AN APPLICATION OF STATISTICAL METHODS TO DETERMINE THE APPROPRIATE SIZE AND LOCATION OF CLASSIFICATION REFERENCE AREA

Kang-Ming Lu, Ph.D. student  
Hsien-Te Lin, Professor  
Chen-Yi Sun, Assistant Professor  
Department of Architecture  
National Cheng-Kung University  
Tainan City, Taiwan (R. O. C.)  
gunman0410@yahoo.com.tw

ABSTRACT

Image classification techniques are usually applied to land-use and land-cover researches. One of the most common means of expressing classification accuracy is the preparation of a classification error matrix. Error matrixes compare the relationship between known reference area data and the corresponding results of a classification. Both the representativeness of reference area data and the classification technique influence the overall accuracy. Representativeness of reference area data means the similar component proportion to study area data. Therefore, the purpose of this paper is to find out statistic methods to assist in determining the most representative reference area. In order to assess different methods to determine reference area we tried to make a simulated urban ground truth data which includes 6 categories: road, building, bare land, vegetation, water, and shadow by supervisors Maximum-likelihood method. We tried to apply the Chi-square goodness of fit test to assess the representativeness of reference areas with different sizes and locations, and calculated the indexes $x$ (Chi-square), $P_G$ (the probability of Chi-square cumulative distribution function), and $AA$ (absolute error of classification overall accuracy). The results indicated that the lower $x$ or the higher $P_G$ presents higher representativeness of reference area, and there is a negative functional relationship between $x$ and $P_G$. These results led to the conclusion that the appropriate size and location of the reference area could be determined exactly and efficiently by the index $P_G$.

Keywords: Image classification; reference area; land-use and land-cover; Chi-square goodness of fit test; Chi-square cumulative distribution function

INTRODUCTION

Information on land-cover and land-use has become an integral part of the environmental supervision and developmental planning process (Scott J. et al., 2003). It has aided in the advancement of more effective land-use planning and researches, such as impervious surface (Goetz et al., 2003; Sawaya et al., 2003; Lu et al., 2006; Yuan et al., 2007), land-cover and land-use (Myint, S. W., 2001; Changshan Wu, 2004; George Xian, Mike Crane, 2005), urban issues (Wu, C. S., 2004; Lu, D. S. et al., 2005; Lu, D. S. et al., 2006; Yuan, Fei et al., 2007; Rebecca, L. et al., 2007), and heat-island (Kato, Soushi, 2005; Kato, Soushi, 2007). Therefore, surface classification accuracy has been an important issue in remote-sensing literature.

There are two important factors to affect classification accuracy. The first one is classification method itself, and the second one is the method of classification accuracy assessment. One of the most common means of expressing classification accuracy is the preparation of a classification error matrix. Error matrix compares, on a category-by-category basis, the relationship between known reference data (ground truth of reference area) and the corresponding results of an automated classification (Lillesand, T. M. et al., 2004). In the other hand, there are different approaches have been applied to promote the classification accuracy, which include incorporation of geographic data (Harris & Ventura, 1995), census data (Meserv, 1998), texture features (Myint, 2001; Lu & Weng, 2005;), and structure or contextual information (Stuckens et al., 2000) into remote sensing spectral data. According to the work of reviewing papers, it seemed to be that little attention has been given to the point of classification accuracy assessment. Therefore, the purpose of this paper is to explore a little further into the issue of how to determine the appropriate size and location of
reference area to reduce the error of classification accuracy assessment.

The representative of reference area is one of the most important factors that affects the classification accuracy assessment. If there is a reference area with similar surface component to study area, it could be defined as a representative one. Therefore, we defined the term “the representative reference area” as the one which surface component is similar to study area. Based on the definition, we tried to use the Chi-square goodness of fit test to calculate the proportion of Chi-square cumulative distribution \( (F_k(x)) \) as an index \( (P_G) \) to assess the representative of reference area.

The issue of this paper is divided into the following points: (1) How to determine the appropriate size of reference area; (2) How to determine the appropriate location of reference area. In order to solve the points above, we designed a series of experiments to discuss the relationship between the representative of reference area and indexes mentioned above. Finally, we tried to propose an objective method to determine the appropriate size and location of representative reference area.

