input. These products were then compared to manually-prepared products produced for the evaluation. Operational suitability, in terms of the usefulness and the acceptability of the prototype data element features and the automated terrain analysis products, was determined by visual analyses conducted by the Terrain Analysis Center at ETL, with support from military terrain analysts.

This paper addresses the aspect of the data base evaluation during which objective measures were formulated to compare and to quantify the differences between digital and manual data features and products. The same techniques were used to evaluate features and products at degraded resolutions to determine the minimum acceptable data resolution necessary which could satisfy Army requirements. The

end result is a viable objective method that can be used to statistically capture the subjective analysis performed in visual evaluation.

A thorough, in-depth statistical analysis of the elevation data was also performed and is documented in Herrmann, et al. (1984). The remainder of this paper, however, will concentrate on the statistical analysis of the feature data.

THE STATISTICAL EVALUATION OF FEATURE DATA

The selection of the statistical measure used for evaluation of feature data was driven by two considerations. The first relates to a common methodology, Circular Probable Error (CPE), applied to feature data. In this case, one obtains a sample of locations (monuments) between which one measures the distances in the two feature map representations to be compared. These distances (errors) are used to estimate the CPE statistics. The problem with this approach is that it ignores the fundamental character of the feature data. Feature data is categorical or nominal and not interval as is required for the CPE approach. This means that the variables do not take on numerical values measured on a continuum, but rather they are simply categories or names of the type of features. A more consistent approach is to use the methods of categorical analysis to obtain accuracy metrics. The second consideration flows from the way individuals normally evaluate a product visually. The most common reaction when confronted with a feature map produced by the DTAS was to overlay the map produced directly from the digital data onto an existing hard copy map, and see how it looked. We felt it was useful to find a statistical approach which was the realization of that visual comparison process. These two considerations led to the choice of a statistical technique noted in some cartographic contexts (see Chrisman (1982), Congalton and Mead (1981)) called Kappa. The details of this method will be discussed below.

A Measure of Agreement

The ideas discussed above were the impetus behind the formulation of the feature data metrics for the data base evaluation. The approach is best illustrated by Figure 1. The figure contains two realizations of a feature classification of a particular region. The solid polygon represents the region designated as category 1 by the first product and the interior of the dashed line represents the region designated as Category 1 by the second product. As one can readily discern, the two classifications do not agree completely. The disagreement can be described in the matrix shown in TABLE 1. Let p_{11} be the fraction of the total area displayed in which the first product or source shows category 1 and so did the second; let p_{12} be the fraction of the total area displayed in which the first source shows category 1 but the second shows category 2; etc. Let a subscript of "+" indicate summation over that index in the matrix. Of course $p_{++} = 1$, since that is the total area of interest.

One obvious measure of agreement is the sum down the diagonal, $p_0 = p_{11} + p_{22}$. This measure ignores the magnitudes of the "marginal" probabilities (the fractions shown as row and column sums in TABLE 1). Cohen (1960) suggested a measure based on the table just described which removes the effect of chance from the measure.

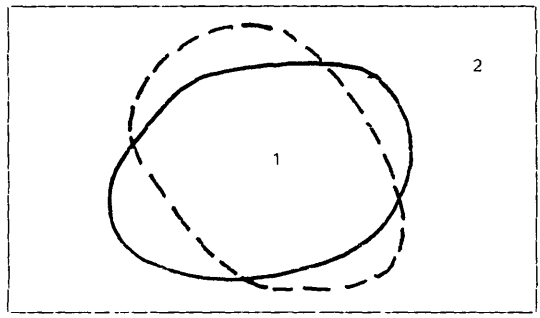

Figure 1

TABLE 1

		Classified by Second Source		
		1	2	
Classified by First Source	1	p_{11}	p_{12}	p_{1+}
	2	p_{21}	p_{22}	p_{2+}
		p_{+1}	p_{+2}	$p_{++} = 1$

The expected probabilities for each cell in the matrix are computed using the marginals as follows: $p_{1+} \times p_{+1}$ is the expected probability for cell (1, 1), etc. Thus the "expected" fraction of agreement is: $p_e = p_{1+}\ p_{+1} + p_{2+}\ p_{+2}$. The Kappa statistic is defined as: $k = (p_o - p_e)/(1 - p_e)$ and is interpreted as the proportion of agreement over and above chance agreement. A detailed explanation of this measure can be found in either Bishop et al. (1975) or Fleiss (1981).

