MANAGING ERRORS IN A GEOLOGICAL DIGITAL DATABASE

ARDEMIRIO DE BARROS SILVA[1,2], WASHINGTON DE JESUS SANT'ANNA DA FRANCA-ROCHA[2,3] AND AURÉLIO AZEVEDO BARRETO NETO[1]

ABSTRACT Before using digital dataset for geological applications it's highly advisable to evaluate the accuracy of the dataset. The inaccuracy and imprecision in spatial dataset can make the results of GIS analysis almost worthless. The evaluation of accuracy in digital dataset is carried out by several indexes such as EO, EC, GI, MI, KIA, among others. In a practical way, existing digital geological databases must be evaluated before acquired. The geological maps derived from digital images processing can also be evaluated in a quantitative way. To establish the rules to select reliable maps is one of the most important goals in a GIS project.

Keywords: GIS, accuracy, geological database

INTRODUCTION

The analysis of spatially located data is one the basic concerns of geologists and is becoming increasingly important in other fields. The need to integrate information, to consider a vast range of combinations, to underline different hypotheses and to replace tedious manual techniques, led to the formulation of Geographical Information Systems (GIS). Analysis for resources management is commonly done by means of map overlay. This procedure has been used for over a half century, and forms the basis of GIS technology. GIS has been developed independently for many purposes, including forestry, property and land parcel, transport, agriculture and environment, epidemiology, civil engineering and mineral exploration. The future of GIS will depend on better maps, on methodological application developments and continuing improvements in hardware and software.

Burrough (1986) initiated a discussion of the problems and sources of errors. Antenucci *et al*. (1991), Bonham-Carter (1994), Burrough and McDonnell (1998), Silva (1999), among others discuss the ways to keep error to a minimum through careful planning and methods for estimating its effects on GIS solutions.

Until quite recently, geologist using GIS paid little attention to the problems caused by error, inaccuracy and imprecision in geological datasets. GIS increasingly utilize these findings to perform analyses on spatial databases and to evaluate their quality. It is important to register that all data suffer from inaccuracy and imprecision, but the effects on GIS problems and solutions are not usually considered in great detail. It is now recognized that error, inaccuracy and imprecision in spatial datasets can make the results of a GIS analysis almost worthless.

This paper introduces some basic definitions such as accuracy, precision, and error that lead to establishing criteria to meet the specific demands of a geological project, setting standards for procedures and products. To achieve the objectives some indexes, such as Kappa index, Moran and Geary indexes, errors of commission and omission have to be calculated and an example is shown to emphasize the importance of accuracy evaluation in a spatial database.

THEORETICAL BASIS

It is very important to distinguish from the start the difference between accuracy, precision and error. Accuracy is the degree to which the data or information in a digital database matches true or accepted values, consequently the level of accuracy required for a particular applications varies greatly and highly accurate data can be very difficult and costly to produce and compile. Precision refers to the level of measurement and exactness independent of the existence of true values. This means that although accurate and precise data have small standard deviations the former is related to true or accepted values. Error encompasses both the imprecision of data and its inaccuracies.

Accuracy indexes in digital databases

GIS specialists have been searching for a single value (index) to represent the accuracy of maps, more than an accuracy value for each category within the map. Some authors such as Hellden (1980) and Short (1982) have attempted to find an accuracy index for individual categories which would consider errors by commission and omission.

Errors of omission (EO) represent cases where sample points of a particular category were mapped as something different. Errors of commission (EC) include cases where locations mapped as particular category were found to be truly something else. EO and EC range from 0 to 1, the closest to 0 the highest accuracy is achieved, being defined by the following equations:

$$E_0 = 1 - \frac{C_c}{C_r}$$

and

$$E_c = 1 - \frac{C_c}{C_i}$$

where, in each category, Cc is the number of cell agreements, Cr is the total number of cells in the real map and Ci is the total number of cells in the interpreted map.

For example, to judge the adequacy of the mapping the error of omission is calculated and to determine how to fix the map to increase accuracy the error of commission is calculated. Figures 1 and 2 were used as single example to show how to calculate both commission and omission errors. Table 1 shows a cross validation to be used in calculating EC and EO.