### DATA USED IN THE PRESENT STUDY

#### Study Area

In these experiments, we needed a ground truth data of this study area to be a data base. Because the concern of this paper is the classification assessment accuracy instead of classification accuracy, we decided to use a surface classification map to be the ground truth data of this study area. A square area of 64 km² between urban and suburb area was selected for this study area. It is located within the area of Tainan City and Tainan County, Taiwan and contains diversity urban land-use and land-cover categories, such as urban area, road system, rivers, a seaport, sea, an airport, surrounding rural area, to consider the impact of complex surface on classification accuracy assessment (Fig. 1 (a)).

#### Remote Sensing Data

A FOMOSAT-2 MS image of part Tanina City and Tanina Country, Taiwan, which was acquired on July 12, 2004 under clear weather condition, was used in this research (Figure 1.(a)). The FOMOSAT-2 MS image has red, blue, green, and NIR bands with 8-m spatial resolution (Table 1.). The size of the image of this study area is 1,000,000 pixels (1,000 by 1,000).

#### The Simulated Ground Truth Data (the Classification Map)

We classified the FOMORSAT-2 MS image of study area into 6 categories, road, building, bare land, vegetation, water, and shadow area, by supervisors Maximum-likelihood method, and then got the surface classification map (Figure 1.(b)) of the study area. It was used as a simulated ground truth data of this study area to proceed with these experiments under. The proportions of road, building, bare land, vegetation, water, and shadow in this study area are separately: 13.46%; 15.59%; 6.44%; 44.41%; 12.84%; 7.62% (Table 2.).
Table 1. Major characteristics of FOMOSAT-2 sensors

<table>
<thead>
<tr>
<th>Subsystem</th>
<th>Spectral Range</th>
<th>Spatial Resolution</th>
<th>Swath Width</th>
<th>FOMOSAT-2</th>
</tr>
</thead>
<tbody>
<tr>
<td>PAN</td>
<td>0.52–0.82μm</td>
<td>2 m</td>
<td></td>
<td>24 km</td>
</tr>
<tr>
<td>blue</td>
<td>0.45–0.52μm</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>green</td>
<td>0.52–0.60μm</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>red</td>
<td>0.63–0.69μm</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NIR</td>
<td>0.76–0.90μm</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The Setting of Producer’s Accuracy in this Study

In order to discuss the relationship between the error of classification overall accuracy and \( P \), we needed the producer’s accuracies data of classification. We set separately the producer’s accuracies of road, building, bare land, vegetation, water, and shadow are: 65%; 75%; 80%; 90%; 85%; and 70% according to the our experience of classifying satellite images (Table 2.).

Table 2. The surface composition and supposed producer’s accuracy of study area

<table>
<thead>
<tr>
<th>Study area</th>
<th>Pixels (pixels)</th>
<th>Road</th>
<th>Building</th>
<th>Bare land</th>
<th>Vegetation</th>
<th>Water</th>
<th>Shadow</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>134,572</td>
<td>155,948</td>
<td>64,382</td>
<td>444,086</td>
<td>128,436</td>
<td>72,576</td>
<td>1,000,000</td>
<td></td>
</tr>
<tr>
<td>Proportion (%)</td>
<td>13.46</td>
<td>15.59</td>
<td>6.44</td>
<td>44.41</td>
<td>12.84</td>
<td>7.26</td>
<td>100.00</td>
<td></td>
</tr>
<tr>
<td>Producer's accuracy (%)</td>
<td>65.00</td>
<td>75.00</td>
<td>80.00</td>
<td>90.00</td>
<td>85.00</td>
<td>70.00</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

The Setting of Reference Area

We set the shape of reference area to a single square area for the sake of simplifying the calculation in this study. The length of the side of reference area was set to an independent variable \( n_s \), a positive integer that specifies the number of pixels of the side.

ANALYTICAL METHODS
This Experiment Design

According to the definition of the term “the representative reference area” above in this study, we tried to use the Chi-square goodness of fit test to calculate Chi-square cumulative distribution \( F_k(x) \) (Abramowitz, Milton et al., 1965; Mood, A.M., 1974; Jonhson, N.L, 1994) as an index \( P_G \) to assess the representative of reference area.

