CALIBRATION APPROACHES FOR AREA ESTIMATION ERRORS 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@ncgia.ucsb.edu

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

ABSTRACT:

This paper presents and compares approaches of estimating true area and calibrating area estimate error in 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. Emphasis has been placed on the relationship between the area difference caused by samples and difference after calibration. The estimated true area proportion from the contingency table and two calibration approaches showed significant 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 area estimate bias and more sensitive to sampling bias.

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 (Dymond 1992, Congalton and Green 1999, Lunetta and Lyon 2004). However, the categorical maps stored and used in GIS are never error-free. The quality of a categorical map is determined by whether the category labelled on the map matches what exists in the real world. The errors in a categorical map can be caused by numerous factors in the processes leading up to its creation, such as measurement error, uncertainty/ambiguity in interpretation and category definition, classification bias and digitizing error (Ehlschlaeger and Goodchild 1994). Those errors can bias area estimates from categorical maps (Goodchild 2003), especially for small, rare categories and their changes (Czaplewski 1992, Congalton and Green 1999).

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 proportion of each category on the map is different from the proportion 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 Chrisman 1989, Goodchild et al. 1992, Ehlschlaeger and Goodchild 1994, Moody and Woodcock 1996, Pontius 2000, Goodchild 2003, and Kyriakidis et al. 2004). 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 (Longley et al. 2005), 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). 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 will be 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 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 (see Table 1).

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 bias 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 (Bauer et al. 1978, Prisley and Smith 1987, Hay 1988, Czaplewski and Catts 1991, Czaplewski 1992, and Walsh and Burk 1993) have employed these methods to calibrate area estimators for misclassification errors in remote sensing. There is no consensus on whether one calibration model is superior to another (Brown 1982, Heldal and Spjotvoll 1988, Czaplewski and Catts, 1991, Van Deusen 1996).

One important factor substantially affecting the performance of different calibrations is the sampling data used to generate the contingency table (Van Deusen 1996, Congalton and Green 1999). Sampling errors can be propagated into errors in calibrated area estimates. Czaplewski and Catts (1991) compared two calibration methods by using a Monte Carlo simulation to evaluate the effects of sample size, detail of the classification system, and classification accuracy under random sampling. They concluded that the inverse calibration estimator was consistently superior. However, their conclusion was based on the average infeasibility, bias and dispersion in all simulations and ignored the performance of each individual simulation for each class. It is also not clear how the sampling errors of individual classes affect the calibration results and how sensitive these calibration methods are to different samples.

The objective of this paper is to illustrate different methods to calibrate area estimation error in categorical maps, and compare the various calibration methods by emphasizing the relationship between sampling bias and estimate bias from calibration. The paper is organized as follows. First the area estimation bias and area estimates from contingency tables are described. Two multivariate calibration approaches are then reviewed. An empirical study was conducted by a Monto Carlo simulation to generate random samples. Different methods were compared, based on results generated from two maps containing 10 different classes each with known errors between them. The paper concludes with a discussion of results and applications of the calibration methods.

2. AREA ESTIMATION BIAS AND THE CONTINGENCY TABLE

For the convenience of illustration, the raster categorical map, in which data are represented by pixels (or cells), is used in the following description. Suppose we have a categorical map of *k* categories in which each pixel is assigned into exactly one category. For any class *i* (*i = 1, … k*), its quantitative error on the map is the difference between its proportion on the map and its real proportion on the ground.

Assume that a sample of *N* pixels is chosen from the map in order to evaluate the quantification error of the map. Each observation of the sample is checked by comparing the categories on the map with their true categories, and the results are summarized in a typical contingency table such as that in Table 1 in quantitative terms. The reference data (columns in Table 1) represent truth, while the classified or map data (rows in Table 1) represent the data obtained from the map. The diagonal numbers represent the number of pixels in the samples correctly assigned to their categories, or agreement between the reference and classified data, and the off-diagonal numbers represent the wrongly assigned samples, or lack of agreement between the reference and classified data.

