CONTROLE DE QUALIDADE DE BASES DE DADOS ESPACIAIS ATRAVÉS DE UMA AMOSTRAGEM DE ZERO-DEFEITOS COM RETIFICAÇÃO

Quality control of a spatial database by a zero-defect sampling with rectification procedure

José Alberto Quintanilha¹ Linda Lee Ho² Roberto C. Quinino³

¹Lab. Geoprocessamento - Depto Engenharia de Transportes – EPUSP

Av. Prof. Almeida Prado, Travessa 2, no. 83 – São Paulo – SP – Brasil - 05508-900

jaquinta@usp.br

²Depto Engenharia de Produção – EPUSP – Brazil (lindalee@usp.br)

³Depto de Estatística – ICEX- UFMG- Brazil (roberto@est.ufmg.br)

RESUMO

Qualidade é comumente usada para indicar superioridade de um bem manufaturado ou o grau de excelência de um produto, serviço ou desempenho. Uma base de dados pode ser vista como resultado de um processo de produção e a confiabilidade desta impacta o seu valor e a sua utilidade. Neste sentido, procedimentos de amostragem podem ser aplicados para avaliar se a base de dados satisfaz critérios especificados pelo usuário. Neste artigo mostra-se um procedimento para se extrair uma amostra de tamanho ótimo de um arquivo digital preparado para um sistema de informações geográficas, obtido através de um processo de conversão de dados. Dispositivos de geometria similar a um quadrado ou retângulo (quadrats) foram utilizados como unidade amostral num processo de amostragem com critério de aceitação contendo zero defeitos através de retificação. O procedimento foi implementado através do software Matlab e foi desenvolvida uma aplicação, ilustrativa, sobre dados digitais referentes a quadras de uma parte da cidade de São Paulo.

Palavras chaves: SIG, qualidade, base de dados espaciais, função custo, zero defeitos, retificação, amostragem em área.

ABSTRACT

Quality is commonly used to indicate the superiority of a manufactured good or as the degree of excellence of a product, service or performance. Since a database can be viewed as a result of a production process, and the reliability of the process imparts value and utility to the database, so sampling procedures can be applied to evaluate if the database met the specifications made by the user. In this paper, we present the optimum sample size to be extracted in a digital file generated from a conversion process. A zero-defect acceptance sampling with rectification was considered and quadrats as area sampling frames. The procedures are implemented in a program using the software Matlab and illustrated by an application to a digital data related to the blocks of a region of São Paulo downtown.

Keywords: GIS, quality, spatial database, cost function; zero-defect; rectification; area sampling.

1. INTRODUCTION

Quality is commonly used to indicate the superiority of a manufactured good or as the degree of excellence of a product, service or performance. In manufacturing processes, quality may be stated as a desirable goal to be achieved by managements and by the control of the production process (usually employing tools as control charts, for example). These same issues may be easily extended or adapted to evaluate the quality of databases, since a database can be viewed as a result of a production process, and the reliability of the process imparts value and utility to the database.

In manufacturing, the characteristics to be evaluated are easily identified and usually classified in two main groups: attributes (conforming or nonconforming) or variables (some measurement of interest). In data quality, users are faced by some problems: what are the dimensions of geographical data quality since features of the real world are represented in the database by objects, points, lines, polygon or areas (for example, rivers or roads are represented by lines). According to VEREGIN (1999), the conventional view is that geographical data is "spatial". The terms "geographical data" and "spatial data" have been used interchangeably. However, this approach is not adequate since it ignores the inherent coupling of space and time (geographical entities are actually events unfolding over space and time) and geography is connected by themes (not space). Space (or space-time) is just the framework inside which theme is measured. In the absence of theme, only geometry is present. So a better definition of geographical data may include the three dimensions: space, time and theme (where-when-what). These three dimensions are the basis for all geographical observation and data quality must concern on them by components as accuracy; precision; consistency; completeness.

To evaluate the quality of digital products is not an easy task and different aspects of the quality of a spatial (sometimes cartographic) database have been discussed in the literature. Some contributions may be listed. For example, REINGRUBER and GREGORY (1994); CHENGALUR-SMITH; BALLOU and PAZER (1999) have pointed out the influence of the spatial database quality on the decision process. Control cartographic objects in a quality evaluation of spatial database process were subjects of interest. See for example: LEUNG and YANG (1998); SHI and LIU (2000) and VEREGIN (1999 and 2000). Related to spatial database building process, the next contributions may be listed: COUCLELIS (1992); NUGENT (1995); LIU, SHI and TONG (1999), QUINTANILHA (2002), QUINTANILHA and HO (2002).

