Land Cover Area Estimation with a Sample of Very High Resolution Images

Francisco Javier Gallego
JRC-EC, e-mail: javier.gallego@jrc.ec.europa.eu

Abstract

This paper examines the potential accuracy of land cover area estimations based on samples of units with a size that is suitable for data collection through the analysis or interpretation of Very High Resolution (VHR) images. Many VHR sensors produce images with a size between 10x10 km and 20x20 km. The feasibility of area estimates based on a sample of images of 10x10 km is being assessed in the framework of the Geoland2 project (SATCHMO core mapping service). A simple assessment can be carried out using a square grid as sampling frame and a land cover map, such as CORINE Land Cover, as pseudo-truth, but the estimated efficiency is biased because the real land cover has a different spatial structure. To estimate the sampling efficiency we propose a method based on intra-cluster correlations computed from correlograms. The results can be expressed in terms of “equivalent number of points” of each site. We obtained poor efficiency results for area estimation of major land cover types in the European Union (EU) with simple random sampling; for example for the estimation of the total arable land area, each site of 10 x 10 km is equivalent to 3 unclustered points. However there are indications that the sampling scheme can be much more efficient for land cover change. Only sampling error is analyzed in the paper. This implicitly assumes that the observations in each sampled site have no error; the impact of non-sampling error still needs to be analysed.

Keywords: Satellite images, Area Estimation, Land Cover.

1. Introduction

When fine or medium resolution satellite images are used in a large area, it may be necessary to consider only a sample of images. This can happen for example if we want to estimate the changes on forest area in the tropical belt or worldwide. For relatively large regions, e.g. between 1 and 5 million km², a full coverage of medium-high resolution images can be affordable for some projects, such as Image2006 in the EU (Müller et al, 2007), but many projects with more scarce resources need to limit themselves to a sample of images.

Some authors have expressed doubts on the possibility of obtaining valid global estimates of forest area change from a sample of Landsat TM images (Tucker and Townshend, 2000); Czaplewsiki (2003) explains that the accuracy of estimates depends more on the
number of sampling units than on the sampling rate: it makes sense to sample 10% of 1000 images, but sampling 10% of 40 images leads to poor estimates. Cihlar et al (2000) apply a purposive sampling with the restriction that that the land cover profile of the sample in an existing land cover map behaves close to the population profile, but most authors prefer using random or systematic sampling. Stehman (2005) analyses the issue of image sampling from the perspective of the overall error obtained by combination of the bias (systematic error) and sampling error, highlighting that a sampling approach can give better estimates if it reduces the bias. The current increase of sensors that provide Very High Resolution images has boosted the interest on sampling sites with a size between 10x10 km and 20x20 km. In Europe this approach is being assessed for example by the SATCHMO Core Mapping Service of the Geoland2 Project (http://www.gmes-geoland.info) that is exploring the estimation of land cover area change with a sample of approximately 200 sites of 10x10 km in EU27 (Figure 1).

![Figure 1: Sample of 200 sites in EU27 in the Geoland2-SATCHMO Project.](image)

### 1.1 A few approaches to sample satellite images.

The TREES-2 project used a sample of 100 Landsat TM images (full or quarter scenes) to estimate rainforest changes between 1993 and 1997 (Achard et al, 2002). This project used hexagonal tessellation to ensure equal probability sampling (Brink and Eva, 2009, Richards et al., 2000), but the system turned out to be too complex and difficult to understand. Gallego (2005) builds a tesselation with Voronoi polygons computed from the Landsat-TM path-row pattern; the method can be used for other sensors with fixed frame scheme (nadir pointing), but not for sensors with off-nadir capacity or for studies...
combining different satellites. In this case the easiest solution is using a tessellation of squares in an equal-area projection. The choice of 10 x 10 km square tiles can be also justified in terms of cost-efficiency optimisation with medium-high resolution images: if images are cheap or free and the main cost component is photo-interpretation, approximately proportional to the area covered, large sampling units might not be cost-efficient. This happens for example with the FRA2010 remote sensing survey (www.fao.org/forestry/fra/remotesensingsurvey/en/) that foresees the use of TM images, but has chosen a sampling frame with tiles of 10 x 10 km and 20 x 20 km located with a systematic pattern every degree of latitude-longitude across the world (Eva et al, 2010).

2. Test area and data.

The study area for the efficiency indicators presented in this paper is the set of 15 countries that were members of the EU in 2001 (EU15), excluding some islands and overseas territories. The total area is 3.2 Million km².

