Pontus Olofsson, Christopher E. Holden, Eric L. Bullock
Boston Education in Earth Observation Data Analysis/
Department of Earth & Environment, Boston University

4. Estimation

4.1 Introduction

In this Module we select sample of reference observations of the study area with the aim of estimating the area of forest change. We will also use the sample for estimating accuracy of the map and map classes. If stratifying, any map can be used as stratification but the instructions will refer to the land cover change map that was constructed in the third Module, *Change Detection*. The basic idea here is the assumption that the mapped areas of land cover (or change in land cover) are biased because of image classification errors, which are identified by comparing the map to a sample of reference observations. Area estimates and accuracy are then inferred by analyzing the sample. This process includes three main steps: 1) design and selection of sample; 2) response design: interpretation of sample and decision of agreement of reference and map observations; and 3) analysis of sample.

4.2 Sampling design

The sampling design is the protocol for selecting the subset of spatial units (e.g., pixels or segments) that will form the basis of the analysis of area and accuracy. Is it recommended that the sampling design is a probability sampling design, which incorporates randomization in the selection protocol and is defined in terms of inclusion probabilities such that the inclusion probability is known and greater than zero for each unit in the sample. A variety of probability sampling designs are applicable, with the most commonly used designs being simple random, stratified random, systematic and clustered. When choosing a design, three main decisions are whether to use clusters, whether to use strata, and whether to use a systematic or simple random protocol. The primary motivation for cluster sampling is to reduce the cost of data collection – for example, if the map is large and high resolution data need to be collected for each unit in the sample, a clustered design will allow for collection only for the primary sampling units and not for the entire population (cluster designs as defined in this text include 2-stage designs). However, the use of clusters is recommended only if cost savings or practical advantages are substantial as it results in a more complex analysis and because the potential correlation among units within a cluster (i.e., intracluster correlation) often reduces precision relative to a simple random sample of equal size. The use of strata is usually motivated by the fact that land cover change is a small proportion of the total map and if not stratifying the sample, a very large sample

might be required to implement the analysis. A stratified design is therefore usually a good choice, especially if the aim is to estimate land cover change.

That map that was created in Module 3 contained a certain number of classes (including *forest cover loss* and *gain*) but the theory and methodology is generic and could be applied to any thematic map regardless of how the map was made and regardless of the nature and number of map categories. As the aim is to estimate the area of forest change, it is recommended to use the map classes as strata. This will ensure that a sufficient sample size for estimation can be allocated to the change classes.

After settling on a sampling design -- stratified random in this case -- we need to determine the total sample size and allocation of the sample to strata. Please refer to **Subsection S4.1.1 in Supplementary Module S4, *Methods: Estimation*.** to complete this step. After sample size and allocation have been determined, the sample needs to be selected. This can be done in several ways but many software lack good support for a selecting sample. Therefore, we have written our own program: usage instructions are provided in **Subsection S4.1.2 in Supplementary Module S4, *Methods: Estimation*.**

4.3 Response design

Once we designed the sample and a stratified random sample is selected, it needs to be interpreted using a suitable source of reference data and we need to decide if the map and reference observations agree. This step is referred to as the response design.

First, we need to identify the reference data sources. Ideally, we would have plots revisited in the field but this is rarely attainable so we will need to collect reference observations by careful examination of the sample units in satellite data. The more data we have at our disposal the better. If you have no additional data you can use the Landsat data for collecting reference observations but the *process has to be more accurate than the process used to create the map being evaluated*. Careful manual examination can be regarded as being a more accurate process than automated classification. In addition to Landsat data, you can use whatever data available in Google Earth™. Please refer to **Subsection S4.2.1 in Supplementary Module S4, *Methods: Estimation*** for instructions for preparing the sample for interpretation including export of the sample to Google Earth™. As the estimates are based on the sample, **it is important that the labels are correct** and it is recommended that **three interpreters** examine each unit independently.

Once the sample has been interpreted, the agreement between map and reference labels needs to decided; this could potentially be a complicated task but in this case we are using the map classes as strata which makes the decision straightforward. The agreement is preferably expressed in the form of an **error matrix**, which is a simple cross-tabulation of the map labels against the reference labels for the sample units. The error matrix organizes the acquired sample data in a way that summarizes key results and aids the quantification of accuracy and area. The main diagonal of the error matrix highlights correct classifications while the off-diagonal elements show omission and commission errors. The cell entries and marginal values of the error matrix are fundamental to both accuracy assessment and area

estimation. Refer **Subsection S4.2.2 in Supplementary Module S4,** *Methods: Estimation* for constructing an error matrix from the sample data.

4.4 Analysis

With the construction of an error matrix the estimation becomes straightforward. At the heart of the analysis is the implementation of an unbiased area estimator. Different estimators can be implemented but with a sample stratified by discrete map classes, **stratified estimation** has proven useful. A stratified estimator of area includes the area of omission but excludes the error of commission, and is easily implemented from the data in the error matrix. Using the error matrix one can also estimate the accuracy of the map and the map classes. Please refer to **Subsection S4.3 in Supplementary Module S4,** *Methods: Estimation* for instructions for area and accuracy estimation. Note that stratified estimation can be used with simple or systematic random samples too.