Consider a situation that a digital file design to a spatial database is generated from a conversion process (for example documents or maps or some others cartographical products in paper format and converted to a digital file). This file will be used in a geographical information application and it is necessary to evaluate it if the specifications (for example specification limits and restrictions for spatial features, attribute values considerations and other relevant aspects) settled by the users are met.

Similar to the evaluation of manufacturer process, a sample of database is randomly selected using some area sampling frame (as we are dealing with spatial data, quadrats are the most common frame). Each sampling unit is evaluated to verify if it satisfies criteria previously fixed. A rule is chosen to decide if the database meets the specification or not. In this paper we will consider the following acceptance sampling scheme:

1 -Consider an area covered by T sheets in a fixed scale. Each sheet can be divided in *n* independent quadrats [see: KISH, (1965); SHAW and WHEELER, (1985)] of a fixed format (in our case, a square) and size.

2 – A random sample of m < n quadrats is extracted from each sheet.

3- The subset of files corresponding to the *m* quadrats are examined and if all information in each file are conforming, then the examined sheet is accepted; otherwise all *n* quadrats of the sheet are inspected, corrected and then the file of the examined sheet is accepted.

Figure 1 illustrates the described sampling procedure. Such sampling scheme is known as zerodefect with rectification and it is usually used to evaluate high quality manufactured processes by attributes. In those processes, we have batches instead of sheets or cartographical products and items or products are examined in place of files related to the quadrats.

In technical literature some papers about zerodefect with rectification can be found. We may mention the contributions from HAHN (1986), BRUSH; HOADLEY and SAPERSTEIN (1990), GREENBERG and STOKES (1992, 1995) and ANDERSON; GREENBERG and STOKES (2001). In those papers, the main objective is to present estimator for the number of non-conforming items in an accepted batch (here we have non-conforming features in an accepted sheet).

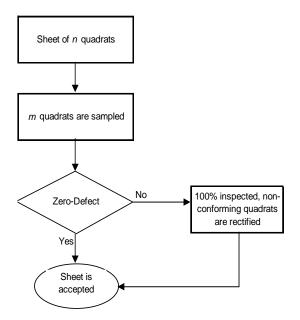


Fig. 1: Acceptance sampling: zero defect with rectification.

In ANDERSON; GREENBERG and STOKES (2001), they introduced the possibility of the classification criteria presents diagnosis errors in zero-defect with rectification procedure. That is, an examined item/product is classified as non-conforming but in reality it is conforming or an item/product can be classified as conforming but it is non-conforming.

Similarly, when we are evaluating a spatial database, a subset of file related to a quadrat is examined and classified as non-conforming but in reality it is conforming or that subset can be classified as conforming but it is non-conforming. [More details about diagnosis errors, see, JOHNSON; KOTZ and WU, (1991)]. Such diagnosis errors can occur either in the inspection or in the rectification stages. MARKOWSKI and MARKOWSKI (2002) presented a methodology to minimize the impact of such diagnosis errors in the acceptance sampling.

However before extracting the sample of files, it is important to design how large must be the sample in order to meet some criteria (statistical and/or economical ones). In this paper, we will consider the determination of an optimum sample size m such that minimizes a cost function. The components of such function include the inspection cost, the costs due to the presence of non-conforming quadrats subset of files in accepted sheets and the costs due to diagnosis errors. Economical models found in the literature do not include the possibility of the diagnosis errors in the inspection stage. The determination of the sample size including such errors in the inspection stage is the focus of this paper.

In Section 2, we introduce the notation and hypothesis considered in this paper. The expected cost function is developed in Section 3 and such procedure is illustrated by a numerical sample in Section 4. We finish this paper with discussions and extensions in future works.