Before calculating \( P_G \) index, we need to calculate Chi-square \( (x) \) first. In the other hand, we defined a \( \triangle AA \) index, that means the absolute error of classification overall accuracy of reference area, with intent to discuss the relationship between \( P_G \) and the error of classification overall accuracy.

There are two independent variables in this experiment design. The first one is \( n_s \), with the range from 1 to 1000 pixels, and for this reason we had 1,000 groups of experiments with different \( n_s \) to process. In each group of experiments, we used a moving window method to move the reference area with specific \( n_s \) from the upper left corner to the lower right one methodically through the layer of \( L_G \), the layer of the ground truth data in this study area. The indexes, \( P_G, \triangle AA, x \), of each moving window were calculated in each moving step, and were recorded separately in these layers: \( L_{PG}, L_{\triangle AA}, L_x \). The plane coordinate of recorded point in each layer was the corresponding one of the upper left corner of each moving window (Figure 2.). The data type of these layers, \( L_{PG}, L_{\triangle AA} \) and \( L_x \), are double. Because the values of \( P_G, \triangle AA, \) and \( x \) are all greater than 0.0, we set their initial value as -1.0 to separate the recorded data and unrecorded data after every groups of experiments.

\[ P_G = F_k(x) = \frac{\gamma(k/2, x/2)}{\Gamma(k/2)} \] (1)

where \( k \) is a positive integer that specifies the number of degrees of freedom, \( x \) is Chi-square, \( \Gamma(k) \) is the Gamma function, and \( \gamma(k, x) \) is the lower incomplete Gamma function and. There are 6 surface categories in this study.

**Figure 2.** The flow char of this experiment design.

The Indexes \( x, P_G, \) and \( \triangle AA \)

In probability theory and statistics, the Chi-square distribution is one of the most widely used theoretical probability distributions in inferential statistics. The common Chi-square goodness of fit test of an observed distribution to an expected one is one of the best-known situations in which the Chi-square distribution are used (Abramowitz, Milton et al., 1965; Mood, A.M., 1974; Jonhson, N.L, 1994). In this paper, we calculated Chi-square cumulative distribution \( F_k(x) \) as an index \( P_G \) to assess the representative of reference area.

\( P_G \) is the most important index in this study, and it is calculated on the flowing equation (Abramowitz, Milton et al., 1965; Mood, A.M., 1974; Jonhson, N.L, 1994)

\[ P_G = F_k(x) = \frac{\gamma(k/2, x/2)}{\Gamma(k/2)} \] (1)

where \( k \) is a positive integer that specifies the number of degrees of freedom, \( x \) is Chi-square, \( \Gamma(k) \) is the Gamma function, and \( \gamma(k, x) \) is the lower incomplete Gamma function and. There are 6 surface categories in this study.
area, and the value of $k$ is 5. In this study, Chi-square $x$ means the sum of the weighted error squares of each observed surface proportion in reference area where the expected one indicates study area. Here $x$ is calculated as (Abramowitz, Milton et al., 1965)

$$
x = \sum_{i=1}^{k+1} \frac{(O_i - E)^2}{E}
$$

(2)

where $O_i$ is the observed proportion of surface $i$ in reference area, and $E$ is the expected one of surface $i$ in study area.

In the other hand, the Gamma function (represented by the capitalized Greek letter $\Gamma$) is an extension of the factorial function to real and complex numbers in mathematics. For a complex number $k$ with positive real part the Gamma function is defined by (Abramowitz, Milton et al., 1965):

$$
\Gamma(k) = \int_0^{\infty} t^{k-1} e^{-t} \, dt
$$

(3)

Moreover, the incomplete gamma function is defined as an integral function of the same integrand in mathematics. There are two varieties of the incomplete gamma function: the upper incomplete gamma function is for the case that the lower limit of integration is variable (i.e. where the "upper" limit is fixed), and the lower incomplete gamma function can vary the upper limit of integration. The lower incomplete Gamma function $\gamma(k, x)$ is calculated based on the following equation (Abramowitz, Milton et al., 1965):

$$
\gamma(k, x) = \int_0^x t^{k-1} e^{-t} \, dt
$$

(4)

