Quality Assessment of Polygon Labeling

Gerardo Bocco and Hugo Riemann

Abstract
The high costs in time and resources of data validation makes the correct selection of a simple but robust adjustment method crucial. In this paper we describe an approach suitable for a polygon labeling assessment. The approach uses the geometric distribution where the sample size is a function of the desired confidence level of the database. It allows the detailed random verification of a selected map. All the area is equally exposed to testing and, if needed, to correction.

Introduction
Quality and error in GIS databases have been widely researched, with emphasis in either location, labeling, or operational errors (Walsh *et al.*, 1987; Congalton, 1991; Gopal and Woodcock, 1994). The main interest has been in validation of data entry, error detection (location or labeling), error propagation, and inaccuracies in data presentation to users (Aronoff, 1989; Lunetta *et al.*, 1991; Lanter and Veregin, 1992; Thapa and Bossler, 1992; Kraus, 1994; Veregin, 1994).

Aronoff (1989) has suggested that an error-free database does not exist; he proposes quality management rather than complete error eradication, following any type of cost-benefit strategy.

Statistical sampling and adjustments of results to describe error probabilities have been widely used in quality evaluations. A typical case is accuracy assessment of spectral classifications (van Genderen and Lock, 1977; van Genderen *et al.*, 1978; Walsh *et al.*, 1987; Jansen and van der Wel, 1994; Fitzgerald and Lees, 1994), wherein probabilities of correctness are established for spectral classes in error matrices.

The evaluation of the quality of polygon labeling after digitizing is of an operational nature, and of interest here. A set of samples of polygons of all categories in the digital database was verified against the same polygons as depicted in the original map. In contrast to the assessment of spectral classifications, where a percentage of error is permitted, in this case the detection of a single labeling error assumes its eradication. When the maps to be digitized are highly complex, labeling quality assessment is mandatory.

The high costs in time and resources of data validation makes the correct selection of a simple but robust adjustment method crucial. In this paper we describe an approach suitable for polygon labeling assessment in a geographic information system environment.

We used the cases of an urban land-use map of the city of Mexicali, state of Baja California, and a vegetation map of the Baja California Peninsula. Both are relatively complex: the urban map has 556 polygons distributed in five categories (Table 1) and the vegetation map has 1629 polygons distributed in 20 categories (Table 2). The entire exercise was performed in ILWIS (Anonymous, 1992), a PC-based GIS with vector and raster capabilities.

Method
The verification of polygon labeling can be described in terms of a binomial case of a success-failure criterion. It can be represented by a Bernoulli experiment, with two possible outcomes, correct (the polygon labeled as class A in the digital database belongs to class A in the original map) or incorrect (the given polygon was incorrectly labeled). This verification routine has usually been implemented by means of a binomial distribution (Congalton, 1991). A condition of this model is that, in order to estimate the probability of finding errors in a given sample, the sample size must be defined in advance.

In this paper we propose instead the use of the geometric distribution (Mood *et al.*, 1974), a special case of the binomial model. The geometric model requires as many Bernoulli experiments as needed to find the first "success." In this context, the first labeling error is a success when the sampling procedure stops and the map is rejected.

Defining as the random variable as "X" that describes the number of Bernoulli experiments needed to find the first labeling error, we had

$$P[X=x] = P[F, F, F, \ldots, S] = q^{x-1}p, \; x = 1, 2, \ldots \quad (1)$$

where F and S were, respectively, failure and success; $q = 1 - p$ was the probability of not finding a labeling error; and x was the number of experiments until the first error was found. The first and second moments of this distribution (Bhattacharyya and Johnson, 1977:154–155) were defined as

$$\mu = 1/p \quad (2)$$

$$\sigma^2 = q/p^2 \quad (3)$$

Procedure
The maps to be tested were displayed on a computer monitor using scenes of 640 by 480 pixels, with resolutions large enough to visualize up to the smallest digitized polygon. Further, each scene was divided into quadrants. From the center of each quadrant, a new point was defined as follows: Two numbers were selected at random to represent the number of cells to move in the x and y directions, from the center to the right and upwards (with even numbers) and to the left and downwards (with odd numbers).

From these points, four search areas of 100 by 100 cells each were defined. In these areas the labeling of all polygons was verified, and their categories were compared to those of the original paper map from which the polygons had been

G. Bocco is with the Instituto de Ecología, Universidad Nacional Autónoma de México, AP 27-H, Sta. María, 58090 Morelia, Michoacán, México.

H. Riemann is with the Departamento de Estudios Urbanos y del Medio Ambiente, El Colegio de la Frontera Norte, P.O. Box L, Chula Vista, CA 91912.

Photogrammetric Engineering & Remote Sensing,
Vol. 63, No. 4, April 1997, pp. 393–395.

0099-1112/97/6304-393$3.00/0

TABLE 1. URBAN LAND-USE CATEGORIES. CITY OF MEXICALI, MEXICO.

Category	Code	Number of polygons
Residential	1	28
Commerce & services	2	275
Infrastructure	3	189
Industry	4	27
Vacant	5	37

TABLE 2. VEGETATION CATEGORIES. BAJA CALIFORNIA, MEXICO.

