STATISTICAL TEST FOR EVALUATION OF THE ACCURACY OF DIGITAL MAPS FOR GEO-SPATIAL INFORMATION SYSTEM

Prof. Bock-Mo Yeu
Dept. of Civil Engineering, Yonsei University

Prof. Hyun Kwon
Dept. of Civil Engineering, Suwon Technical University

Researcher Seok-Kun Lee
Dept. of Civil Engineering, Yonsei University

Researcher Dong-Bin Shin
Korea Research Institute for Human Settlements

Commission II, Working Group 5

KEY WORDS: Statistics, Production, Accuracy , Quality, Project, Geo_Spatial Databases

ABSTRACT

Digital maps of Geo-Spatial Information System are very important as they are the foundation of the system and the necessity to implement and develop digital map database is increasing . But because of various different needs and requirement of different users, each has different needs of accuracy. A standard for statistical test for accuracy evaluation such as the hypothesis test is very essential under such changing environment. In this study a statistical method for evaluating accuracy of digitized maps is introduced . Using hypergeometric distribution and considering consumer risk and producer risk, a statistical method was used to decide the optimal sample size.

1. INTRODUCTION

In Korea, with the development and growing interest of Geo-Spatial Information System, the need for standards in digital maps are being discussed and frameworks are being established by government bodies such as the National Geographic Institute. Meanwhile maps are being digitized, trying to satisfy the standard as much as possible. It is well known that the database implementation cost comprises most of any GSIS projects and reports have been made that some amounts up to 80% of the whole project implementation cost.(K.Thaph, J.Bossler, 1992). Big cost for initial production also implies big cost for editing and correction of erroneous data and sometimes irreparable damages. All these means that everything has to start with the correct data from the initial stage. Despite this emphasis on the importance of accurate data, the importance of the quality of data from the producer and the consumer point of view is not getting enough attention.

Topographic base maps are directly produced in digital format nowadays by using analytical stereo plotters and test points are selected from the produced maps and compared to ground surveying results and residuals are computed and used as an evaluation factor for the map. Although effective, due to extra costs involved in the testing procedure producers tend to look for other methods of evaluation. Also researches are done recently to evaluate the classification accuracy of thematic maps by statistical evaluation and sampling method using the binomial distribution, considering the risk of both the consumers and producers (Michael E, G., 1979) (G.H. Rosenfield et. al., 1982) (S.Aronoff, 1982). Despite these efforts more researches should be done on the evaluation of accuracy of digitized maps and statistical testing of these results. This study aims to introduce a method of determining the number of sample map sheets required to evaluate the accuracy of digital maps considering consumer and producer constraints.

2. STATISTICAL EVALUATION OF DIGITIZED MAPS

2.1 Producer risk and consumer risk

The accuracy of all the maps can be assumed as one hypothesis. The testing of the hypothesis involves determining whether the digital map is accurate or not but because the position of the producer and consumer is different, different results are obtained depending on whose position is considered when establishing the hypothesis.

The hypothesis from the position of the producer

H_o : The accuracy of the digitized maps are as accurate or more accurate than the requirement.

H_1 : The accuracy of the digitized maps are less accurate than the requirement

If Q_0 is the required accuracy, the following equations can be made.

$$H_0 : Q \geq Q_0$$

$$H_1 : Q < Q_0$$

If the test is made with 5% test level , i.e. 95% confidence level, this means that if the result is not included in the left 0.05 part of the of Fig. 2.1, the test is accepted. But from the position of the consumer the following equations can be made.

$$H_0 : Q < Q_0$$

$$H_1 : Q \leq Q_0$$

And the test is accepted if the result is not included in the right 0.05 part of the Fig. 2.1.

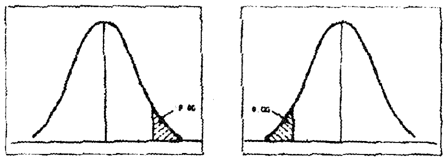

Fig 2.1 Test of the left end and right end

Even though the required accuracy is the same the result can be different depending on the position when the hypothesis is established. If the minimum required accuracy is Q_L, the hypothesis is established from the position of the consumer and this hypothesis is of error type 1, i.e. the probability that the H_0 is not rejected despite the map is less accurate is 0.05 and this is the risk of the consumer. From the producer position, it is advantageous that the probability of Q_H which is higher than Q_L being rejected be minimized. This is error of type 2, this means minimizing the probability of rejecting H_1 despite being accurate (producer risk) and this can be achieved by increasing the consumer risk or by increasing the number of samples.

Illustrating this with figures, Q_L is fixed to the minimum required map accuracy and the hypothesis H_0 is accepted as long as the probability that accuracy is lower than Q_L is not less than 5%. The critical level Q_T is computed with the assumption that map accuracy has the accuracy of Q_L and as the probability that Q_T has more that 5% of the samples. If the ratio of accurate samples is larger than or is equal to Q_T, H_0 is rejected and the map accuracy is accepted as being more accurate or is equal to than Q_L.

The producer risk is the probability that the maps are not selected, i.e the probability of not being selected despite $Q_H > Q_L$ and this is shown as the region smaller than Q_T. This can be seen from the first graph of Fig. 2.2, when the number of samples is small, the risk of the producer is more than 50%, and this means that it is difficult for the map to be evaluated as being accurate. This risk can be reduced by increasing the number of samples as in the second graph of Fig. 2.2. Increasing the number of samples also means increasing the cost of production, so an optimum level should be selected.