2.1 CORINE Land Cover (CLC).

CLC has been produced in 3 occasions (around 1990, 2000 and 2006) by photo-interpretation of medium-high resolution images (10-30 meters). For the simulations presented in this paper we have used CLC2000, produced by photo-interpretation of Image2000, a mosaic of Landsat ETM+ images (Multispectral+Panchromatic) resampled with 12.5 m resolution (JRC-EEA, 2005). The nomenclature of CLC2000 has 44 classes. The photo-interpretation was made separately in each country with the same rules. The minimum mapping unit is 25 ha. CLC layers can be downloaded from the data service of the European Environment Agency (http://www.eea.europa.eu/data-and-maps/data). CLC provides an additional layer on land cover change between the reference date of the initial CLC (very roughly around 1990) and CLC2000. The layer is also available in the EEA database and has been used and has been used as main source of data to estimate land cover change areas in the EU, since more reliable data do not seem to exist (Feranec et al. 2010). With 44 classes in CLC, the theoretical number of types of change is 44 x 43, too large for a meaningful analysis. Changes have been grouped into a very small number of types for the preliminary assessment presented in this paper (Table 1).

<table>
<thead>
<tr>
<th></th>
<th>artificial</th>
<th>crops</th>
<th>pasture</th>
<th>heterogeneous</th>
<th>forest &amp; wood</th>
<th>Natural</th>
</tr>
</thead>
<tbody>
<tr>
<td>artificial</td>
<td>0</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>crops</td>
<td>1</td>
<td>0</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>pasture</td>
<td>1</td>
<td>4</td>
<td>0</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>heterogeneous</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>forest &amp; wood</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>natural</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 1: Groups of land cover change categories

Major change types
0  No change
1  New artificial
2  Agricultural expansion
3  Agricultural abandonment
4  Other changes
2.2. LUCAS (Land Use/Cover Area-Frame Statistical survey)

LUCAS is a major EUROSTAT tool to estimate land cover area and change. In 2001 it was carried out in the 15 countries that were member states of the EU. The sampling scheme had a two-stage systematic design (Gallego and Delincé, 2010): Primary Sampling units (PSU) were selected with a grid of 18 km without stratification. Each PSU is a cluster of 10 points following a 5x2 rectangular pattern with a 300 m step. The point is conceived as a circle of 3 m diameter. The field work was carried out in the spring of 2001 except in the UK and Ireland, where it was postponed to 2002 because of the foot-and-mouth disease. LUCAS has a double nomenclature: land cover (57 classes) and land use (14 classes). The LUCAS nomenclature is quite detailed for crops (35 classes in 2001) and relatively coarse for other land cover classes (Bettio et al., 2002).


An easy way to assess the potential accuracy of a sample of sites for land cover area estimation is using a land cover map as pseudo-truth. With this assumption we have a full knowledge of the sampling population and standard errors with different sampling strategies can be computed by simple simulation or applying basic formulas of sampling theory. This approach has been applied in the framework of Geoland2-SATCHMO with CLC2000: A square grid of 10x10 km cells was overlaid on EU15 defining a sampling frame slightly above 32,000 cells after excluding those that are more than 50% outside the target area.

3.1. Comparing cluster sampling with point sampling.

A useful indicator of the efficiency of cluster sampling is the ratio between the variance of the estimated mean with a sample of \(n\) points and the estimated mean with a sample of \(n\) clusters: 

\[
Q = \frac{V_{n_{pt}}(\bar{y})}{V_{n_{clus}}(\bar{y})}
\]

We call this ratio \(Q\) the “equivalent number of points of a cluster”, meaning that a sample of \(n\) clusters gives the same sampling accuracy as a sample of \(n \times Q\) points (we disregard the finite population factor). For simple random sampling \(Q\) is nearly constant when \(n\) changes, unless the sampling rate \(n/N\) is high. The “equivalent number of points” is a simple concept that facilitates the understanding by users of the efficiency of a cluster sampling plan, when the main cost component is proportional to the number of clusters.