2. NOTATION AND HYPOTHESIS

Consider an area covered by a sheet. This sheet can be divided in *n* independent quadrats of a fixed format and size. A random sample of *m* quadrats is selected and *p* is the probability of a quadrat be non-conforming. The value of *p* is equal to zero with probability $(1-\pi)$ or it can vary from one sheet to another sheet according to a Beta distribution (a, b) with probability π (the probability of p > 0). Let:

 $e_1 \rightarrow$ the probability of a quadrat from a subset of file be wrongly classified as non-conforming when it is conforming;

 $e_2 \rightarrow$ the probability of a quadrat from a subset of file be wrongly classified as conforming when it is non-conforming;

 $c_0 \rightarrow$ the cost to inspect a quadrat from a subset of file ;

 $c_1 \rightarrow$ the cost of a non-conforming and non-rectified quadrat subset of file in an accepted sheet;

 $c_2 \rightarrow$ the cost to judge erroneously a quadrat from a subset of file as conforming when it is non-conforming;

 $D_{1i} \rightarrow$ the number of non-conforming quadrats from subsets of files in a sample of size *m* in the sheet *i*;

 $D_{2i} \rightarrow$ the number of non-conforming quadrats from subsets of files in (n-m) non-sampled quadrats in sheet *i*;

 $D_i = D_{1i} + D_{2i} \rightarrow$ the number of nonconforming quadrats from subsets of files in sheet *i*;

 $Y_{1i} \rightarrow$ the number of non-conforming quadrats from subsets of files observed in a sample of size *m* in the sheet *i*;

 $Y_{2i} \rightarrow$ the number of non-conforming quadrats from subsets of files observed in (*n*-*m*) non-sampled quadrats in the sheet *i*;

 $Y_i = Y_{1i} + Y_{2i} \rightarrow$ the number of nonconforming quadrats from subsets of files observed in the sheet *i*.

3. COST FUNCTION AND DETERMINATION OF THE OPTIMUM SAMPLE SIZE *m*

In this Section, an expected cost function per sheet (E_m) is developed employing the earlier notations and hypothesis. The total medium cost to evaluate T sheets is TE_m . And *m* such that minimizes E_m will also minimize the total medium cost. So hereafter, the index *i* will be is suppressed in the expression of the expected cost function per sheet. The expected cost function E_m is compounded by three

parts:
$$E_m = E_m^1 + E_m^2 + E_m^3$$
. (1)

The first one (E_m^1) is related to inspection cost. It is compounded by costs to inspect *m* quadrats and the possibility to inspect the (n-m) non-sampled quadrats. Such factor is conditioned to the presence of at least one non-conforming quadrat in the initial inspection of *m* quadrats. So E_m^1 is given by

$$E_m^1 = c_0 m + c_0 (n-m)U$$

where $U = P(Y_1 > 0) = 1 - P(Y_1 = 0)$.

To obtain the value of 1-U, we have to consider two scenarios:

1 – In the random sample of m quadrats, all are conforming and they must be correctly classified as conforming and the probability of this event is given by:

$$(1-\pi)(1-e_1)^m$$
 (2)

2 – In *m* examined quadrats, D_1 are nonconforming quadrats, but all of *m* quadrats must be classified as conforming. The probability of this event conditioned on fixed values of *p* and D_1 is given by

$$\pi \binom{m}{D_1} p^{D_1} (1-p)^{m-D_1} (1-e_1)^{m-D_1} e_2^{D_1}$$

Unconditioned the above expression for all values of p and D_1 , it results

$$\sum_{D_{1}=0}^{m} \int_{0}^{1} {m \choose D_{1}} p^{D_{1}} (1-p)^{m-1} \frac{p^{a} (1-p)^{b}}{\beta (a,b)} (1-e_{1})^{m-D_{1}} e_{2}^{D_{1}} dp$$
$$= \pi \sum_{D_{1}=0}^{m} {m \choose D_{1i}} \frac{Beta(a+D_{1};m-D_{1}+b)}{Beta(a,b)} ((1-e_{1})^{m-D_{1}} e_{2}^{D_{1}})$$
(3)

Summing up equations (2) and (3), we have the probability of 1 - U.

