CALIBRATING AREA ESTIMATE BIAS ON CATEGORICAL MAPS USING THE CONTINGENCY TABLE

D.M. Chen[a, *] and M. Goodchild[b]

[a]Department of Geography, Queen's University, Kingston, ON K7L 3N6, CANADA – chendm@post.queensu.ca
[b]Department of Geography, University of California, Santa Barbara, CA 93106-4060, USA – good@geog.ucsb.edu

KEY WORDS: Categorical map, Area estimation, Quantitative error, Sampling, Contingency table, Calibration model

ABSTRACT:

This paper presents and compares approaches of estimating true area on the ground and calibrating quantitative errors of area estimate on categorical maps from the contingency table. Results directly estimated from the contingency table and those from two calibration methods were compared on two maps of 10 different land cover classes with known errors between them. The estimated true area percentage from the contingency table and two calibration approaches showed obvious improvement when compared with uncalibrated values. However, there is no significant difference among the estimates from the contingency table and the two calibration methods. Although the inverse method led to mean estimates closer to the true values for all classes than other methods, comparing the individual area estimates for each class showed that the inverse method did not always produce the most accurate estimate. Homogeneous classes with high classification accuracy have a better chance of achieving more accurate estimates from calibration than heterogeneous classes. Compared with large classes, classes covering a small percentage of a map are more vulnerable to the quantitative error and more sensitive to sampling error.

1. INTRODUCTION

The categorical map (or data) with nominal classes generated from remotely sensed data or other data sources is one of the major map types stored in GIS. The categorical map is often used to visualize and calculate how much area each category takes in the map region. Often these area estimates taken from the categorical maps are treated as unbiased estimates of the true area for each category and are used for various types of resource management or input of other quantitative models or applications (Congalton and Green 1999, Lunetta and Lyon 2004). However, the categorical maps stored and used in GIS are never error-free. In the evaluation of errors in a categorical map, two types are usually distinguished: quantification error and location error (Pontius 2000). Quantification error summarizes how the area percentage of each category on the map is different from the area percentage of each category in reality, while location error occurs when the classes do not occur in the correct locations, whether or not total areas are correct. With the growing attention to the error and uncertainty issue in GIS and remote sensing, many efforts have been made to measure, model and visualize location and quantification errors of categorical maps (such as Goodchild et al. 1992, Pontius 2000, and Goodchild 2003). In this paper, we focus on quantification errors in categorical maps, where their error is identified by the total areas in which categories are misclassified.

The contingency table (also called the error matrix or the confusion matrix) is a common and effective way to represent quantification errors for categorical maps, and is usually the first step used to evaluate the accuracy of categorical maps, especially for maps generated from remotely sensed data (Congalton and Green 1999, Jensen 1996). The contingency table is generated by comparing the ground truth of selected samples with their classes on a map. Depending on the representation format of maps, the sampling unit can vary. For a raster map generated from remotely sensed data, the samples are a number of sampling pixels based on a sampling strategy, while the samples would be selected points or polygons in a vector map. The accuracy of the results from these samples is then extrapolated to the entire map. The contingency table the entire map. The contingency table records the comparison results in a square array of numbers set out in rows and columns that express the number (or percentage) of samples assigned to one category in the map relative to one category in reality.

Information on the contingency table can be used to compute other accuracy measurement indexes and update the area estimates of map categories (Lewis and Brown, 2001). Several calibration-based models (such as Tenenbein 1972, Card 1982, Grassia and Sundberg 1982) have been developed to improve the area estimate accuracy in the statistical literature and have been applied to remote sensing applications. In general, there are two classes of statistical calibration methods (the classical model and the inverse model) based on linear algebraic equations to treat quantification error using the information from a contingency table (the details are explained in following section). Previous studies (Prisley and Smith 1987,Czaplewski and Catts 1991, and Walsh and Burk 1993) have employed these methods to calibrate area estimators for misclassification errors in remote sensing.

The objective of this paper is to illustrate different methods to calibrate quantitative errors on categorical maps, and compare the various calibration methods by emphasizing the relationship between sampling error and accuracy of area estimate from calibration. Two multivariate calibration approaches are reviewed. An empirical study was conducted by a Monto Carlo simulation to generate random samples. The area estimate directly from the sampling and two calibration methods were compared, based on results generated from maps with known errors. The paper concludes with a discussion of results and applications of the calibration methods.

2. CALIBRATION METHODS

Two calibration methods have been developed to calibrate the area estimate difference by using misclassification probabilities from a contingency table generated from samples. The following explains the principles and steps involved in these two methods.