Category	Code	Number of polygons
Agriculture	1	201
Pine-oak forest	2	42
Chaparral	3	172
Agave shrub	4	225
Montane shrub	5	73
Sarco-crassulaceous shrub	6	120
Sarcophyllous shrub	7	165
Secondary sarco-shrub	8	232
Larrea-Mezquite shrub	9	31
Microphyllic shrub	10	4
Larrea shrub	11	25
Yucca shrub	12	11
Crassulaceous shrub	13	12
Mezquite forest	14	67
Grassland	15	108
Dry tropical forest	16	11
Sandy desert vegetation	17	24
Halophytic vegetation	18	42
Mangrove	19	21
Bare rock or soil	20	48

TABLE 3. LABELING ERROR FOR THE DIGITAL MAP OF MEXICALI, MEXICO.

Category	x	p	q	$P[X=x]$	C (%)
Residential	9	0.11	0.89	0.043	95.7
Commerce & services	24	0.04	0.96	0.016	98.4
Infrastructure	23	0.04	0.96	0.016	98.4
Industry	3	0.33	0.67	0.148	85.2*
Vacant	8	0.13	0.88	0.049	95.1

x = number of verified polygons per category; $p = (1/x)$; $q = 1\text{-}p$;
C = confidence (1-P [$X=x$]), using the first sampling procedure.
* = category where a labeling error was found.

digitized. The procedure stopped when all polygons in the search area were verified or when the first labeling error was detected.

Results and Discussion

Using the procedures described, 67 polygons or approximately 12 percent of the total number of polygons were tested for the urban land-use map (Table 3). Only in the class "Industry" was an error detected; the rest of the classes were successfully verified. The error was corrected and a second sampling procedure was initiated (Table 4). In this case 64 polygons or 11.5 percent of the total were verified, and no errors were detected. The sampling was stopped, and the overall confidence was set at 96.6 percent (Table 4).

The method was further tested on the vegetation map. We verified 148 polygons or 9 percent of the total number of polygons (Table 5). An error was detected in the class "Mezquite forest," the polygon was properly labeled, and a second sampling procedure was initiated. In this case, 153 polygons were tested and no errors were detected. The overall confidence after sampling was 94 percent (Table 6).

This approach allowed the detailed random verification of a given map. All the area is equally exposed to testing.

TABLE 4. LABELING ERROR FOR THE DIGITAL MAP OF MEXICALI, MEXICO, USING THE SECOND SAMPLING PROCEDURE.

Category	x	p	q	$P[X=x]$	C (%)
Residential	15	0.07	0.93	0.0254	97.46
Commerce & services	14	0.07	0.93	0.0273	97.27
Infrastructure	18	0.06	0.94	0.0210	97.90
Industry	7	0.14	0.86	0.0567	94.33
Vacant	10	0.1	0.90	0.0387	96.13

TABLE 5. LABELING ERROR FOR THE DIGITAL VEGETATION MAP OF BAJA CALIFORNIA PENINSULA, MEXICO USING THE FIRST SAMPLING PROCEDURE.

Category	x	p	q	P [$X=x$]	C (%)
Agriculture	13	0.08	0.92	0.029	97.06
Pine-oak forest	9	0.11	0.89	0.043	95.67
Chaparral	5	0.20	0.80	0.082	91.81
Agave shrub	17	0.06	0.94	0.022	97.77
Montane (cloud) shrub	9	0.11	0.89	0.043	95.67
Sarco-crassulaceous shrub	8	0.12	0.88	0.049	95.09
Sarcophyllous shrub	11	0.09	0.91	0.035	96.50
Secondary sarco-shrub	13	0.08	0.92	0.029	97.06
Larrea-Mezquite shrub	5	0.20	0.80	0.082	91.81
Microphyllic desertic shrub	4	0.25	0.75	0.105	89.45
Larrea shrub	5	0.20	0.80	0.082	91.81
Yucca shrub	4	0.25	0.75	0.105	89.45
Crassulaceous shrub	4	0.25	0.75	0.105	89.45
Mezquite forest	3	0.33	0.67	0.148	85.19*
Grassland	11	0.09	0.91	0.035	96.50
Dry tropical forest	4	0.25	0.75	0.105	89.45
Sandy desert shrub	4	0.25	0.75	0.105	89.45
Halophytic vegetation	9	0.11	0.89	0.043	95.67
Mangrove	5	0.20	0.20	0.082	91.81
Bare rock or soil	5	0.20	0.80	0.082	91.81

TABLE 6. LABELLING ERROR FOR THE DIGITAL VEGETATION MAP OF BAJA CALIFORNIA PENINSULA, MEXICO USING THE SECOND SAMPLING PROCEDURE.