Table 1 - Cross validation table using existing data in the Figures 1 and 2.

	granites (1)	**gabbros (2)**	**arenites (3)**	**basalts (4)**	**TOTAL**	**EC**
granites (1)	9	0	0	4	13	(1 – 9/13) 0,31
gabbros (2)	4	19	0	2	25	(1 – 19/25) 0,24
arenites (3)	2	3	8	5	18	(1 – 8/18) 0,55
basalts (4)	0	0	1	7	8	(1 – 7/8) 0,12
TOTAL	15	22	9	18	64	-
EO	(1 - 9/15) **0,4**	(1 - 19/22) **0,14**	(1 – 8/9) **0,11**	(1 – 7/18) **0,61**	-	(1 – 43/64) **0,33**

Kappa Index of Agreement (KIA) (Rosenfield and Fitzpatrick-Lins 1986, Congalton 1991, Silva 1999) is defined as the proportion of agreement after chance agreement is removed:

$$KIA = \frac{P(A) - P(E)}{1 - P(E)}$$

as,

$$P(E) = \frac{Q}{C_T}$$

$$P(A) = \frac{D_R}{C_T}$$

and

$$Q = \sum_{I=1}^{n} C_R \left[\frac{C_i}{C_T} \right]$$

1 – Instituto de Geociências, Universidade Estadual de Campinas, Campinas, 13083-970,São Paulo, Brazil. E-mail: abarros@uefs.br
2 – Departamento de Ciências Exatas, Universidade Estadual de Feira de Santana, Km 03, BR-116, Feira de Santana, 44031-460, Bahia, Brazil. E-mail: wrocha@uefs.br
3 - Instituto de Geociências, Universidade Federal da Bahia. Rua Caetano Moura, 123, Salvador, 40210-340, Bahia, Brazil. E-mail: aurelio@ige.unicamp.br

Where P(A) is the proportional agreement between grid cells, P(E) is a proportional measure of the agreement by chance only; D_R corresponds to the total of cells that match with interpreted and real data, C_T corresponds to the total of examined cells, and Q is a result of weighting the chance in all categories (n is the number of categories, C_R is the real number of cells for one category and C_I is the interpreted number of cells for the same category).

Substituting equations above we derived the expression of KIA for practical purposes:

$$KIA = \frac{(D_R - Q)}{(C_T - Q)}$$

KIA varies from 0 to 1, the closest to 1 the highest accuracy is obtained. Figures 3 and 4 show interpreted and actual geological maps, respectively. Table 2 shows the cross validation agreement derived from the existing data in the Figures 3 and 4.

Figure 1 - Interpreted geological map. (1) granite; (2) gabbro; (3) arenite; (4) basalt.

1	**1**	**1**	**2**	**2**	**3**	**3**	**3**
1	**1**	**1**	**2**	**2**	**3**	**3**	**4**
1	**1**	**1**	**2**	**2**	**3**	**3**	**3**
2	**2**	**2**	**2**	**2**	**3**	**3**	**2**
2	**2**	**2**	**2**	**2**	**3**	**2**	**2**
4	**4**	**4**	**4**	**4**	**4**	**2**	**2**
3	**3**	**3**	**4**	**1**	**1**	**2**	**2**
3	**3**	**3**	**3**	**1**	**1**	**2**	**2**

Figure 2 - Actual geological map. (1) granite; (2) gabbro; (3) arenite; (4) basalt.

1	**1**	**1**	**1**	**2**	**2**	**3**	**3**
1	**1**	**1**	**1**	**2**	**2**	**3**	**3**
1	**1**	**1**	**2**	**2**	**2**	**3**	**1**
2	**2**	**2**	**2**	**2**	**3**	**1**	**1**
2	**2**	**2**	**2**	**2**	**3**	**2**	**1**
4	**4**	**4**	**4**	**4**	**4**	**2**	**2**
3	**3**	**4**	**4**	**4**	**4**	**2**	**2**
4	**4**	**4**	**4**	**4**	**4**	**4**	**4**

Figure 3 - Interpreted geological map. (1) granite; (2) gabbro; (3) arenite; (4) basalt.

