원격탐사 영상자료를 이용한 토지피복도 제작을 위한 지상자료 획득 방법

Methodology of ground-truthing for land cover mapping using remote sensor data

이규성*, 김선화, 신정일
Kyu-Sung Lee, Sun-Hwa Kim, Jung-Il Shin
인하대학교 지리정보공학과

Abstract: 토지피복분류, 식생분류, 식물피복도 분류 등 원격탐사 영상자료의 주된 이용 분야에서 지상자료는 매우 중요한 부분을 차지하고 있다. 가령 감독분류를 위한 training site에 대한 측정이나 또는 분류 정확도 검증을 위한 측면에서도 지상측정은 반드시 필요한 부분이다. 본 논문에서는 피복분류 과정에서 반드시 필요한 지상측정을 위한 표본조사에서 유의하여야 할 통계학적 측면에서 고려해야 할 사항을 검토한다.

1. Introduction

Ground truth data, often referred to 'reference data', are crucial part in remote sensing data analysis. Ground truth data are used to assist image classification/interpretation and to calibrate and assess the accuracy of both remotely sensed data and the products of remotely sensed data, such as land cover map and vegetation coverage map. The main objective of this study is to suggest suitable methods of collecting ground-truth data to be used for analysis of remotely sensed data for land cover and/or vegetation coverage mapping.

2. Spatial sampling

Since the collection of ground-truth is highly labor and time intensive work, the number of ground sample unit should be minimized. To minimize the sample size, sampling design for the ground data collection should be statistically sound enough to represent the spatial variability of cover types over the area of interest. Ground sample size should be determined to cover the training data for image classification, assessing classification accuracy, and generating a continuous surface map, such as vegetation percent coverage map. Prior to sample design, we must clarify the relations between the study area, sample site, and subplot.

1) Study area (spatial population)

Population can be a whole area for land cover classification case, while it can be only for a particular land cover (grassland, forest, etc...) for the mapping of percent vegetation coverage.

2) Sample site (spatial sample unit)

Sample site, often called 'sample plot', is the place where the actual ground-truth data are measured or collected. When designing an appropriate sampling scheme for collecting ground-truth data, the major concern is choice of the size, shape, and distribution of sample unit. The sample site can be either point (soil sample) or polygons (vegetation) with various shape of rectangular, circle, or irregular boundary.

3) Subplot (sub sample)

When the size of a sample site is too large, it is rather difficult to collect or measure ground-truth data even within a single sample site. For example, if a sample site has an area of 1ha(100x 100m^2), which seems a reasonable size for the analysis of medium resolution Landsat TM data, it would be very difficult to measure the accurate percent vegetation coverage for a sample site. In case for relatively large area of a sample unit, we may need subplot within a sample unit. Use of subplot would be very similar to two-stage sampling in statistical sampling. After collecting actual ground-truth data from subplots, we need to compile them to derive a representative value for the sample site.