A final methodological comment is required when the variables used for classification are something other than nominal, such as ordinal, in nature. A disagreement in this context may be more or less critical. An example is that classifying a point as type 1 in one method and type 2 in another may be much less problematic than if the classification types were, respectively, 1 and 7. To remedy this situation Cohen (1968) also introduced the weighted Kappa. A complete explanation of this concept is beyond the scope of this paper, though the reader is referred to Fleiss (1981) for an excellent explanation. The interpretation is essentially the same, i.e., the fraction of weighted agreement over and above chance. The "weightings" are often displayed in a matrix such as the one shown in

TABLE 2. The values there show for example that an exact agreement between "control" and "product" maps is given full weight as demonstrated by 1's down the main diagonal. A disagreement of only one category (such as 2 for control and 3 for product) is given a weight of 1/2. Of course, the weight matrix for the unweighted Kappa is an identity matrix.

TABLE 2
KAPPA WEIGHTS

Control	PRODUCT						
	1	2	3	4	5	6	7
1	1	.5	0	0	0	0	0
2	.5	1	.5	0	0	0	0
3	0	.5	1	.5	0	0	0
4	0	0	.5	1	.5	0	0
5	0	0	0	.5	1	.5	0
6	0	0	0	0	.5	1	.5
7	0	0	0	0	0	.5	1

The Operational Issues

The first step in the evaluation of a set of feature polygons or product polygons is to obtain a control product for comparison. To that end ETL undertook the task of hand digitizing the feature and product overlays which were to be evaluated. The included sets of feature and product polygons were:

- Soil
- Slope
- Vegetation
- Cross Country Movement (CCM).

CCM is an analytical model that predicts off-road speed potential based on vehicle characteristics and the terrain (soil, slope, and vegetation).

The second step in the process was to register the matching feature or product, from either the HTC or AC source, to the digitized control set of polygons. The computational task required to compute Kappa or weighted Kappa is to obtain the cross-classification matrix such as the one shown in TABLE 1. The matrix values p_{ij} are estimates for the areas of that type of agreement or disagreement in the maps. One method is to use polygon processing

capabilities to identify the various regions and then to calculate the enclosed areas. A second method is to use an underlying grid structure where each grid point represents a region of area surrounding it. For this application, the grid approach was chosen.

The grid structure allowed for the estimation of fractional agreement by comparing the files position by position in the grid. For details of the mechanics of the grid approach and an example of an agreement matrix, see Herrmann et al. (1984).

Any human errors in polygon labeling in the hand digitized, HTC, or AC data sets that were detected during the creation phase or visual analysis phase of this evaluation were corrected. Also, as the statistical analysis phase was executed, the agreement matrices were examined for any abnormally large numbers off the main diagonal which could be an indication of mislabeled polygons. The input data was closely scrutinized and any mislabeled polygons corrected. Then the statistical analysis was performed again.

One aspect of this approach that has not yet been considered in the literature is the evolution of "rules of thumb" to help interpret the Kappa results. In order to help remedy this situation, this study included two separate attempts to begin this evaluation process. First we generated a "benchmark" value as follows. The soil polygons for Yakima were digitized two separate times. The value of Kappa derived from those two sets of polygons was used as a reference point from which to assess the other values of Kappa that occurred. As shown in TABLE 3, the benchmark was 0.967, a very high value. It has the interpretation that approximately 97% of the grid points (an estimate of area) were in agreement over and above what would be expected to agree by chance alone. The second thing done to evolve the rules of thumb for interpeting Kappa is described below in the section on granularity.

Results

Following the methods described above Kappa and weighted Kappa coefficients were computed and are shown in TABLES 3 and 4. The former table contains coefficients for features. They are shown grouped by feature type (soil, slope, and vegetation) crossed with region (Yakima and Fort Lewis). Within that classification, four values of HTC data granularity (12.5m, 25m 50m and 125m) and the lone AC format are shown. The latter table contains the four products at different HTC granularities and the one AC product. The major cells shown in the table contain weighted Kappa statistics broken down by three different weighting schemes. The first one, unweighted, is identical to the method used for the feature polygon evaluation in that only those points along the main diagonal are considered to agree. The second weighting scheme takes into account those grid points where the two sources differ by one speed category, and weights these points 1/2 as great as a perfect match (see TABLE 2). In the final weighting scheme, proposed by Cicchetti and Allison, a linear approach is taken. Weights are assigned with respect to relative positioning in the matrix as applied through the formula: $W_{ij} = 1 - [1 - (i-j)/(h-1)]$, where h = the number of categories. Therefore, in a seven by seven matrix the two minor diagonals closest to the main diagonal are assigned weights of 5/6, the next two diagonals are assigned weights of 4/6, etc. These weightings are crossed with location (Yakima and Fort Lewis) in TABLE 4.