1	**1**	**2**	**4**	**4**
1	**2**	**2**	**4**	**4**
1	**3**	**2**	**4**	**2**
3	**3**	**3**	**2**	**2**

Figure 4 - Actual geological map. (1) granite; (2) gabbro; (3) arenite; (4) basalt.

1	**2**	**2**	**2**	**4**
1	**3**	**2**	**2**	**2**
3	**3**	**3**	**2**	**2**
3	**3**	**3**	**3**	**2**

Statistical test for spatial dependency Spatial analysis usually requires continuous surfaces generated through point values via some gridding techniques, giving the interpreter the ability to view and manipulate the continuous distribution of discrete sampling. Errors can be found that depend on factors such as sample point density and characteristics of the spatial autocorrelation. Autocorrelation is a property that mapped data posses whenever it exhibits organized patterns. Thus, autocorrelation deals simultaneously with both location and attribute information. One of the most valuable methods to measure autocorrelation is to calculate the Geary (GI) and Moran indexes (MI). The interpretation of the GI and MI indicates if the dataset can be gridded or not. If 0<GI<1 and IM>0 the dataset is similar, regionalized and have smooth limits, therefore can be gridded. The expressions to calculate GI and MI are shown below.

Table 2 - Cross validation agreement

	1	**2**	**3**	**4**	**TOT**
1	2	1	1	0	4
2	0	4	3	0	7
3	0	0	4	0	4
4	0	4	0	1	5
TOT	2	9	8	1	11

$$GI = \left(\frac{n-1}{2\sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}} \right) \left[\frac{\sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}(x_i - x_j)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right]$$

and

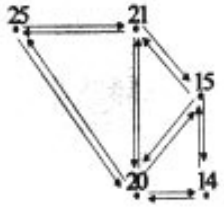
$$MI = \left(\frac{n}{\sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}} \right) \left[\frac{\sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right]$$

Where n = total of spatial data, x_i, x_j = spatial data associated with columns and rows and C_{ij} = total of binary connectivity.

Figure 5 shows how to calculate the GI and the MI from irregular points data and Figure 6 shows how to calculate the GI and the MI from regular points data using Hook case technique.

DISCUSSION Managing errors in GIS datasets is now recognized as a substantial problem that needs to be addressed in the design and use of such systems. Failure to control and manage error can limit severely or invalidate the results of a GIS analysis. Standards are not arbitrary. They should suit the demands of accuracy and precision to meet the demands of a project. No matter what the project, standards should be set from the start. Regular checks and tests should be employed through a project to make sure that standards are being followed. This may include the regular testing, such as EO, EC, MI, GI and KIA, of all data added to the dataset. GIS datasets should be checked regularly against reality, this involves checking maps and positions in the field or, at least, against sources of high quality. These statistical tests can be employed to compare true values and those recorded in the dataset. As in example of what this means, consider remote sensed images and a lithological interpretation derived from supervised classification on false color composite.

Interpreting the results of remote-sensing surveys often begins by searching the data planes for spatial correlations and anticorrelations between anomalies of different kinds. Such relationships can provide clues for building geological models. The area covered by remote-sensing surveys is, very often, large. This implies that the success of the computer-assisted interpretation of these images depends on the presence of distinctive signatures for the land covers, classes of interest in the band set being used and, also, the ability to distinguish these signatures from other spectral response patterns that may be present. Correlation and anticorrelation show up in the distribution of colors in the image. Where all three components are strongly correlated, their digital numbers contribute equally to color, and grays characterize the areas of correlation. Usually, small areas called training sites are defined, a statistical characterization is achieved for each information class and then the system classifies the image by examining the reflectance for each pixel and making a decision about which of the signatures it resembles most. Following the supervised classification, an intensive fieldwork is undertaken to build a geological map that represents the training sites. From that point, the coefficients of agreement (EO, EC, KIA) are calculated as a measure of thematic classification accuracy and the standards to be established depend on the characteristics of the GIS project. Recent works have shown that some projects use EO and EC less than 0.25 and KIA higher than 0.85, others uses EO and EC less than 0.40 and KIA higher than 0.60, the former can be related to base metals prospecting and the latter can be related to land-use assessment.