Table 1: A typical contingency table for a map of *k* classes

	Reference Data						
	Class	**Class 1**	**…**	**Class j**	**…**	**Class k**	**Total**
Classified data	**Class 1**	N_{11}	…	N_{1j}	…	N_{1k}	N_{1+}
	…	…	…	…	…	…	…
	Class i	N_{i1}	…	N_{ij}	…	N_{ik}	N_{i+}
	…	…	…	…	…	…	…
	Class k	N_{k1}	…	N_{kj}	..	N_{kk}	N_{k+}
	Total	N_{+1}	…	N_{+j}	…	N_{+k}	N

In Table 1 N_{ij} denotes the number of observations mapped to category i $(i = 1, 2, …, k)$ and found to actually be in category j $(j = 1, 2, …, k)$ in the reference data. Let $\sum_{j=1}^{K} N_{ij} = N_{i+}$ be the total number of observations in the category i on the map, and $\sum_{i=1}^{K} N_{ij} = N_{+j}$ be the number of observations actually in the category j on the reference. Let P_{ij} denote the percentage of observations of N_{ij}, P_{i+} denote the total percentage of observations classified into class i in the thematic map, and P_{+j} be the total percentage of observations of class j in the reference data. Mathematically,

$$P_{ij} = N_{ij}/N \qquad (1)$$

$$P_{i+} = \sum_{j=1}^{K} P_{ij} = \sum_{j=1}^{K} \frac{N_{ij}}{N} = \frac{N_{i+}}{N} \qquad (2)$$

and $$P_{+j} = \sum_{i=1}^{K} P_{ij} = \sum_{i=1}^{K} \frac{N_{ij}}{N} = \frac{N_{+j}}{N} \qquad (3)$$

Let AC'_i denote the proportion of the samples that is class i on the map, and AC_i denote the proportion of the samples that is class i on the ground, then

$$AC_i = P_{+i} \text{ and } AC'_i = P_{i+} \qquad (4)$$

Let $A'i$ $(i = 1, … k)$ denote the proportion of class i on the entire map, which usually is known, and A_i represent the unknown true proportion of class i on the ground. If the sampling data can completely represent the probability of each class on the ground, the contingency table generated from samples can be used to accurately estimate the proportion of class i on the whole map.

$$A_i = AC_i = P_{+i} = \sum_{j=1}^{K} P_{ji} = \sum_{j=1}^{K} \frac{N_{ji}}{N_{+i}} \qquad (5)$$ and

$$A'_i = AC'_i = P_{i+} = \sum_{j=1}^{K} P_{ij} = \sum_{j=1}^{K} \frac{N_{ij}}{N} = \frac{N_{i+}}{N} \qquad (6)$$

However, it is impossible that A_i will exactly equal to AC_i. In other words, sampling bias often occurs and causes the estimates obtained from the contingency table to differ from those on the map and ground. Let Es'_i denote the area estimate difference on the map, which is the difference between area proportion on the whole map for class i and that estimated from the samples, and Es_i denote the difference between area proportion of class i on the ground and that estimated from the samples. They can be represented as

$$Es'_I = A'_i - AC'_i \qquad (7)$$
$$Es_i = A_i - AC_i \qquad (8)$$

Therefore, the quantitative error of class i is

$$A_i - A'_i = AC_i - AC'_i + Es_i - Es'_i = P_{+i} - P_{i+} + Es_i - Es'_i$$
$$= \sum_{j=1, j\neq i}^{k} P_{ji} - \sum_{j=1, j\neq i}^{k} P_{ij} + Es_i - Es'_i \qquad (9)$$

Where $\sum_{j=1, j\neq i}^{K} P_{ij}$ represents commission errors for class i and $\sum_{j=1, j\neq i}^{K} P_{ji}$ represents omission errors for class i. Equation (9) shows that in order to have a true percentage of area estimates for any category on a map there should be

a) no omission or commission errors, and sampling is unbiased ($\sum_{j=1, j\neq i}^{k} P_{ji} = \sum_{j=1, j\neq i}^{k} P_{ij}$ = 0 and $Es_i = Es'_i = 0$); or

b) omission errors are the same as or equal to the commission errors in the contingency table, and the sampling is unbiased ($\sum_{j=1, j\neq i}^{k} P_{ji} = \sum_{j=1, j\neq i}^{k} P_{ij}$ and $Es_i = Es'_i = 0$); or

c) no omission or commission errors are in the contingency table, and the difference between the proportion of class i on the map and its estimate from samples is the same as that in the ground; ($\sum_{j=1, j\neq i}^{k} P_{ji} = \sum_{j=1, j\neq i}^{k} P_{ij}$ = 0 and $Es_i = Es'_i$); or

d) the difference between omission errors and commission errors is equal to the difference between Es_i and Es'_i ($\sum_{j=1, j\neq i}^{k} P_{ji} - \sum_{j=1, j\neq i}^{k} P_{ij} = Es'_i - Es_i$).