In this study, index $AA$ is defined to the absolute error of classification overall accuracy of a reference area to discuss the relationship between $P_G$ and the error of classification overall accuracy in detail. It is expressed as

$$
\Delta AA = |A_{Oi} - A_E|
$$

(5)

where $AO_i$ is the observed overall accuracy in specific reference area $i$, and $AE$ is the expected overall one in study area. $AO_i$ and $AE$ were separately calculated as the following functions:

$$
A_{Oi} = \frac{1}{n} \sum_{i=1}^{k} n_i \cdot A_i \times 100\%
$$

(6)

$$
A_E = \frac{1}{N} \sum_{i=1}^{k} N_i \cdot A_i \times 100\%
$$

(7)

where $n$ is the size of specific reference area $i$ ($n = n_i^2$ pixels, in this study), $N$ is the size of study area ($N = 1,000,000$ pixels, in this study), and $Ai$ is the hypothetic producer’s accuracy of surface $i$ (Table 2. & Table 3.). The table 3. shows that $AE$ was calculated by the error matrix of study area based on ground truth data $Ni$ and hypothetic producer’s accuracy $Ai$. $AO_i$ could be calculated in the same way as Table 3.
Determine the Size of Reference Area by $P_G$ and $H_G$

In the basic sample view, if the sample size was expanded until the population one, then the population characteristic of interest would be known exactly (Thompson, S. K., 1992). In the other word, if the size of reference area was expanded until the size of study area, then the each surface proportion in study area would be know exactly. But the bigger the sample size is, the higher the cost of survey is. Therefore we tried to propose statistic methods using the index $P_G$ and the threshold $H_G$ to determine the appropriate size of reference area.

We know that the range of $P_G$ is between 0 and 1. When $P_G$ is closed to 1, the surface proportions of reference area will be similar to the one of study area. According to the basic sample view, the bigger the reference area is, the more exact of classification accuracy assessment is. For the reasons already stated above, we advise the following steps to fine out the smallest size of representative reference area. First, we defined $H_G$ as a threshold of $P_G$, and it was set to 0.99 (or 0.95). Second, keep processing each experiment group (Figure 2.) with an increasing independent variable $n_s$ until the $P_G$ is bigger than $H_G$ for the first time. If it happens, the latest value of $n_s$ is the smallest size of a square representative reference area side. If it never happened, we would recommend you to low down the $H_G$ and try to process the same experiment procedure again until it happens.

In order to observe the relationship between $n_s$ and the $P_G$, we had processed 1,000 groups of experiments and calculated the mean, standard deviation (Std), maximum (Max.), and minimum (Min.) in this study.

Determine the Location of Reference Area by $P_G$

Before determining the location of reference area by $P_G$, we suggest determining the size first. It could be determined by the method in this paper or according to the survey workload which researchers can afford. We set $n_s$ to 100 pixels according to actual workload we can afford (survey area is 800 m²). Second, we calculated index $P_G$ in every possible reference area by moving window method (Figure 2.). Finally, the appropriate location of reference area could be objectively decided according to the highest $P_G$. The indexes, $x$ and $AA$, were also calculated to discuss the relationship between $P_G$ and themselves.

RESULTS

The Appropriate Size of Reference Area

After finishing 1,000 groups of experiment, we got the results of relationship between $P_G$ and $n_s$ in this case (see Figure 3.). The Figure 3. can be divided into 4 zones according whether Max. is bigger than $H_G$ (0.99): (a) where the range of $n_s$ is from 1 to 18, and there is an immediate sharp increase of Max.; (b) $n_s$ is from 19 to 159, and Maxs. are all bigger than $H_G$; (c) $n_s$ is from 160 to 808, and Maxs. are all smaller than $H_G$; (d) $n_s$ is from 809 to 1,000, and Maxs. start being bigger than $H_G$ again.

Figure 3. tells us three important points. First, the appropriate size of $n_s$ is 19 because it is the smallest one which Max. is bigger than $H_G$. Second, increasing the size of $n_s$ to get higher Max. is available only in zone (a).
Finally, the trends of Means and Stds. show us that the smaller the $n_s$ is, the smaller the number of reference area with high $P_G$. It means that we can hardly find out the representative reference area with high $P_G$ by simple random sampling or subjective one when $n_s$ is too small.