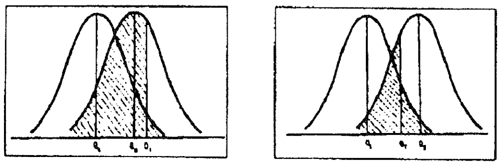

Fig 2.2 Producer risk and consumer risk

2.2 Determining number of samples using hypergeometric distribution

Hypergeometric distribution is widely used for quality control purposes and is the probability model for a fixed amount of observations with two values. If number of maps satisfying the required accuracy is M and inaccurate maps are defined as D, the probability f(x) that x amount of accurate maps are selected from a sample of n is as follows.

$$f(x) = {}_{N}C_{I} \cdot {}_{D}C_{n-x} / {}_{N+D}C_{n}$$

where $x = 0,1,2,\ldots,n$ and $x < n$ and $n < M$

Applying this to the consumer and producer risk, the number of maps satisfying the minimum required accuracy Q_L when n samples are selected from N number of maps is as follows.

$$M = N.Q_L$$

And the number of maps satisfying the required accuracy is as follows.

$$D = N - M$$

Consumer risk, i.e. the probability from a 95% confidence test that the number of inaccurate maps (n_y) from sample maps (n), can be defined as follows.

$$CR = [\Sigma \ ({}_{M}C_{n-y} \cdot {}_{D}C_{y}) / ({}_{N}C_{n})] \leq 0.05 \quad (\text{for } y = 0 \text{ to } n_y)$$

If the consumer risk is 5%, the producer risk will depend on the whole number of maps N and the size of sample n and is defined as follows.

$$PR = [\Sigma \ ({}_{M}C_{n-y} \cdot {}_{D}C_{y}) / ({}_{N}C_{n})] \quad (\text{for } y = n_y + 1 \text{ to } D)$$

From this equation the optimum size of samples considering producer risk can be known when the consumer risk is fixed.

3. EXPERIMENTS

In some recent data conversion projects, standard sampling methods were suggested with superimposition tests for 10 - 20 % of the whole number of maps. The following is an application of the aforementioned method of considering both the consumer risk and producer risk using the hypergeometric distribution.

If the consumer risk is fixed at 5%, the minimum required accuracy can be known invariable to the size of the sample and from the producer position, an appropriate size of sample can be determined at an appropriate producer risk.

From Fig. 3.1, it can be seen that generally, if 90% for minimum required accuracy, 95% for digitizing level and 30 % for the producer risk are selected, the appropriate size of samples are 83 for 100, 104 for 200 and 121 for 500. As can be seen the size of sample decreases relatively as the number of whole maps increase. For a size of 100 sheets, the sample size is about 80% but for 500 sheets the sample size is only 25% to make the same test.

From the consumer position, 90% is selected for minimum required accuracy and 97% or 99% is selected for digitizing level, and 1 map sheet is the criteria for not passing the test from a sample size of 47 from the total 100 sheets, as can be seen from Fig. 3.1, the producer risk is 21.8% which makes it difficult to pass the test is the overall quality is not high. It can be assumed that the both the consumer and the producer can come to a agreement on the size of the sample referring to this graph.

no.of map	consumer	producer
15	0.109	
14	1.055	
13	4.542	
12	11.561	
11	19.457	
10	22.939	
9	19.577	
8	12.324	
7	5.777	
6	2.02	
5	0.525	0
4	0.1	9.182
3	0.014	27.983
2	0.001	33.886
1	0	20.835
0	0	6.872

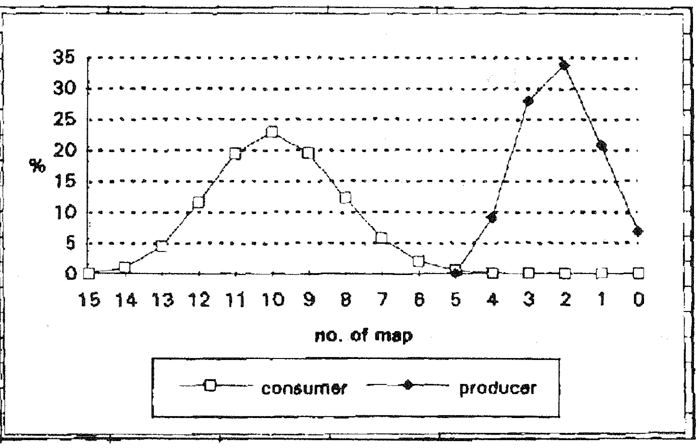

Fig. 3.1 Sample size considering consumer risk and producer risk

4. CONCLUSION

The following conclusion can be made through this study on the evaluation of digitized maps of Geo-Spatial Information System.

1) An appropriate level of accuracy can be evaluated through a statistical test considering both the consumer and producer risk.

2) The hypergeometric distribution is suitable for the determination of the sample size and the ratio of sample size decreases with the increase of the total number of sheets.

5. REFERENCES

1. K.Thaph, J.Bossler, Accuracy of Spatial Data Used in Geographic Information System, PE&RS, VOL.48, No. 6, 1992, pp. 835-841
2. Michael E, G., Testing Land-Use Map Accuracy: Another Look, PE&RS, Vol. 45, No. 10, 1979, pp. 1371-1377.
3. G.H. Rosenfield et. al., Sampling for Thematic Map Accuracy Testing, PE&RS, Vol., 48., No. 1, 1982, pp.131-137.
4. S.Aronoff, Classification Accuracy: A User Approach, PE&RS, Vol. 48, No. 8, 1982, pp. 1309-1313.