However, none of these cases occurs frequently in reality. It should be noted that the Es'_i can be easily obtained for every sampling, but the Es_i is usually impossible to calculate because the real area of each class is unknown. Therefore, it is also difficult to judge if even the above cases are true.

3. CALIBRATION METHODS

Since commission and omission errors often exist on the map and the sampling data cannot completely represent the proportions of different classes on the ground, calibration becomes necessary when the proportion of a class on a map doesn't match with the estimated proportion from the samples. Two calibration methods mentioned above 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 ground, the conditional probability that it is classified as class i on the classified map is $\frac{P_{ii}}{P_{i+}}$, and the conditional probability (commission error) that it is classified as another class j ($j=1, ..., k,$ and $j != i$) is $\frac{P_{ji}}{P_{j+}}$. So the proportion of pixels classified as class i on both the map and the ground is $\frac{P_{ii}}{P_{i+}} * A'_i$, and the total proportion of pixels that is misclassified to other classes is $\sum_{j=1, j\neq i}^{K} (\frac{P_{ji}}{P_{j+}} * A'_j)$. So, for any class i, the proportion in the ground A_i is the sum of both and can be calibrated as:

$$A_i = \frac{P_{ii}}{P_{i+}} * A'_i + \sum_{j=1, j\neq i}^{K} (\frac{P_{ji}}{P_{j+}} * A'_j) = \sum_{j=1}^{K} (\frac{P_{ji}}{P_{j+}} * A'_j) \qquad (10)$$

It can also be expressed in matrix algebra as:

$$A = \begin{bmatrix} A_1 \\ A_2 \\ ... \\ A_k \end{bmatrix} = P_j A' = \begin{bmatrix} \frac{P_{11}}{P_{1+}}, \frac{P_{12}}{P_{1+}}, ..., \frac{P_{1k}}{P_{1+}} \\ \frac{P_{21}}{P_{2+}}, \frac{P_{22}}{P_{2+}}, ..., \frac{P_{2k}}{P_{2+}} \\ ..., \\ \frac{P_{k1}}{P_{k+}}, \frac{P_{k2}}{P_{k+}}, ..., \frac{P_{kk}}{P_{k+}} \end{bmatrix} \begin{bmatrix} A'_1 \\ A'_2 \\ ... \\ A'_k \end{bmatrix} \qquad (11)$$

We can illustrate this method using hypothetical data. Suppose that we have a map with two classes of urban and non-urban, in which 40% is urban and 60% is non-urban, while real proportions of urban and non-urban on the ground are 46% and 54%, respectively. The quantitative errors of urban and non-urban classes on the map are 6% (46% - 40%) and 6% (60% - 54%). When the sampling can truly represent the proportion of each class on the map, an unbiased estimate of the true proportion of urban and non-urban classes can be estimated directly from the contingency table by adding the number of pixels in each column and then divided by the total number of samples.

However, when the sampling can not accurately represent the proportion of each class, the correct proportion of each class cannot be obtained from the contingency table. Table 2 is an example of a contingency table obtained from unrepresentative samples.

	Reference Data			
Classified Data	Class	Urban	Non-urban	Total
	Urban	36	4	40
	Non-urban	10	50	60
	Total	46	54	100

Table 2: An example of a contingency table from an unrepresentative sample

Table 2 tells us that on the classified map, 83.33% (30/36) of the urban land on the map is correctly assigned to the class 'urban', and 12.5% (8/64) of the non-urban land on the map should be assigned to 'urban'. Therefore, the proportion of urban land in the ground can be estimated as the sum of both:

$$AreaofUrban = \frac{30}{36} * 40\% + \frac{8}{64} * 60\% = 40.833\%$$

Here, the 40% and 60% are the known proportions of urban land and non-urban land on the map, respectively.