The second component (E_m^2) in (1) is due to the possibility of a quadrat be classified as conforming but in reality it is a non-conforming one and it is given by.

$$E_m^2 = c_1 E \bigg[I_{[Y_1=0]} D + e_2 I_{[Y_1>0]} D \bigg]$$

where $I_{[\bullet]}$ denotes an indicator function and $E(\bullet) \rightarrow$ the expected value of a random variable. Such result can produce alteration in the expenses when the sheet is accepted or rejected in the inspection stage. As $D=D_1+D_2$, the above expression can be written as

$$E_m^2(a, b, \pi, e_1, e_2) = c_1 E[D] - c_1(1 - e_2) E[I_{[Y_1 > 0]}(D_1 + D_2)]$$

= $c_1 E[D] - c_1(1 - e_2) E[I_{[Y_1 > 0]}D_1] - c_1(1 - e_2) E[I_{[Y_1 > 0]}D_2]$

where

$$\begin{split} & E \Big[I_{[Y_{li}>0]} D_{l} \Big] = \\ & \pi \sum_{D_{l}=0}^{m} \binom{m}{D_{l}} \frac{Beta(a+D_{l};m-D_{l}+b)}{Beta(a,b)} (1 - (1 - e_{l})^{m-D_{l}} e_{2}^{D_{l}i}) D_{l} \end{split}$$

and

$$E\left[I_{[Y_1>0]}D_2\right] = \pi(n-m)\sum_{D_1=0}^m \binom{m}{D_1} \frac{Beta(a+D_1;m-D_1+b)}{Beta(a,b)} (1-(1-e_1)^{m-D_1}e_2^{D_1})$$

The last part (E_m^3) in (1) is due to the consequence in classifying a quadrat as non-conforming when it is a conforming one. In this case, the sheet is rejected and consequently all quadrats are classified as non-conforming but there is a possibility to be rectified them unnecessarily. After this introduction the components of such cost function per sheet are:

$$E_{m}^{3} = c_{2}e_{1}E\left[(n - D_{i})I_{[Y_{1} > 0]}\right]$$

= $c_{2}e_{1}nU - c_{2}e_{1}E\left[I_{[Y_{1} > 0]}D_{1}\right] - c_{2}e_{1}E\left[I_{[Y_{1} > 0]}D_{2}\right]$
(5)

The optimum value of m (m°) is such that minimize E_m and can be obtained by direct search substituting values of m = 1, ..., n in E_m .

4. NUMERICAL EXAMPLE

The example described in this section is based on an application to a digital data related to the blocks of a small region of São Paulo downtown, Brazil. The attribute of the interest was to verify if the presence/absence of block drafts were or not correctly located.

It is known that the area recovered by sheets and each one is made up by n=5000 quadrats. They will be inspected by a zero-defect with rectification procedure and the inspection consists of checking visually the presence or absence of block drafts on the screen or by plot. In this context it is reasonable the occurrence of misclassifications.

Let us consider the following costs: $c_0 = \$1.00$; $c_1 = \$100.00$; $c_2 = \$500.00$. According to the user's experience, the diagnosis errors were at most 0.1% (that is: $e_1 = e_2 = 0.001$) and the *p* is equal to zero with probability 0.90 (that is: $\pi = 0.1$) or it follows a Beta distribution with a = 0.335, b = 3.01.

The goal is to find the optimum value of m (m°) such that minimizes E_m . A program using the software Matlab was developed to find the optimum value m° (Appendix 1). Such program provides us the optimum sample size (m°) equal to 26 which

(4)

corresponds an expected cost of \$847.89 per sheet. In the absence of diagnosis errors the optimum sample increased to $m^{\circ} = 61$ which corresponds to a decreased expected cost of \$472.70 per sheet. Note that even small diagnosis errors can alter significantly the expected cost as also the optimum sample sizes. Figure 2 illustrates the behavior of E_m in function of m.

For illustration purpose, the optimum sample size was also obtained for other values of c_1 =5.00; 10.00 and 20.00. The other parameters are remained constant. The results are in Table 1.