The first method is known as the "inverse" or "inverse prediction" estimator (Czaplewski, 1991). For any pixel of class *i* on the classified map, the conditional probability that it is classified as class *i* on the ground is P_{ii}/P_{i+}, and the conditional probability that other class *j (j=1, ..., k,* and $j \neq i$*)* on the map is classified to class *i* on the ground is P_{ji}/P_{j+}. So the percentage of pixels on the map classified as class *i* on the ground is $(P_{ii}/P_{i+}) * AM_i$, and the total percentage of other classes on the map that is misclassified to class *i* on the ground is $\sum_{j=1, j\neq i}^{k} (\frac{P_{ji}}{P_{j+}} * AM_j)$. So, the percentage of any class *i* on the ground AG_i is the sum of both and can be calibrated as:

$$AG_i = \frac{P_{ii}}{P_{i+}} * AM_i + \sum_{j=1, j\neq i}^{k} (\frac{P_{ji}}{P_{j+}} * AM_j) = \sum_{j=1}^{k} (\frac{P_{ji}}{P_{j+}} * AM_j) \qquad (1)$$

The second method is known as a classical estimator and was first introduced into the statistical community by Grassia and Sundberg (1982). It is an alternative calibration to the first method. For any class *i* on the ground, it is estimated that $(P_{ii} / P_{+i}) * AG_i$ of its pixels are classified as class *i* on the map and $(P_{ij} / P_{+j}) * AG_j$ of class *j (j =1, …, k, and* $j \neq i$*)* on the ground are misclassified as class *i* on the classified map. So the total of the percentage AM_i for class *i* on a classified map can be estimated as:

$$AM_i = \frac{P_{ii}}{P_{+i}} * AG_i + \sum_{j=1, j\neq i}^{K} (\frac{P_{ij}}{P_{+j}} * AG_j) = \sum_{j=1}^{K} (\frac{P_{ij}}{P_{+j}} * AG_j) \qquad (2)$$

This method can be expressed in matrix algebra as:

$$AM = \begin{bmatrix} AM_1 \\ AM_2 \\ \dots \\ AM_k \end{bmatrix} = P_i * AG = \begin{bmatrix} \frac{P_{11}}{P_{+1}}, \frac{P_{12}}{P_{+1}}, \dots, \frac{P_{1k}}{P_{+1}} \\ \frac{P_{21}}{P_{+2}}, \frac{P_{22}}{P_{+2}}, \dots, \frac{P_{2k}}{P_{+2}} \\ \dots, \\ \frac{P_{k1}}{P_{+k}}, \frac{P_{k2}}{P_{+k}}, \dots, \frac{P_{kk}}{P_{+k}} \end{bmatrix} \begin{bmatrix} AG_1 \\ AG_2 \\ \dots \\ AG_k \end{bmatrix} \qquad (3)$$

The matrix inverse is used to solve the true percentage *AG* as:

$$AG = (P_i)^{-1} AM \qquad (4)$$

From a statistical point there is no preference between the inverse estimator and classical estimator (Brown 1982, Heldal and Spjotvoll 1988).

4. AN EMPIRICAL STUDY

To investigate the effectiveness of calibration methods and the impact of sampling error, two land-cover maps covering the same area in the western region of the city of Kingston, Ontario, Canada, were used as reference and classified maps in this study. The two maps were generated by classifying a 4 m multispectral IKONOS image with two different classification methods. One of the land-cover maps generated from a texture classifier was treated as reference data. There are 10 land cover categories on the map. The class percentage of different classes ranges from 1.38% to 21.91% with different levels of spatial autocorrelations. The class categories and their percentages on both maps, and their individual accuracy measured with Kappa are listed in Table 1. The quantitative error, the difference between the area percentage of each land-cover category on these two maps, ranges from 0.499% for lawn and artificial grass to 4.29% for natural grass. Without calibration 15.77% of the study area would be counted in wrong land-cover categories. This range of quantitative errors can often occur in classified maps generated from remotely sensed data.