Similar to the proportion of non-urban land:

$$AreaofNonUrban = \frac{56}{64} * 60\% + \frac{6}{36} * 40\% = 59.167\% \cdot$$

Note the sum of A_i is 1, or 100% of the study area.

The second method is known as a classical estimator and was first introduced into statistical communities by Grassia and Sundberg (1982). It is an alternative calibration to the first method. For any class i, it is estimated that $(P_{ii} / P_{+i}) * A_i$ proportion of the true pixels of class i are classified as class i, and $(P_{ij} / P_{+j}) * A_j$ of class j ($j = 1, \ldots, k$, and $j != i$) are misclassified as class i during classification. So the total of the proportion A'_i for class i on a classified map can be estimated as:

$$A'_i = \frac{P_{ii}}{P_{+i}} * A_i + \sum_{j=1, j\neq i}^{K} \left(\frac{P_{ij}}{P_{+j}} * A_j\right) = \sum_{j=1}^{K} \left(\frac{P_{ij}}{P_{+j}} * A_j\right) \quad (12)$$

The estimated area for all classes is listed and then the true area for each class i (A_i (i=1, .., N)) is solved.

This method can be expressed in matrix algebra as:

$$A' = \begin{bmatrix} A_1' \\ A_2' \\ \ldots \\ A_k' \end{bmatrix} = P_i A = \begin{bmatrix} \frac{P_{11}}{P_{+1}}, \frac{P_{12}}{P_{+1}}, \ldots, \frac{P_{1k}}{P_{+1}} \\ \frac{P_{21}}{P_{+2}}, \frac{P_{22}}{P_{+2}}, \ldots, \frac{P_{2k}}{P_{+2}} \\ \ldots, \\ \frac{P_{k1}}{P_{+k}}, \frac{P_{k2}}{P_{+k}}, \ldots, \frac{P_{kk}}{P_{+k}} \end{bmatrix} \begin{bmatrix} A_1 \\ A_2 \\ \ldots \\ A_k \end{bmatrix} \quad (13)$$

The matrix inverse is used to solve the true proportion A as:

$$A = (P_i)^{-1} A' \quad (14)$$

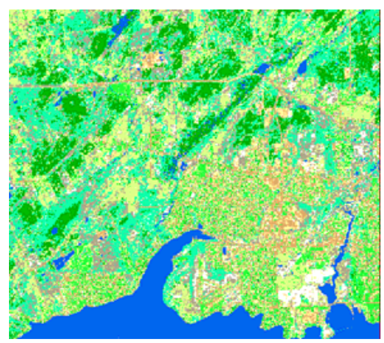
We can use the contingency table in Table 2 to show how this estimation method works. Assume the true proportion of urban land and non-urban land are A and $1\text{-}A$, respectively, that contingency table shows that in the A proportion of urban land on the ground, 30/38 of areas are correctly classified to A while in the $(1\text{-}A)$ proportion of the non-urban land, 6/62 of them are misclassified as urban. Therefore, the 40% of the urban land on the classified map is the summary of both, which is:

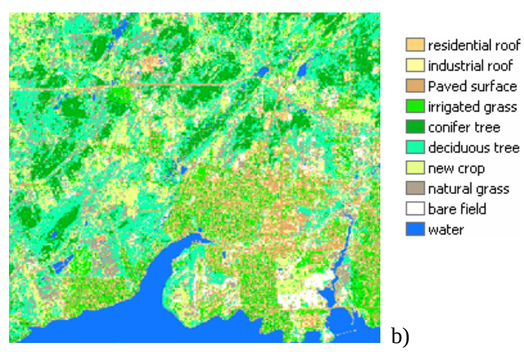
$$40\% = A * 30/38 + (1-A) * 6/62$$
So $A = 43.7745\%$

From a statistical point there is no preference between the inverse estimator and classical estimator ((Brown 1982, Heldal and Spjotvoll 1988). As discussed above, the key factor causing the inaccuracy of estimating the true area proportion in the use of contingency table is the unrepresentative samples. Two calibration methods basically try to reduce the area estimate difference caused by sampling. It would be useful to separate the area estimate difference caused by sampling and difference caused by classification errors on the map.