TABLE 3
FEATURE KAPPA STATISTICS*

		Spacing	Yakima	Ft. Lewis
Soil	HTC	12.5	.961	.946
		25	.960	.942
		50	.954	.941
		125	.924	.920
	AC	N/A	.944	.910
Slope	HTC	12.5	.900	.954
		25	.900	.948
		50	.895	.939
		125	.858	.891
	AC	N/A	.921	.953
Vegetation	HTC	12.5	.960	.938
		25	.956	.937
		50	.957	.922
		125	.937	.855
	AC	N/A	.964	.928

*Benchmark = .967

TABLE 4
CCM WEIGHTED KAPPA STATISTICS

		Spacing	Yakima	Ft. Lewis
Unweighted	HTC	12.5	.925	.709
		25	.919	.702
		50	.916	.699
		125	.883	.661
	AC	N/A	.911	.774
"1/2" Adjacent Diagonal Weighting	HTC	12.5	.928	.790
		25	.921	.782
		50	.918	.778
		125	.888	.739
	AC	N/A	.913	.825
Cicchetti & Allison Weighting	HTC	12.5	.931	.801
		25	.925	.793
		50	.922	.789
		125	.893	.747
	AC	N/A	.918	.829

The two tabulations just described were then used to address the following issues (shown in order of importance):

- the extent to which the two prototype data bases support feature and product generation

- the appropriateness of the DMA data bases in producing CCM
- the granularity of data required to produce acceptable features and products.

The following discussion of these issues is presented in the reverse order, for reasons of logical development. We will then discuss the extent to which the Kappa statistic fulfilled those goals.

Granularity. It would be desirable to have statistical measures which bring with them accepted "rules of thumb" for interpreting their magnitude. Unfortunately, because the Kappa statistic has not been used extensively in this area, it was necessary to evolve acceptable levels of Kappa by close reference to the visual analysis. The visual analysis by ETL technical personnel was accomplished completely independently of the statistical analysis. The hope was that when the two analyses were compared they would be mutually supportive. Indeed, this was the case. The visual analysis also allowed one to evolve a working level of Kappa below which one is likely to see unacceptable results. Refer first to TABLE 3. The most striking thing about the table is the uniformly high percentages of agreement (Kappa coefficients) shown. Only a few values on the page are below .900. The visual analysis of features found essentially that the computer generated feature plot for Yakima slope (with k = .858) and for Fort Lewis vegetation (k = .855) were the only unacceptable features. Those were in fact the two lowest Kappa coefficients. The next lowest value was .891 for Fort Lewis slope which was acceptable. The logical conclusion is that somewhere between .850 and .900 the level of agreement becomes unacceptable.

The specifics of the granularity study are as follows. It would be quite justifiable to assert that spacing as high as 50 meters between raw data values can be tolerated in producing acceptable feature overlays. However, spacing of 125 meters between data points is not universally acceptable, though when the number of polygons is small the results (as one would expect) were reasonable.

There were also some general things to be said about the differences between Yakima and Fort Lewis which were detected by the percentage agreement figures. First of all, in Yakima where the terrain is much more rugged than Fort Lewis, the slope coefficients were uniformly worse at each granularity by approximately 5 percentage points. However, in Fort Lewis, where vegetation is complex, the coefficients were uniformly worse than Yakima where vegetation is sparse. The difference between Yakima and Fort Lewis widened as the data was degraded.

CCM. The results for the CCM analysis are shown in TABLE 4. The first point to make is that the Fort Lewis results are uniformly not acceptable. There are, however, several pieces of information which allow one to explain the discrepancy for Fort Lewis. The problem is categorical representation of stem spacing and stem diameter in the DMA data bases. The DMA data format has the categorical values for stem spacing broken down into groups that are no smaller than .5 meters wide and stem diameter broken into groups that are no smaller than 2 centimeters wide. Consequently, the algorithm could only use a mid-range estimate for the stem spacing

and diameter instead of an exact value which was available to the analysts creating the manual products. The results were quite dramatic as the Fort Lewis CCM (from the HTC source) got no better than 80% agreement (even when one gives partial credit by non-exact agreements as is done in the weighting schemes).