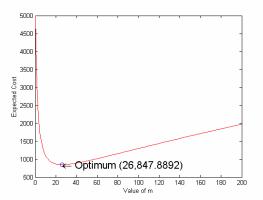


Fig. 2: Values of m versus expected cost $(c_0=1.0; c_1=100.00 \text{ and } c_2=500.00).$

TABLE 1 - OPTIMUM SAMPLE SIZE AS A

FUNCTION	OF C1
TUNCTION	$OI \ OI$

	With diagnosis errors		No diagnosis errors	
Cost c_1	m^0	cost	m^0	cost
5	1	248.85	3	216.19
10	4	380.79	8	285.31
20	8	507.08	18	343.38
100	26	847.89	62	472.71

5. CONCLUSIONS AND FINAL REMARKS

Diagnosis errors can cause a significant impact in determining the optimum sample size in a zero-defect with rectification procedure. As illustrated in this study, even small diagnosis errors as $e_1=e_2=0.001$ can alter significantly the value of the optimum m (m⁰). In this sense, it is important to incorporate them in the modeling and evaluated them in an economic perspective.

Extensions of this study can be made in two directions. One is to change the initial criteria in the sampling inspection for a limit other than zero, that is, k > 0. Another alternative is the possibility to examine the quadrats independently and repetitively to decrease the impact of the diagnosis errors. In this case, a quadrat is classified as conforming if the number of the

conforming classifications after the repetitive tests is higher than a specified integer a. In this scenery, the objectives are to determine the optimum values of: sample size m^0 , the criteria a and the limit k such that minimize the total expected cost.

ACKNOWLEDGMENTS

The authors would like to acknowledge the anonymous referees for suggestions, FAPESP (Proc. No. 2002/11996-5) and CNPq for partial financial support.

REFERENCES

ANDERSON, M. A.; GREENBERG, B. S. & STOKES, S. L. Acceptance sampling with rectification when inspection errors are present, **Journal of Quality Technology**, v. 33 (4), p.493-505, 2001

BRUSH, G. G., HOADLEY, B. & SAPERSTEIN, B. Estimating outgoing quality using the quality measurement plan, **Technometrics**, v. 32, p.31-41, 1990.

CHENGALUR-SMITH, I. N.; BALLOU, D. P. & PAZER, H. L. The impact of data quality information on decision making: an exploratory analysis. **IEEE Transactions on Knowledge and Data Engineering**, v.11 (6), p.853-864, 1999.

COUCLELIS, H. Geographic knowledge production through GIS: towards a model for quality monitoring. In: **Two perspectives on data quality**, Santa Barbara, National Center for Geographical Information and Analysis, NCGIA Technical Report 92-12, 1992.

GREENBERG, B. S. & STOKES, S. L., Estimating nonconformance rates after zero defect sampling with rectification, **Technometrics**, v. 34 (2), p.03-13, 1992.

GREENBERG, B. S. & STOKES, S. L., Repetitive testing in the presence of inspection errors, **Technometrics**, v.37 (1), p.102-111, 1995.

HAHN, G. J., Estimating the percent nonconforming in the accepted product after zero defect sampling. **Journal of Quality Technology**, v.18 (3), p.182-188, 1986.

JOHNSON, N. L., KOTZ, S.; & WU, X., **Inspection** errors for attributes in quality control, Chapman and Hall, London, 1991

KISH, L. **Survey sampling**. John Wiley: New York, Chapter 9: Area sampling, p.301-358, 1965.

LEUNG, Y. & YAN, J., A location error model for spatial features, **International Journal of Geographical Information Science**, 12 (6), p.607-620, 1998.

LIU, D. J.; SHI, W. Z. & TONG, X. H., Accuracy analysis and quality control of spatial data in GIS. Shanghai Science and Technology Documentation Press, 1999.

MARKOWSKI, E.P. & MARKOWSKI, C.A., Improved attribute acceptance sampling plans in the presence of misclassification error, **European Journal** of Operation Research, v.139 (3), p.501-510, 2002.

NUGENT, J. L., Quality control techniques for a GIS database development project, **Photogrammetric Engineering and Remote Sensing**, v.61 (5), p.523-527, 1995.

QUINTANILHA, J. A. Qualidade em bases de dados para sistemas de informações geográficas **In: Dados geográficos: aspectos e perspectivas** Salvador, Bahia: Quarteto Editora, (in Portuguese), 2002.

QUINTANILHA, J. A. & HO, L. L. Estimating the average number of errors to evaluate the quality of the spatial database from a conversion process In: 5th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences, 2002, Melbourne, **Proceedings** of Accuracy 2002. p.397-403, 2002.

REINGRUBER, M.; & GREGORY, W.W., The data modeling handbook: a best-practice approach to building quality data models, John Wiley and Sons, New York, 1994.