4. AN EMPIRICAL STUDY

To investigate the effectiveness of calibration methods discussed above, two maps with 10 land cover categories were used as the reference and classified maps. Both maps cover the same area in the western region of the city of Kingston, Ontario, Canada. The two land cover maps were generated by classifying a 4 m multispectral IKONOS image with two different classification methods. In order to check the efficiency of area estimates from contingency table and calibration method, we need to have a controlled condition in which the true misclassification probability and proportion difference are known. Since there are major practical problems in obtaining an accurate depiction of the land cover on the ground for the whole region in practice, hypothetical reference data were used in this study. One of the land cover maps generated from texture classifier was treated as reference data. Since the purpose of this study is not on classification accuracy, assuming one map as reference will not bias any following result. In this way, we have a clear idea of the exact area proportion of each class in both reference and classified maps. The two maps are shown in Figure 1 and the class categories and their proportion on both maps, and their individual accuracy measured in Kappa are listed in Table 3.


a)



b)

Figure 1. The classified map (a) and hypothetical reference map (b) used in this study

The quantitative error, difference between the area proportion of land cover category on these two maps, ranges from 0.499% for lawn and artificial grass to 4.29% for natural grass. The total quantitative error between the proportions of land cover categories in these two maps can be expressed as $\sum_{i=1}^{10} |A_i - A'_i|$, which is 15.77%. It means that without calibration 15.77% of areas in the study area would be counted into wrong land cover categories.

Random sampling was used in accuracy assessment to obtain the contingency table. Statistically, random sampling is the most unbiased sampling method (Stehman 1992). Czaplewski and Catts (1991) suggested that at least 500 samples should be used when calibrating the area estimate bias. 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 the negative estimates of percentages would appear in the results of calibration methods, and those are inadmissible in practice, all simulations with any negative estimates from the two calibrations were discarded.

Class ID	Land cover type	Proportion in reference map (%)	Proportion in classified map (%)	Accuracy in Individual Kappa
1	Residential roof	4.7904	2.7847	0.868
2	Industrial/Commercial roof	1.3763	2.0476	0.546
3	Paved surface	8.8499	6.6212	0.958
4	Lawn and artificial grass	8.5868	8.087	0.901
5	Coniferous tree	14.1183	12.3097	0.961
6	Deciduous tree	21.91116	20.5688	0.863
7	New crop and pasture	17.3783	18.888	0.723
8	Nature grass	9.1542	13.4473	0.551
9	Bare field	3.7037	4.5862	0.696
10	Water surface	10.1305	10.6595	0.944

Table 3: Land cover classification scheme and their proportions in two maps

In total, 100 feasible contingency tables were created. The following estimated proportions were calculated for each feasible contingency table:

1) P_{i+} *(i=1, …, 10),* the percentage of samples of each class in the map from the contingency table;
2) P_{+i} *(i=1, …, 10),* the percentage of samples of each class in the reference data from the contingency table;
3) The estimated value from the inverse calibration method for each class (AIN_i).
4) The estimated value from the classical calibration method for each class (ACL_i).
5) The ratio between P_{i+} and $A_i^{'}$ ($RPM_i = P_{i+} / A_i^{'}$, where $A_i^{'}$ is the proportion of class *i* in the map); this ratio measures how close the sampling data can represent the proportion on the map. If it equals 1, the proportion in the samples can accurately represent the proportion on the map. If it is greater than 1, the proportion in the samples overestimates that on the map. If it is less than 1, it underestimates it.
6) The ratio between P_{+i} and A_i ($RPR_i = P_{+i} / A_i$; where A_i is the true proportion of class *i*); this ratio measures how close the sampling data can represent the true proportion on the reference map.
7) The ratio between the estimate from the inverse method and the true proportion ($RIV_i = AIN_i / A_i$);
8) The ratio between the estimate from the classical method and the true proportion ($RCL_i = ACL_i / A_i$).