The contention that the error is due to stem spacing and diameter is reinforced by the other results. For the Yakima CCM, one would expect that the vegetation values would not hamper the results since Yakima contains no significant vegetation; and indeed, the Yakima CCM Kappa values are above 90% in most cases. In fact, the resulting values fall roughly between the Kappa values for soil and slope in Yakima which is what one would expect.

A second piece of evidence is the CCM results for Fort Lewis. It just happened that the DMA data in the AC format included one additional category of stem spacing than does the HTC data; therefore, one could expect just slightly better results for the AC data. Indeed, for each weighting scheme used, the AC CCM showed the best level of agreement among all done in that group.

There are two points to emphasize. First that the CCM algorithm is quite sensitive for some variables and therefore would be much improved if the coding scheme allowed a more accurate description of these variables. Second, it is likely that when the data is provided that the results will be quite good since the Yakima CCM products (which do not rely on vegetation) were, with the exception of the 125 meter spacing, well above 90% agreement and thus acceptable products.

AC and HTC Comparison. Both data sources are quite good in representing the features with neither data format showing a uniformly superior performance. For example, for the soil features, HTC results were slightly better than those for the AC source. However, for slope the AC source was about the same or better than the HTC. For vegetation, it was mixed with AC better for Yakima and HTC better for Fort Lewis. Neither pattern of the differences nor their magnitudes were significant enough to suggest that one data base was "better" than the other for features or feature dependent products (like CCM); and both were acceptable. The fact that the Fort Lewis CCM was superior for the AC source is worth noting; but neither product was really in the acceptable range. Indeed, the upgrade of stem spacing and diameter to numeric quantities would be helpful for both data bases.

Conclusions

The use of the Kappa and weighted Kappa statistics was a very useful and credible methodology for analyzing the relative strengths and weaknesses of the two data bases; and as such provided the Army with a good case for the conclusions reached in the prior paragraphs. The major area of work will be in educating users in map products about the meaning of the statistics and "rules of thumb" for interpreting the strength of the results.

FURTHER RESEARCH

There are several areas of further research that can be mentioned here. The Kappa statistic needs to be applied in more situations so that rules of thumb about the magnitude can be

evolved. In addition, other modifications to the Kappa which are more spatially motivated need to be developed. The current form of the statistic does not take the size and extent of the boundary into consideration. Therefore, as pointed out, the regions with many polygons are, because the opportunity is there, more likely to have lower levels of agreement.

ACKNOWLEDGEMENTS

We gratefully acknowledge the support of the U.S. Army Engineer Topographic Laboratories for this work as well as programming contributions of Mary Armstrong and James Span of the IIT Research Institute. Finally, we thank Jacqueline Rota and Chris Magagna for their accurate typing and continuing patience during preparation of this paper.

REFERENCES

Bishop, Y.M.M., Feinberg, S.E. and Holland, P.W. 1975, Discrete Multivariate Analysis, MIT Press, Cambridge, MA.

Chrisman, N.R., 1982, Beyond Accuracy Assessment: Correction of Misclassification: Proceedings ISPRS Commission IV, p. 123-132.

Chrisman, N.R., 1984, Alternatives for Specifying Quality Standards for Digital Cartographic Data: Issues in Digital Cartographic Data Standards Report # 4, ed. H. Moellering, National Committee for Digital Cartographic Data Standards, ACSM, Columbus, Ohio.

Cohen, J., 1960, A coefficient for nominal scales: Educ. Psychol. Meas., 20, p. 37-46.

Cohen, J., 1968, Weighted Kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychological Bulletin, 70, p. 213-220.

Congalton, R.G. and Mead, R.A. 1981, A Quantitative Method to Test for Similarity Between Photo Interpreters: Technical papers 47th Annaul Meeting ASP, p. 263-266.

Fleiss, J. L., 1981, Statistical Methods for Rates and Proportions, Second Edition, John Wiley & Sons, New York.

Greenwalt, C. R. and Shultz, M.E. 1962, Principles of Error Theory and Cartographic Applications ACIC Technical Report No. 96, DMA-AC, St. Louis, MO.

Hermann, R. et. al., 1984 Army Digital Topographic Data Requirements, ETL-GSL-2. U.S. Army Engineer Topographic Laboratories, Fort Belvoir, VA.