Category	x	p	q	P[X=x]	C (%)
Agriculture	11	0.09	0.91	0.035	96.50
Pine-oak forest	9	0.11	0.89	0.043	95.67
Chaparral	6	0.17	0.83	0.067	93.30
Agave shrub	14	0.07	0.93	0.027	97.27
Montane (cloud) shrub	7	0.14	0.86	0.057	94.33
Sarco-crassulaceous shrub	8	0.12	0.88	0.049	95.09
Sarcophyllous shrub	9	0.11	0.89	0.043	95.67
Secondary sarco-shrub	14	0.07	0.93	0.027	97.27
Larrea-Mesquite shrub	6	0.17	0.83	0.067	93.30
Microphyllic desertic shrub	4	0.25	0.75	0.105	89.45
Larrea shrub	5	0.20	0.80	0.082	91.81
Yucca shrub	5	0.20	0.80	0.082	91.81
Crassulaceous shrub	5	0.20	0.80	0.082	91.81
Mezquite forest	8	0.12	0.88	0.049	95.09
Grassland	11	0.09	0.91	0.035	96.50
Dry tropical forest	5	0.20	0.80	0.082	91.81
Sandy desert shrub	6	0.17	0.83	0.067	93.30
Halophytic vegetation	9	0.11	0.89	0.043	95.68
Mangrove	5	0.20	0.20	0.082	91.81
Bare rock or soil	6	0.17	0.83	0.067	93.30

The procedure stops under one of two circumstances; either when the first labeling error is detected, or when the sample is large enough to satisfy a specified confidence level. This value can be defined by evaluating the threshold in the increase of the confidence accumulated with an increased number of Bernoulli experiments.

Conclusions

This method is particularly useful when applications require the systematic (manual or automatic) digitizing and subsequent labeling of polygons from analog maps, a normal procedure in many working environments. The procedure is simple and straightforward. The size of the sample, that controls the cost of the testing exercise, becomes a function of the desired accuracy level for the digital database. Appropriate cost-benefit strategies are thus facilitated, and pre-defined levels of quality are insured.

Acknowledgments

We thank Ann Stewart, Carlos Valenzuela, and two anonymous reviewers for their constructive criticism. Research on which this paper is based was funded by El Colegio de la Frontera Norte, Tijuana, Mexico.

References

Anonymous, 1992. *ILWIS Operating Manual*, ITC, Enschede, The Netherlands.

Aronoff, S., 1989. *Geographic Information Systems: A Management Perspective*, WDL, Ottawa.

Bhattacharyya, G.K., and R.A. Johnson, 1977. *Statistical Concepts and Methods*, John Wiley and Sons, New York.

Burrough, P., 1986. *Principles of Geographical Information Systems for Land Resources Assessments*, Clarendon Press, Oxford.

Congalton, R.G., 1991. A review on assessing the accuracy of classifications of remote sensing data, *Remote Sensing of Environment*, 37:35–46.

Fitzgerald, R.W., and B.G. Lees, 1994. Assessing the classification accuracy of multisource remote sensing data, *Remote Sensing of Environment*, 47(3): 362–368.

Gopal, S., and C. Woodcock, 1994. Theory and methods for accuracy assessment of thematic maps using fuzzy sets, *Photogrammetric Engineering & Remote Sensing*, 60(2):181–188.

Jansen, L.L., and F.J. van der Wel, 1994. Accuracy assessment of satellite derived land-cover data, *Photogrammetric Engineering & Remote Sensing*, 60(4):419–426.

Kraus, K., 1994. Visualization of the quality of surfaces and their derivatives, *Photogrammetric Engineering & Remote Sensing*, 60(4): 457–462.

Lanter, D.P., and H. Veregin, 1992. A research paradigm for propagating error in layer-based GIS, *Photogrammetric Engineering & Remote Sensing*, 58(6):825–833.

Lunetta, R.S., R.G. Congalton, L.K. Fenstermaker, J.R. Jensen, K.C. McGwire, and L.R. Tinney, 1991. Remote sensing and geographic information system data integration: Error sources and research issues, *Photogrammetric Engineering & Remote Sensing*, 57(6):677–687.

Mood, M.A., F.A. Graybill, and D.C. Boes, 1974. *Introduction to the Theory of Statistics*, McGraw Hill, Tokyo.

Thapa, K., and J. Bossler, 1992. Accuracy of spatial data used in geographic information systems, *Photogrammetric Engineering & Remote Sensing*, 58(6):835–841.

van Genderen, J.L., and B.F. Lock, 1977. Testing land-use map accuracy, *Photogrammetric Engineering & Remote Sensing*, 43(9): 1135–1137.

van Genderen, J.L., B.F. Lock, and P.A. Vass, 1978. Remote sensing: statistical testing of thematic map accuracy, *Remote Sensing of Environment*, 7:3–14.

Veregin, H., 1994. Integration of simulation modeling and error propagation for the buffer operation in GIS, *Photogrammetric Engineering & Remote Sensing*, 60(4):427–435.

Walsh, S.J., D.R. Lightfoot, and D.R. Butler, 1987. Recognition and assessment of error in GIS, *Photogrammetric Engineering & Remote Sensing*, 53(10):1423–1430.

(Received 18 April 1995; revised and accepted 13 February 1996; revised 18 April 1996)