	25	21	20	15	14
25	0	1	1	0	0
21	1	0	1	1	0
20	1	1	0	1	1
15	0	1	1	0	1
14	0	0	1	1	0

x_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
25	6	36
21	2	4
20	1	1
15	-4	16
14	-5	25

x_i/x_j	25	21	20	15	14
25	0	$(25-21)^2$	$(25-20)^2$	0	0
21	$(21-25)^2$	0	$(21-20)^2$	$(21-15)^2$	0
20	$(20-25)^2$	$(20-21)^2$	0	$(20-15)^2$	$(20-14)^2$
15	0	$(15-21)^2$	$(15-20)^2$	0	$(15-14)^2$
14	0	0	$(14-20)^2$	$(14-15)^2$	0

$(x_i-\bar{x})/(x_j-\bar{x})$	6	2	1	-4	-5
6	0	6*2	6*1	0	0
2	2*6	0	2*1	2*-4	0
1	1*6	1*2	0	1*-4	1*-5
-4	0	-4*2	-4*1	0	-4*-5
-5	0	0	-5*1	-5*-4	0

$\bar{x} = 19$ $\quad n = 5$ $\quad \sum c_{ij} = 14$

$\sum (x_i - x_j)^2 = 280$ $\quad \sum (x_i - \bar{x})^2 = 82$ $\quad \sum c_{ij}(x_i - \bar{x})(x_j - \bar{x}) = 46$

$$MI = \left(\frac{5}{14}\right)\left(\frac{46}{82}\right) = 0{,}200 \qquad GI = \left(\frac{5-1}{2*14}\right)\left(\frac{280}{82}\right) = 0{,}487$$

Figure 5 - Location of spatial data, statistical analysis, binary connectivity table and GI and MI results for irregular grid

25	21	20
15	14	28
19	23	*16*

	25	21	20	15	14	28	19	23	16
25	0	1	0	1	0	0	0	0	0
21	1	0	1	0	1	0	0	0	0
20	0	1	0	0	0	1	0	0	0
15	1	0	0	0	1	0	1	0	0
14	0	1	0	1	0	1	0	1	0
28	0	0	1	0	1	0	0	0	1
19	0	0	0	1	0	0	0	1	0
23	0	0	0	0	1	0	1	0	1
16	0	0	0	0	0	1	0	1	0

x_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
25	5	25
21	1	1
20	0	0
15	-5	25
14	-6	36
26	6	36
19	-1	1
23	3	9
17	-3	9

x_i/x_j	25	21	20	15	14	26	19	23	17
25	0	$(25-21)^2$	0	$(25-15)^2$	0	0	0	0	0
21	$(21-25)^2$	0	$(21-20)^2$	0	$(21-14)^2$	0	0	0	0
20	0	$(20-21)^2$	0	0	0	$(20-26)^2$	0	0	0
15	$(15-25)^2$	0	0	0	$(15-14)^2$	0	$(15-19)^2$	0	0
14	0	$(14-21)^2$	0	$(14-15)^2$	0	$(14-26)^2$	0	$(14-23)^2$	0
26	0	0	$(26-20)^2$	0	$(26-15)^2$	0	0	0	$(26-17)^2$
19	0	0	0	$(19-15)^2$	0	0	0	$(19-23)^2$	0
23	0	0	0	0	$(23-14)^2$	0	$(23-19)^2$	0	$(23-17)^2$
17	0	0	0	0	0	$(17-26)^2$	0	$(17-23)^2$	0

$(x_i-\bar{x})/(x_j-\bar{x})$	5	1	0	-5	-6	6	-1	3	-3
5	0	5*1	0	5*-5	0	0	0	0	0
1	1*5	0	1*0	0	1*-6	0	0	0	0
0	0	0*1	0	0	0	0*6	0	0	0
-5	-5*5	0	0	0	-5*-6	0	-5*-1	0	0
-6	0	-6*1	0	-6*-5	0	-6*6	0	-6*3	0
6	0	0	6*0	0	6*-6	0	0	0	6*-3
-1	0	0	0	-1*-5	0	0	0	-1*3	0
3	0	0	0	0	3*-6	0	3*-1	0	3*-3
-3	0	0	0	0	0	-3*6	0	-3*3	0