**Figure 3.** The relationship between descriptive statistics indexes of $P_G$ and $n_s$.

**The Appropriate Location of Reference Area ( $n_s = 100$ pixels)**

We set $n_s$ to 100 pixels according to actual workload we can afford before. The $L_{PG}$ of this group of experiments was drawn as Figure 4.(a), and it shows three important points. First, there are two zones of high $P_G$ in this map (Figure 4.(a)). Second, the proportion of reference area with high $P_G$ is very low (see both Figure 3. and Figure 4.(a)), and it may be hard to find out the right one by simple random sampling or subjective sampling. But, finally, we can easily determine the best location of reference area by choosing any one which $P_G$ is greater than or equal to $H_G$ in the two zones (Figure 4.(a) & (b)). It follows from what has been said that the method using index $P_G$ is a useful and efficient one to determine the appropriate location of reference area in this case.
Figure 4. Determining the appropriate location of reference area which $P_G$ is greater than or equal to $H_G$ (in the case where $n_s$ is 100 pixels).

The Relationship Between $P_G$ and $x$ of Reference Area

$$P_G = F_k(x) = \frac{\gamma(k/2, x/2)}{\Gamma(k/2)}$$

Figure 5. The negative functional relationship between $P_G$ and $x$.

Eq. (1), (2), (3), and (4) show the negative functional relation between $P_G$ and $x$. The closer $P_G$ is to 1, the closer $x$ is to 0. In this study where $k$ is 5, $H_G$ was set to 0.99 (or 0.95), and the corresponding $x$ is 0.91 (or 1.67) calculated by the Eq. (1). It presents two things: first, the value 0.99 (or 0.95) is a very strict and useful threshold for $P_G$, because the corresponding $x$ is very low especially in the case where $k$ is 5; second, it presents that using the index $P_G$ to determine the suitable size and location of reference area is an effectual method again.
The Relationship Between $P_G$ and $\Delta AA$ of the Reference Area (in the case where $n_s$ is 100 pixels)

Eq. (5), (6), and (7) show the definition of absolute error of classification overall accuracy $AA$; and Figure 6. shows the relationship between $P_G$ and $\Delta AA$. There are three points we were interested in: (a) the closer $P_G$ is to 1, the lower $\Delta AA$ is; (b) although there have not any definite relational function of $P_G$ and $\Delta AA$ been found in this study, it still can not be ignored that the curve edge (the hidden curve line in Figure 6.) of the distribution data is very similar to $F_k(x)$ function (see Figure 5.); (c) some specific low $P_G$s also lead to very low $\Delta AA$s, and we presumed that these data are specific cases which contain more proportions of high producer’s accuracy surfaces than study area. These results lead to the conclusions, that $P_G$ is more exact and comprehensive than $\Delta AA$ when assessing the representative of reference area, and higher $P_G$ leads to lower $\Delta AA$.

![Figure 6. The relationship between $P_G$ and $\Delta AA$ (in the case where $n_s$ is 100 pixels).](image)

CONCLUSION

The representative of reference area is one of the most important factors to affect the classification assessment accuracy. Based on the theory of Chi-square goodness of fit test, a useful and exact method had been proposed to determine the appropriate size and location of reference area. There are three indexes $P_G$, $x$, and $AA$ were suggested to assess the representative of reference area in this paper, and the definitions of $P_G$, $x$, and $AA$ are: the value of Chi-square cumulative distribution function, Chi-square, and the absolute error of classification overall accuracy of a reference area. In order to test the method mentioned above, we designed a series of groups of experiments to process. According to the results of experiments, there is evidence in plenty to support the following conclusions. First, there is a negative functional relation between $P_G$ and $x$; the closer $P_G$ is to 1, the closer $x$ is to 0. Second, the small the reference area is, the less the number of the representative reference area is. Third, the appropriate size and location of reference area could be exactly and efficiently determined by choosing the one which $P_G$ is greater than or equal to $H_G$. All these conclusions make it clear that $P_G$ is the most useful index of the three indexes to assess the representative of reference area. There are also two suggestions in the end. First, we suggest setting $H_G$ to 0.99 or 0.95 when determining the smallest representative reference area. Second, the appropriate size of reference area could be determined by the method suggested in this paper or the appropriate size which researchers can afford.

ASPRS 2009 Annual Conference
Baltimore, Maryland • March 9-13, 2009
REFERENCES


Thompson, S.K., 1992. Sampling, John Willey and Sons.