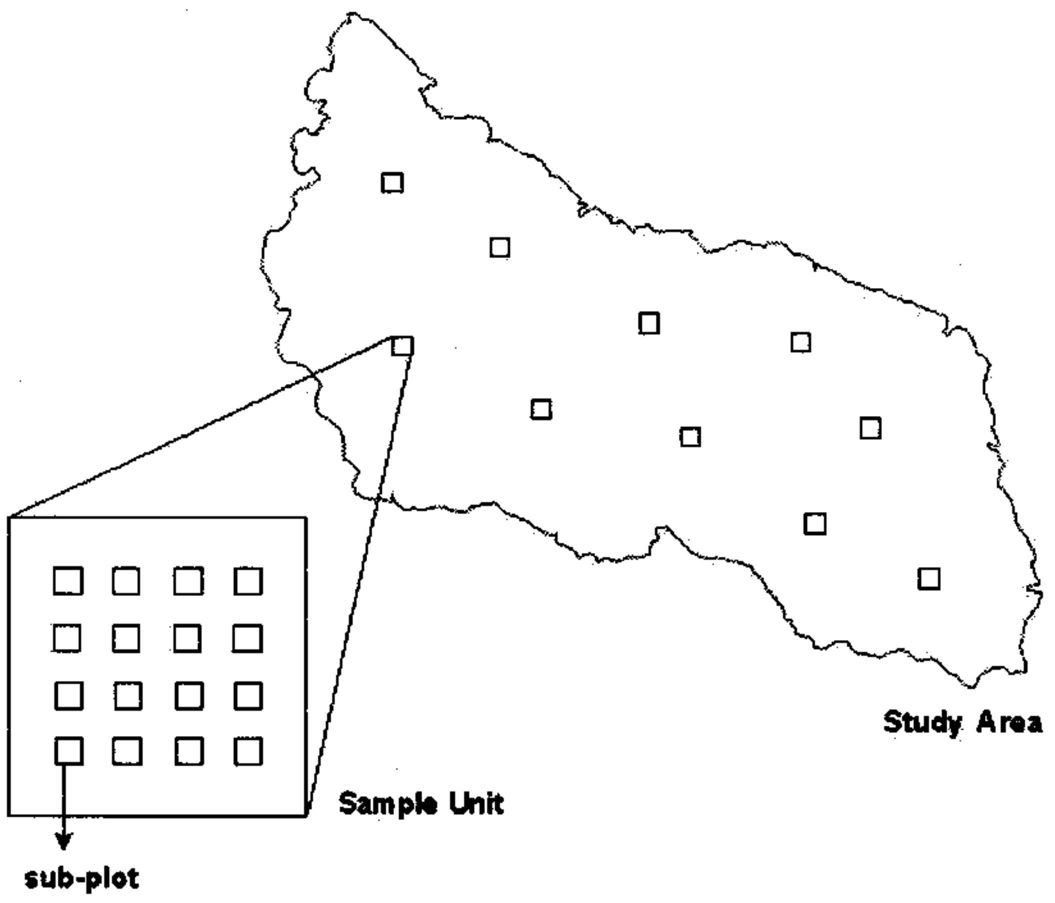

Fig. 1. Definition of sampling unit for the ground measurement.

3. Sample Size

Selecting the suitable number of sample size is dependent on the level of acceptable error and desired level of confidence (Congalton and Green, 1998). Estimation of the sample size can be divided into two different cases of where the ground-truth data is used. The first case is related to the conventional land cover classification where the ground-truth data are primarily used for the assessment of classification accuracy. The second case is related to the quantitative estimation of vegetation coverage where the ground-truth data are used to link the spectral reflectance from remotely sensed data.

1) Sample size for land cover classification.

Ground-truth data for the land cover classification can be utilized for 1) the selection of training fields needed for normal supervised image classification and 2) the accuracy assessment of the classification results. Although there may be several equations for generating the appropriate sample size from statistics books, the one presented by Congalton and Green (1998) is more directly linked to the analysis of remotely sensed data, in particular for the sample size for accuracy assessment of image classification.

The multi-nomial distribution is appropriate for determining image classification accuracy where the ground-truth data is used only for determining right and wrong of each land cover class. The number of sample based on the level of acceptable error and the desired level of confidence is calculated by the following equation (Congalton and Green, 1998).

$$n = C\Pi_i(1-\Pi_i)/B_i^2 \qquad (1)$$

n: total number of sample

C: Chi square value with 1 degree of freedom and $1-\alpha/k$

$1-\alpha$: desired confidence interval

k : number of land cover classes

Π_i : proportion of class i (where i = 1, ⋯, k) within the study area

B_i : desired precision

Assume that there are five land cover categories with the desired confidence level of 95% (1 in 20 chance of rejecting the classification accuracy), that the desired precision of 10% (10% error of the estimated classification accuracy), and that one particular class occupies 53% of the map area (Π_i= 53%). The value of C is determined from Chi square table with 1 degree of freedom and $1-\alpha/k$. χ^2(1, 0.999) = 6.635. The sample size for this case is as follows:

$n = 6.635(0.53)(1-0.53)/(0.1)^2$
$= 165$

There should be k calculations to determine the sample size, one for each Π_i (and B_i for different precision is need for every class) and select the largest n as the desired sample size.