Class ID	Land cover type	Proportion in reference map (%)	Proportion in classified map (%)	Accuracy in Individual Kappa
1	Residential roof	4.79	2.78	0.868
2	Industrial/Commercial roof	1.37	2.05	0.546
3	Paved surface	8.85	6.62	0.958
4	Lawn and artificial grass	8.59	8.09	0.901
5	Coniferous tree	14.12	12.32	0.961
6	Deciduous tree	21.91	20.57	0.863
7	New crop and pasture	17.38	18.89	0.723
8	Nature grass	9.15	13.45	0.551
9	Bare field	3.70	4.59	0.696
10	Water surface	10.13	10.66	0.944

Table 1: Land cover classification scheme and their proportions in two maps

Random sampling was used in accuracy assessment to obtain the contingency table. In this study a sample size of 600 was used. The random samples were generated by a Monte Carlo simulation. The contingency table was generated for each sampling. Since negative estimates of percentages would appear in the results of calibration methods, and since those are inadmissible in practice (Czaplewski and Catts 1991), all simulations with any negative estimates from the two calibrations were discarded. In total, 100 feasible contingency tables were created. The following estimated percentages were calculated for each feasible contingency table:

1. P_{i+} *(i=1, …, 10)*, the percentage of samples of each class on the classified map from the contingency table (ACM_i);
2. P_{+i} *(i=1, …, 10)*, the percentage of samples of each class on the reference data from the contingency table (ACG_i);
3. The estimated value of area percentage from the inverse calibration method for each class *(*AIN_i*)*.
4. The estimated value of area percentage from the classical calibration method for each class *(*ACL_i*)*.
5. The ratio between ACM_i and AM_i ($RPM_i = ACM_i / AM_i$, where AM_i is the area percentage of class *i* on the classified map); this ratio measures how closely the sampling data represent the percentage on the map.
6. The ratio between ACG_i and AG_i ($RPR_i = ACG_i / AG_i$; where AG_i is the true area percentage of class *i* on the reference data); this ratio measures how closely the sampling data represent the true percentage on the reference map. If it equals 1, the percentage in the samples can accurately represent the percentage on the ground. In this case, no ground sampling error exists.

7. The ratio between the estimate from the inverse method and the true area percentage on the reference data ($RIV_i = AIN_i / AG_i$);
8. The ratio between the estimate from the classical method and the true area percentage on the reference data ($RCL_i = ACL_i / AG_i$).

The last two ratios measure how closely the estimates from the calibration are to the true percentage on the reference data. The closer to 1, the better the calibrated estimate is. The averages of the above differences and their standard deviations in 100 simulations were also calculated to check the dispersion of the samples and calibration methods.

5. RESULTS AND DISCUSSIONS

Table 2 summarizes the mean and standard deviation of estimated true percentage of each class obtained from the contingency table and from calibration methods. It is obvious that the estimates from the different methods are not the same. For all classes, the means of all estimated values from the contingency table (ACG_i) and two calibration methods (AIN_i and ACL_i) are closer to the true values than the uncalibrated value directly taken from maps. The estimate from the inverse method (AIN_i) has a mean closer to the true percentage with a smaller standard deviation than those taken directly from the contingency table (ACG_i) and the classical method (ACL_i). The mean estimates of percentage from sampling closely represent the percentage on the map and the true percentage on the ground (or reference data). After calibrating using the inverse method and the classical method, the average AIV_i and ACL_i of class 1 are 4.79% and 5.34%, with standard deviations of 0.90% and 1.46%, respectively. The average estimates of the true percentage from the contingency table and two calibration methods are much closer to the true values of 4.79% than the value of 2.79% on the map without any calibration. However, the difference between the mean of ACG_i (4.75%) and the mean of AIV_i (4.79%) is not obvious or significant. This is the case for all other classes.

Class	1	2	3	4	5	6	7	8	9	10
AGi	4.79	1.38	8.85	8.59	14.12	21.91	17.38	9.15	3.70	10.13
AMi	2.79	2.05	6.62	8.09	12.31	20.57	18.89	13.45	4.59	10.66
ACMi	2.78 ±0.48	2.02 ±0.40	6.44 ±0.83	8.15 ±0.91	12.52 ±1.15	20.34 ±1.47	19.12 ±1.25	13.49 ±1.11	4.71 ±0.66	10.44 ±0.92
ACGi	4.80 ±0.58	1.28 ±0.34	8.76 ±0.84	8.57 ±0.92	14.51 ±1.06	21.70 ±1.39	17.66 ±1.15	9.00 ±0.86	3.78 ±0.62	9.93 ±0.90
AIVi	4.82 ±0.51	1.29 ±0.27	8.92 ±0.53	8.503 ±0.43	14.30 ±0.56	21.86 ±0.68	17.47 ±0.71	8.99 ±0.64	3.69 ±0.41	10.14 ±0.22
ACLi	4.97 ±0.86	1.35 ±0.46	9.06 ±0.82	8.52 ±0.56	14.28 ±0.70	21.94 ±1.04	17.34 ±1.12	8.88 ±1.08	3.52 ±0.59	10.13 ±0.24