The last two ratios measure how close the estimates from the calibration are to the true proportions. In order to evaluate the difference in the sensitivity of various classes on the samples and the relationship between the difference caused by samples and the estimate from calibration methods, the following values were calculated:

1) The difference between P_{i+} *(i=1, …, 10)* and the proportions on the map $A_i^{'}$, which is measured by the absolute difference (*ADPM*) between P_{i+} and $A_i^{'}$. For each simulated table, *ADPM* is calculated as $ADPM = \sum_{i=1}^{N} | P_{i+} - A_i^{'} |$, where N is the total number of classes.
   ADPM indicates how closely the samples represent the proportions on the map.
2) The difference between P_{+i} and the true percentages in the reference data (*ADPR*). For each simulated table, *ADPR* is calculated as
   $ADPR = \sum_{i=1}^{N} | P_{i+} - A_i |$,
   where *ADPR* measures how closely the samples represent the proportion in the reference data.
3) The difference between the estimates obtained from the classical method and the true percentage in the reference data (*ADCL*).
4) The difference between the estimates obtained from the inverse method and the true percentages in the reference (*ADIV*). For each simulation, *ADCL* and *ADIV* are calculated as $\sum_{i=1}^{N} | AI_i - A_i |$, where AI_i is the estimated true proportion for class *i* from the calibration method.

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. Previous studies on the efficiency of the two calibration methods compared the average *ADCL* and *ADIV* with the true proportion without considering *ADPM* and *ADPR* and the sensitivity of individual classes on the area estimate bias. As pointed out in Eq. (7) to (9), the unrepresentative samples are the key to the area estimate bias from the contingency table. When the samples can accurately represent the proportion of each class, the true proportion of each class can be estimated directly from the contingency table by P_{+i} without any calibration methods. Therefore, the evaluation of the efficiency of the two calibration methods should focus on how much different of estimate area proportion is caused by sampling can be reduced by the two calibration methods, rather than the absolute difference between the estimates from the calibration and the true proportions. In other words, the comparison should emphasize the difference between the proportion estimates directly from the contingency table and those from the calibration in order to evaluate the efficiency of the calibration methods and the relationship between the difference caused by sampling and estimates from the calibration methods. A two-tailed independent t-test was used to see whether there was a difference among mean estimates obtained from different calibration methods and those obtained directly from the contingency table. The correlation coefficient was calculated between the ratios of RPR_i, PIV_i, RCL_i and RPM_i.

5. RESULTS AND DISCUSSIONS

Table 4 summarizes the mean and standard deviation of estimated true proportion 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 (P_{+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 proportion with a smaller standard deviation than those taken directly from the contingency table (P_{+i}) and the classical method (ACL_i). For example, the proportions of class 1 (Residential Roof) are 4.79% and 2.785% in the reference map and classified map, respectively. The average area estimate of class 1 on the map (P_{i+}) from 100 simulations is 2.775% with the standard deviation of 0.48% while the average true area estimate of class 1 from the contingency table (P_{+i}) is 4.753% with a standard deviation of 1.13%. From these numbers it can be seen that the mean estimates of proportions from sampling closely represent the proportions on the map and the true proportions on the ground (or reference data). After calibrating using the inverse method and the classical method, the average estimates of class 1 are 4.796% and 5.3387%, with standard deviations of 0.896% and 1.463%, respectively. The average estimates of the true proportion from the contingency table and two calibration methods are much closer to the true values of 4.79% than the value of 2.785% on the map without any calibration. However, the difference between the mean of P_{+i} (4.753%) 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
A_i	4.79	1.376	8.85	8.587	14.118	21.912	17.378	9.154	3.704	10.13
A_i'	2.785	2.048	6.621	8.087	12.31	20.57	18.89	13.447	4.586	10.659
P_{i+}	2.775 ±0.482	2.02 ±0.404	6.438 ±0.833	8.146 ±0.907	12.525 ±1.154	20.342 ±1.472	19.118 ±1.251	13.487 ±1.114	4.708 ±0.656	10.44 ±0.918
P_{+i}	4.797 ±0.584	1.285 ±0.346	8.757 ±0.846	8.572 ±0.925	14.515 ±1.056	21.7 ±1.39	17.66 ±1.155	9.001 ±0.858	3.78 ±0.62	9.932 ±0.902
AIV_i	4.8198 ±0.506	1.2898 ±0.27	8.9212 ±0.534	8.5033 ±0.434	14.304 ±0.565	21.865 ±0.676	17.471 ±0.713	8.9874 ±0.643	3.695 ±0.414	10.14 ±0.2164
ACL_i	4.971 ±0.861	1.352 ±0.461	9.057 ±0.815	8.521 ±0.563	14.284 ±0.701	21.945 ±1.045	17.34 ±1.12	8.8806 ±1.08	3.523 ±0.586	10.13 ±0.245