$\bar{x} = 20$ $\quad \sum_{ij} C_{ij} = 24$ $\quad n = 9$

$\sum (x_i - \bar{x})^2 = 142$ $\quad \sum c_{ij}(x_i - \bar{x})(x_j - \bar{x}) = -144$ $\quad \sum (x_i - x_j)^2 = 611$

$$MI = \left(\frac{9}{24}\right)\left(\frac{-144}{142}\right) = -0{,}380 \qquad GI = \left(\frac{5-1}{2*14}\right)\left(\frac{280}{82}\right) = 0{,}717$$

Figure 6 - Location of spatial data, statistical analysis, binary connectivity table and GI and MI calculations for regular grid

The majority of the GIS projects make use of continuous surfaces (Digital Elevation Models, Geophysical and Geochemical grids) to be integrated with other data. First of all it is highly recommended that, before the use of this data, the GI and MI be calculated. This is because in the majority of the cases the geography of the models is vitally dependent on the chosen algorithm. Better results are achieved when the GI is close to 0 and the MI is higher then 0. For instance, in generating reliable continuous surfaces from hypsometric maps, previous works have demonstrated that GI varies from 0.10 to 0.25 and MI varies from 1.0 to 1.5.

CONCLUSIONS This paper has reviewed some procedures to assess inaccuracies in a digital geological database. Such procedures improve the performing of confidence analysis over a database in a Geographical Information System. Despite the rugged equations and mathematical details, calculating those indexes can be easily implemented by means of overlapping digital maps and building cross tables.

A very useful approach in applying accuracy measurements can be seen in lithological data collections via remote sensors. Training sites were used to associate distinct spectral signatures to lithological diversity. The interpretation derived from supervised classification must be tested to judge the quality of the map produced and this must take into account the level of accuracy measured in field control sites.

The specification of an accuracy index meets the needs of the user of map documents, who often has only qualitative statements about within-map unit variations. Different levels of accuracy are requested for different GIS applications but increasing accuracy is more cost effective, for this reason the best choice must be matched to the consequences of errors.

Acknowledgements To two anonymous referees of RBG for the critical review of the manuscript.

References

Antenucci J. C., Brown K., Croswell P. L., Kevany M. J., Archer H.1991.*Geographic Information Systems, A guide to the technology*. New York,Van Nostrand Reinhold. 301 p.

Aronoff S. 1989. *Geographic Information Systems: A management perspective*. Ottawa, WDL Publications. 294 p.

Bonham-Carter G.F. 1994. *Geographic Information Systems for Geoscientists: Modelling with GIS*. New York, Pergamon/Elsevier, 398 p.

Burrough P.A. 1986. *Principles of Geographical Informations Systems for Land Reources Assessment*. London, Oxford, 194p.

Burrough P.A. & McDonnell R. 1998. *Principles of Geographical Informations Systems*. London, Oxford, 333p.

Congalton R.G. 1991. A Review of Assessing the Accuracy of Classifications of Remotely Sensed Data. *Remote Sensing and the Environment,* **37:**35-46

Helden U. 1980. A Test of Landsat 2 imagery and digital data for thematic mapping illustrated by an environmental study in Northern Kenya. Sweden, Lund University. Natural Geographic Institute Report No 47.

Rosenfield G.H. & Fitzpatrick-Lins K. 1986. A coefficient of agreement as a measure of thematic classifications accuracy. *Photogrammetric Engineering and Remote Sensing,* **52:**223-227.

Short N.M. 1982. *The Landsat tutorical workbook - basics of satellite remote sensing*. Greenbelt , Md., Goddard Space Flight Center, NASA Reference Publication 1078.

Silva A. B. 1999. *Sistemas de Informações Geo-referenciadas: Conceitos e Fundamentos* Campinas, UNICAMP, 250 p.

Contribution IGC-195
Received March 13, 2000
Accepted for publication May 7, 2000