Table 2: The mean and standard deviation of the different estimates from 100 simulations. (AG_i=the true percentage of class i on the ground; AM_i = the percentage of class i on the classified map; ACM_i = the estimated AM_i from the sample; ACG_i = the estimated AG_i from the sample; AIV_i: the estimated AG_i from the inverse method; ACL_i: the estimated AG_i from the classical method. The number after the sign ± is the standard deviation)

Comparing the mean differences of each class from different methods and samples (t-test) shows that the true area estimates (ACG_i, AIV_i, and ACL_i) from the contingency table and two calibration approaches show significant improvement ($p > 0.05$) when compared with those from the map (AM_i) and the samples (ACM_i). However, there is no significant difference among the estimates from the contingency table and two calibration methods. Although the two calibration methods consistently led to more accurate means of the estimates for all classes, this did not mean that the estimates from calibration were superior to those taken directly from the contingency table in every simulation. This can be seen clearly in Table 3, which summarizes the comparison of individual estimates from the contingency table and those taken after calibration for each class in all simulations. In all estimates from the inverse method, 70% of them were closer to their true area values than those taken directly from the contingency table, while 65.1% of them were more accurate than those taken using the classical method. It appears that classes that are more homogeneous and more accurately classified have a higher probability of achieving more accurate estimates from the calibration. The two classes with the highest probability of having more accurate AIV_i, and ACL_i than ACM_i were class 4 (irrigated grassland) and class 10 (water), while the two most heterogeneous classes (class 1 of Residential roof and class 2 of Commercial/industrial roof) had the least chance of having more accurate estimates after calibration.

Class	AIV_i is more accurate than ACG_i	AIV_i is more accurate than ACL_i	ACL_i is more accurate than ACG_i
1	57	63	41
2	64	66	43
3	67	66	50
4	81	65	75
5	72	67	66
6	73	68	55
7	67	70	54
8	66	59	46
9	67	66	49
10	87	59	87
Total	70	65.1	56.5

Table 3. The comparison of individual estimates from two calibration methods and those directly from the contingency table for individual class (The number shows the percentage of simulations in which one method led to a more accurate estimate than the other)

To check how accurate the estimate from each individual simulation was, the individual ratios of *RPM*, *RPR*, *RIV* and *RCL* of each class in 100 simulations are plotted in Figure 1. The ratio value of 1 means the estimate is the same as the true percentage. Ratios greater than 1 mean that the estimates overstate the true values, while those less than 1 represent underestimated values.

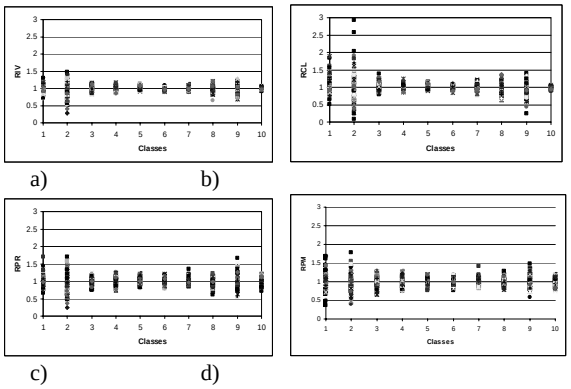

Figure 1.The different ratios for individual classes in 100 simulations: a) the ratio between the estimate from the inverse method and AGi,; b) : the ratio between the estimate from the classical method and AG_i; c) :

the ratio between ACG_i and AG_i; d) the ratio between ACM_i and AM_i.

From Figure 1 it can be seen that the sensitivity of estimate accuracy from different methods varies for different classes. The large classes had much less biased estimates, relatively speaking, than the small classes in all four estimates. The larger classes have relatively smaller variations on their estimates than the smaller classes. This is true for all results from different methods. The class that fluctuates the most in all estimates is class 2 (Industrial and commercial roof), which has the smallest percentage at 1.37%. In all simulations, its highest and lowest ratios of estimates from the contingency table are 1.47 and 0.26. The most stable estimates vary in three different methods. Class 10 (water, 10.13%) has the most stable estimates in both the inverse and classical methods. Its highest and lowest ratios obtained from simulations by using the inverse method are 1.07 and 0.92, while the ratios from the classical method range from 0.90 to 1.07. It should be noted that class 10 is the fourth-largest class, not the largest one. The estimates of class 6 show the smallest variation from the contingency table. From the map it can be seen that class 10 (water) and class 6 (deciduous tree) are less mixed and fragmented by other classes.