Table 4: The mean and standard deviation of the different estimates from 100 simulations (A_i: the true proportion of class i; A_i': the proportion of class i on the map; P_{i+}: the estimated A_i' from the samples; P_{+i}: the estimated A_i from the samples; AIV_i: the estimated A_i from the inverse method; ACL_i: the estimated A_i from the classical method. The number after the sign ± is the standard deviation)

Without any calibrations, the mean of total area estimate difference from the map is 15.77%. The average proportion difference caused by the sampling on the map is 9.247%. The average total difference of true estimates of all classes has been reduced from 15.77% to 8.82%, 5.01 % and 7.51% for the contingency table, the inverse method and the classical method, respectively. From Table 4 it is obvious that the average estimate from the inverse method was virtually unbiased for all classes compared with those taken from the contingency table and the classical method. This is consistent with the conclusion of Czapleski and Catts (1992). However, the improvement was different for different classes. For the classes with small difference (such as class 1 and 10), the improvement was not obvious for all estimates. This suggests that the calibration of the contingency table will not reduce the area estimate difference substantially when the error of area estimates on the map is relatively small (<0.5% in this study).

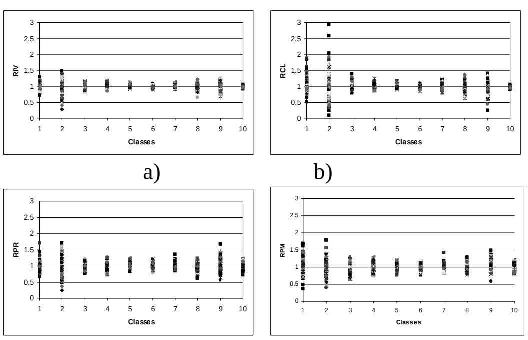
Comparing the mean differences of each class from different methods and samples (t-test) yields that the true area estimates (P_{+i}, AIV_i, and ACL_i) from the contingency table and two calibration approaches show significant improvement ($p > 0.05$) when compared with those from the map (A_i') and the samples (P_{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 5, 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 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 P_{+i} were class 4 (irrigated grassland) and class 10 (water), which are two classes with the most homogeneous patterns and least proportion difference on the map, 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.

Classes	AIV_i is more accurate than P_{+i}	AIV_i is more accurate than ACL_i	ACL_i is more accurate than P_{+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 5: 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) (P_{+i}: the estimated A_i from the samples; AIV_i: the estimated A_i from the inverse method; ACL_i: the estimated A_i from the classical method)

What have been discussed above are the average values from 100 simulations. In practice, it is impossible to have 100 realizations of reference data to calculate the mean estimates as we did in this study using Monte Carlo simulations. Usually only one realization of the reference data is available to generate the contingency table. 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 2. The ratio value of 1 means the estimate is the same as the true proportion. Ratios greater than 1 mean that the estimates overstate the true values, while those less than 1 represent underestimated values. The closer the ratios are to 1, the less bias the estimates have.

From Figure 2 it can be seen that the sensitivity of each class to the area estimate bias varies. The large classes had much less biased estimates, relatively speaking, than the small classes in all four estimates. For example, the four classes (class 5, 6, 7 and 10) with the largest proportions had the minimum ratios of 0.913, 0.903, 0.87 and 0.916, and the maximum estimates of 1.14, 1.08, 1.16 and 1.068 from the inverse method. On the other hand, the estimates of the three smallest classes (class 1, 2, and 9) fluctuate from 0.69, 0.26, and 0.669 to 1.308, 1.468 and 1.27 of their true proportions from the same method. 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 proportion at 1.37%. In all simulations, the highest and lowest ratios of estimates from the contingency table are 1.468 and 0.26 of the true values (1.37%) of class 2. 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 1.068 and 0.916, while the ratios from the classical method range from 0.90 to 1.068. 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. It appears that the fluctuating range of estimates for a class is not only affected by its proportion, but also by other factors such as distribution, pattern and size of a class as well as the sampling method and sampling number used.



a) b) c) d)

Figure 2: The different ratios for individual classes in 100 simulations: a) the ratio between the estimate from the inverse method and $A_{i,}$ b) : the ratio between the estimate from the classical method and A_i, c) : the ratio between P_{+i} and $A_{i;}$ d) the ratio between P_{i+} and A_i'.