2) Sample size for vegetation coverage mapping

The calculation of required sample size for the mapping of quantitative estimation of vegetation coverage needs rather different approach. In such case, the ground-truth data are used directly to link the spectral reflectance from remotely sensed data. For this purpose, the ground-measured data is continuous variable that does not follow binomial distribution. The population is divided into N sampling unit. N is determined by the size of each sample unit. If the size of sample unit is assumed about the area equivalent to 3x3 pixels of remote sensing data used, there would be 1,000,000 sample units for the area having 3,000x3,000 image size. Calculation of sample size presented here is based on the equation to estimate population mean with desired level of confidence and error (Scheaffer et al., 1979).

In case of simple random sampling, the required sample size to estimate the population mean (i.e., average percent coverage) is as follows.

$$n = \frac{Nt^2S^2}{NB^2 + t^2S^2} \qquad (2)$$

n:total number of samples required

t: t-value for desired probability of confidence level

B:Allowable boundary of error

S^2 : population variance

N: total number of sampling units in the population

The values of t and B depend on the desired level of confidence and error for the sampling scheme. However, an estimate of the population variance S^2 is unknown in practical situations. The population variance is commonly estimated from the result of previous studies or the guesswork by the analyst. The value of B is also needed as an actual value, rather than percentage. To obtain the value of B, we need to have an approximate population mean value, which can be obtained as the same method for the population variance S^2.

For instance, let's assume an grassland where estimated population mean and variance were 50.9% and 22.1, respectively. With 95% confidence interval (t = 1.96) and 10% acceptable error (B=2.545), the required sample size was calculated to be 279.

One alternative to decrease the sample size is the stratification of the population into subgroups that are more homogeneous. The grassland could be divided into 4 strata by the vegetation percent coverage. If we know the proportion and the variance of each grassland class, we should able to calculate the required sample size or each sub-class.

4. Other considerations

In spatial sampling, one of the most critical parameter is the way sample units are dispersed. Several researchers have expressed opinions about the proper spatial sampling schemes and these can be summarized three common sample distributions that have been applied for collecting ground-truthing: simple random sampling, systematic sampling, stratified sampling (Liang, 2004).

Spatial autocorrelation can be defined the spatial self dependency of a variable by its proximity (Cliff and Ord 1973). This should affect the sample size and interval especially the sampling scheme used for collecting ground-truth data, especially in the way this autocorrelation affects the assumption of sample independence (Liang, 2004). Spatial autocorrelation can be observed by the semi-variogram that is a plot of the average semi-variance value with the various lag distance. In semi-variogram, the lag-distance where the maximum semi-variance is reached is called 'range'. The range distance has important meaning for spatial sampling, in particular for the distance between sample units. There is little or no autocorrelation among a variable of interest beyond the range distance. Therefore, the interval between sample units should be at least longer than the range distance not to violate the assumption of sample independence (Atkinson and Emery 1999).

References

Atkinson, P. M. and D. R. Emery (1999): "Exploring the relation between spatial structure and wavelength: implications for sampling reflectance in the field," *International Journal of Remote Sensing,* 20, pp 2663-2678.

Cliff, A. and J.K. Ord (1973): Spatial Processes, Models, and Applications, London : Pion.

Congalton, R.G. and K. Green (1998): Assessing the Accuracy of Remote Sensing Data : Priciple and Practice, Lewis Publishers, New York, NY.

Liang, S. (2004): Quantitative Remote Sensing of Land Surfaces, Wiley & Sons.

Scheaffer et al., (1979): Elementary Survey Sampling, 2nd Ed., Duxbury Press, North Scituate, MA, USA.