6. SUMMARY AND CONCLUSIONS

In this paper, we presented and compared two methods of calibrating area estimate errors on the categorical map by using the contingency table. One hundred contingency tables generated from 100 sets of 600 random samples were used to test the efficiency of three methods. The individual area estimates, as well as average estimates taken directly from the contingency table and two calibration methods were evaluated. Emphasis has been placed on the relationship between the area estimate bias from samples and estimate bias after calibration.

The mean estimates from all methods were substantially less biased than the uncalibrated estimates taken directly from the map or the samples. However, the differences among the true area estimates taken directly from the contingency table and two calibration methods were not significant. Comparison of the individual area estimates for each class showed that the inverse method produced the most stable area estimates with mean values closer to the true percentages. But this did not guarantee that all estimates from the inverse method were superior to estimates taken using the classical method and taken directly from the contingency table. There is no significant difference among the estimates directly from the contingency table and those taken from calibration methods. In this study only 70% and 56.5% of the estimates from the inverse method and the classical method, respectively, were more accurate than the estimates taken directly from the contingency table. The classes that were homogeneous with less percentage difference on the map had a higher probability of achieving more accurate estimates from the calibration than the heterogeneous classes.

The sensitivity of a class to the area estimate bias is related to the size of the class. Classes with a smaller percentage of coverage on a map are more vulnerable to the area estimate bias than are larger classes. This was also suggested by Czaplewski (1992). However, this type of sensitivity is influenced not only by the percentage of a class, but also by spatial patterns of the class. In this study, the smallest class is the most sensitive, but the most stable class is not the one with the highest percentage but the fourth-largest class (water), with a homogeneous and less fragmented presence on the map. Future studies are needed to systematically evaluate the relationship between the accuracy and precision of area estimates and the percentage and spatial autocorrelation parameter of classes.

REFERENCES

Bauer, M.E., Hixson, M.M., Davis, B.J. and Etheridge, J.B., 1978. Area estimation of crops by digital analysis of Landsat data. *Photogrammetric Engineering & Remote Sensing*, 44, pp. 1033-1043.

Brown, P.J., 1982. Multivariate calibration. *Journal of Royal Statistic Society B*, 44, pp. 287-321.

Card, D.H., 1982. Using known map category marginal frequencies to improve estimates of thematic map accuracy. Photogrammetric Engineering & Remote Sensing, 48, pp. 431-439.

Congalton, R.G. and Green, K., 1999. *Assessing the Accuracy of Remotely Sensed Data: Principles and Practices.* 137 pp. (New York: Lewis Publishers),

Czaplewski, R.L., 1992. Misclassification bias in areal estimates. *Photogrammetric Engineering & Remote Sensing*, 58(2), pp. 189-192.

Czaplewski, R.L. and Catts, G.P., 1991. Calibration of remotely sensed proportion or area estimates for misclassification error. *Remote Sensing of Environment,* 39, pp. 29-43.

Dymond, J.R., 1992. How accurately do image classifiers estimate area? *International Journal of Remote Sensing*, 13, pp. 1735-1742.

Goodchild, M.F., Sun, G., and Yang, S., 1992. Development and test of an error model for categorical data. *International Journal of Geographic Information Science*, 6(2), pp. 87-104.

Goodchild, M.F., 2003. Models for uncertainty in area-class maps. In: W.Shi, M.F. Goodchild and P.F. Fisher (Editors), *The Second International Symposium on Spatial Data Quality*. Hong Kong Polytechnic University, Hong Kong, pp. 1-9.

Heldal, J. and Spjotvoll, E., 1988. Combination of surveys and registers: a calibration approach with categorical variables. *International Statistic Review*, 56, pp. 153-164.

Lewis, H.G. and Brown, M., 2001. A generalized confusion matrix for assessing area estimates from remotely sensed data. *International Journal of Remote Sensing*, 22(16), pp. 3223 - 3235.

Lunetta, R.S. and Lyon, J.G., 2004. *Remote Sensing and GIS Accuracy Assessment* (Boca Raton: CRS Press).

Pontius, R.G., 2000. Quantification error versus location error in comparison of categorical maps. *Photogrammetric Engineering & Remote Sensing*, 66(8), pp. 1011-1016.

Prisley, S.P. and Smith, J.L., 1987. Using classification error matrices to improve the accuracy of weighted land-cover models. *Photogrammetric Engineering & Remote Sensing,* 53, pp. 1259-1263.

Walsh, T.A. and Burk, T.E., 1993. Calibration of satellite classification of land area. *Remote Sensing of Environment,* 46, pp. 281-290.