Comparing Figures 2a, 2b, and 2c, it is obvious that the results from the inverse method are less fluctuant than those from the classical method and the contingency table. This can also be seen clearly from the minimum and maximum ratios of different methods. For class 2, the most fluctuant class, the difference between the minimum and maximum ratios within 100 simulations is 1.208 for the inverse method, 2.83 and 1.45 for the classical method and contingency table, respectively. The same trend exists for other classes.

Czaplewski (1992) suggested that in order to reduce the area estimate bias, stratified sampling should be used. Van Deusen (1996) also emphasized that the known map marginal frequencies should be maintained. In the stratified sampling in which the RPM_i of each class is close to 1, it is more likely that the estimates from the contingency table and two calibration methods will be close to the true proportions. However, this will not guarantee an accurate true estimate for individual sample realizations in practice. In the simulations in which the RPM_i was very close to 1 (0.98 to 1.02), the ratios of RPR_i, RIV_i and RCL_i still fluctuated from 0.74 to 1.25, which means that even with stratified sampling it is still likely that the true estimates from the calibration and contingency table would over- or underestimate true values by 25%.

6. SUMMARY AND CONCLUSIONS

In this paper, we presented and compared 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 proportions. 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 proportion 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 proportion 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 proportion and spatial autocorrelation parameter of classes.

When the proportions estimated from the samples were close to the proportions on the map, the true area estimates from both the contingency table and the calibration methods were more likely to be close to their true values and there was not much difference in the magnitude of proportion difference among them. This suggests that stratified sampling would be used to reduce the area estimate bias. However, it is still likely that the true estimates from the calibration and contingency table would overestimate or underestimate their true values by 25% within a stratified sampling approach for a single sample realization in practice.

ACKNOWLEDGEMENT

This research is supported by a discovery grant from National Science and Engineering Research Council of Canada. The author would like to thank YaXiong Chen's assistance in the programming of calibration methods.

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.

Chrisman, N.R., 1989. Modeling error in overlaid categorical maps. In: M.F. Goodchild and S. Gospal (Editors), *Accuracy of spatial databases.* pp. 21-34 (London:Taylor and Trancis).

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.

Ehlschalaeger, C.R. and Goodchild, M.F., 1994. Dealing with Uncertainty in Categorical Coverage Maps: Defining, visualizing, and managing errors, *Second ACM Workshop on Advances in GIS,* Gaithersburg. MD, pp. 101-106.

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.

Grassia, A. and Sundberg, R., 1982. Statistical precision in the calibration and use of sorting machines and other classifiers. *Technometrics,* 24, pp. 117-121.

Hay, A.M., 1988. The derivation of global estimates from a confusion matrix. *International Journal of Remote Sensing*, 9, pp. 1395-1398.

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

Kyriakidis, P.C., liu, X. and Goodchild, M.F., 2004. A geostatistical Mapping of Thematic Classification Uncertainty. In: R.S. Lunetta and J.G. Lyon (Editors), *Remote Sensing and GIS Accuracy Assessment*. pp. 145-162(Boca Raton: CRS Press).

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.

Longley, P.A., Goodchild, M.F., Maguire, D.J. and Rhind, D.W., 2005. *Geographic Information Systems and Science,* 517 pp. (New York :John Wiley & Sons Ltd).

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

Moody, A. and Woodcock, C.E., 1996. Calibration-based models for correction of area estimates derived from coarse resolution land-cover data. *Remote Sensing of Environment*, 3(58), pp. 225-241.

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.

Stehman, S.V., 1992. Comparison of systematic and random sampling for estimating the accuracy of maps generated from remotely sensed data. *Photogrammetric Engineering and Remote Sensing*, 9(58), pp. 1434-1530.

Tenenbein, A., 1972. A double sampling scheme for estimating from misclassified multinomial data with application to sampling inspection. *Technometrics,* 14(1), pp. 187-202.

Van Deusen, P.C., 1996. Unbiased estimates of class proportions from thematic maps. *Photogrammetric Engineering & Remote Sensing*, 62(4), pp. 409-412.

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