SHAW, G. & WHEELER, D., **Statistical techniques in geography**. S.I.: Spatial indices and pattern analysis, John Wiley and Sons, New York, 1985.

SHI, W. Z. & LIU, W. B., A stochastic process-based model for the positional error of line segments in GIS, **International Journal of Geographical Information Science**, v.14 (1), p.51-66, 2000.

VEREGIN, H., Line simplification, geometric distortion and positional error, **Cartographica**, v.36 (1), p. 25-39, 1999.

VEREGIN, H., Quantifying positional error induced by line simplification,. **International Journal of Geographical Information Science**, v.14 (2), p. 113-130, 2000.

APPENDIX

tic;

c00=1; c11=5; c22=500; pi=0.1; a=0.334782250677; b=3.01304284135; n=5000; e1=0.001; e2=0.001; %_____ p1=1-e1; p1l=e2; c1 = c00;c2=c11; c3=c22: mlim=200;custo=inf; motimo=0: Inbetaab=betaln(a,b); m=0; s1=0; p=0; s2=nchoosek(m,0)*beta(a+1,m+b)/beta(a,b)*(1-(1-b))*(1p11)^m); p=1-(pi*p+(1-p11)^m*((1pi)+pi*exp(betaln(a,m+b)-lnbetaab))); f=c1*m+c1*(n-m)*p+c2*n*a/(a+b)*pic2*pi*p1*s1-c2*pi*p1*(n-m)*s2+c3*n*p*p11c3*p11*pi*s1-c3*p11*pi*(n-m)*s2; vetorf(1)=f; vetorm(1)=m; for m=1:mlim s1=0: p=0; s2=nchoosek(m,0)*beta(a+1,m+b)/beta(a,b)*(1-(1-b))*(1p11)^m); vdi=1:1:(m); vm=(m)*ones(1,m);vtemp=log(1-(1-p11).^(vm-vdi).*(1-p1).^vdi); vtemp1=(betaln(a+vdi,vm-vdi+b)-lnbetaab); clear vcomb; if mod(m,2) == 0metade=m/2-1; else metade=floor(m/2); end for di=1:metade vcomb(di)=logcomb(m,di); end if m==1 vcomb=[0]; elseif m==2 vcomb=[log(2) 0]; elseif floor(m/2) = (m/2);vcomb=[vcomb logcomb(m,metade+1) fliplr(vcomb) 0]; else vcomb=[vcomb fliplr(vcomb) 0]; end temp2=exp(vcomb+vtemp1);

Revista Brasileira de Cartografia Nº 60/01, Abril 2008. (ISSN 1808-0936)

```
p=p+sum(temp2.*((1-p11).^(vm-vdi).*(1-p1).^vdi));
  s1=s1+sum(exp(vcomb+vtemp1+vtemp+log(vdi)));
  s2=s2+sum(exp(vcomb+(betaln(a+vdi+1,vm-
vdi+b)-lnbetaab)+vtemp));
  p=1-(pi*p+(1-p11)^m*((1-
pi)+pi*exp(betaln(a,m+b)-lnbetaab)));
  f=c1*m+c1*(n-m)*p+c2*n*a/(a+b)*pi-
c2*pi*p1*s1-c2*pi*p1*(n-m)*s2+c3*n*p*p11-
c3*p11*pi*s1-c3*p11*pi*(n-m)*s2;
  vetorf(m+1)=f;
  vetorm(m+1)=m;
end
  [custo,motimo]=min(vetorf);
  motimo=motimo-1
  plot(vetorm,vetorf,'r-');
  hold on
  plot(motimo,custo,'bo')
  hold on
  text(motimo,custo,strcat(
                             '\leftarrow'.'
                                            Otimo
(',num2str(motimo),',',num2str(custo),')'),'HorizontalAli
gnment','left','FontSize',16)
fprintf('*****************************Solucao*********
*********************************/n')
  fprintf('M otimo: %6.0f \n',motimo)
  fprintf('Custo: %10.2f \n',custo)
  fprintf('\n')
  fprintf('Maior m pesquisado: %6.0f \n',mlim)
  fprintf('Elapsed
                    time:
                             %6.2f
                                           (%4.2f
                                      s
min)\n',toc,toc/60)
********************************/n')
```