3.2. Using CLC2000 to compute the “equivalent number of points” of a cluster.

When we use a land cover map as pseudo-truth, computing \(Q\) for simple random sampling (srs) is straightforward. Table 2 reports the coefficients of variation obtained for some CLC classes with srs samples of 200 points or 200 sites of 10x10 km. Computing \(Q\) becomes more complicated if we have data for a sample of points but the amount of data by cluster is insufficient. For example if we consider the class “arable land”, computing the value of \(Q\) on the basis of LUCAS data for a given sampling plan
CLC is not easy because the number of points in a 10x10 km site is often 0. At most we have 10 LUCAS points in a site and their geographical layout is unsuitable to estimate the within-site variance. When we consider classes that are not reported by CLC, such as single crops, it is obvious that CLC is not enough to assess the sampling efficiency.

Table 2: Estimated equivalent number of points of a site of 10x10 km for some CLC land cover types

<table>
<thead>
<tr>
<th>% area</th>
<th>cv 200 points (%)</th>
<th>cv 200 sites 10 km (%)</th>
<th>equivalent number of points/site</th>
</tr>
</thead>
<tbody>
<tr>
<td>artificial</td>
<td>4.65</td>
<td>32.0</td>
<td>12.3</td>
</tr>
<tr>
<td>arable</td>
<td>28.64</td>
<td>11.2</td>
<td>6.9</td>
</tr>
<tr>
<td>perm crops</td>
<td>2.92</td>
<td>40.8</td>
<td>21.9</td>
</tr>
<tr>
<td>pastures</td>
<td>12.31</td>
<td>18.9</td>
<td>9.7</td>
</tr>
<tr>
<td>heterogeneous</td>
<td>11.97</td>
<td>19.2</td>
<td>7.6</td>
</tr>
<tr>
<td>total agriculture</td>
<td>43.53</td>
<td>8.1</td>
<td>4.8</td>
</tr>
<tr>
<td>forest and woodland</td>
<td>26.64</td>
<td>11.7</td>
<td>6.5</td>
</tr>
<tr>
<td>bare</td>
<td>1.29</td>
<td>62.0</td>
<td>33.8</td>
</tr>
<tr>
<td>other vegetation</td>
<td>3.46</td>
<td>37.4</td>
<td>18.1</td>
</tr>
</tbody>
</table>

The lower spatial autocorrelation for land cover change leads to a much higher number of equivalent points (Table 3). This means that there is more room for a positive cost-efficiency of VHR images for land cover change estimation than for land cover area estimation.

Table 3: Estimated equivalent number of points of a site of 10x10 km for main land cover change types based on the CLC change layer.

<table>
<thead>
<tr>
<th>% area</th>
<th>cv 200 points (%)</th>
<th>cv 200 sites 10 km (%)</th>
<th>equivalent number of points/site</th>
</tr>
</thead>
<tbody>
<tr>
<td>New artificial</td>
<td>0.27</td>
<td>136.7</td>
<td>21.8</td>
</tr>
<tr>
<td>New agriculture</td>
<td>0.15</td>
<td>183.5</td>
<td>34.7</td>
</tr>
<tr>
<td>Agricultural abandonment</td>
<td>0.21</td>
<td>154.6</td>
<td>25.2</td>
</tr>
<tr>
<td>Other changes</td>
<td>1.64</td>
<td>54.8</td>
<td>15.6</td>
</tr>
</tbody>
</table>

3.3. Equivalent number of points and intra-cluster correlation.

If we take an srs of $n$ unclustered points we get $V_{srs}(\bar{y}) = S^2/(n-1)$, where $S^2$ is the population variance. For a sample of $n$ clusters (single stage, i.e. all the $M$ units in each sampled cluster are observed), the variance of the estimate $\bar{y}$ can be written (Cochran, 1977, ch. 9):

$$V_c(\bar{y}) = \frac{N-n}{nNM} S^2 \left(1 + (M-1)\rho_M \right)$$  \hspace{1cm} (1)

where $\rho_M$ is the intra-cluster correlation coefficient for clusters of size $M$, estimated by

$$\hat{\rho}_M = \frac{1}{(M-1)(nM-1)S^2} \sum_{i} \sum_{j \neq k} (y_{ij} - \bar{y})(y_{ik} - \bar{y})$$  \hspace{1cm} (2)

where $y_{ij}$ is the value for the elementary unit $j$ within cluster $i$. 


If we assume that the sampling rate is small and the sample size $n$ is sufficiently large we can approximate the “equivalent number of points” $Q$ by:

$$Q \approx \frac{M}{1 + (M - 1)\rho_M}$$  \hspace{1cm} (3)

Which is not far from $1/\rho_M$ if $M$ is large.

### 3.4. Link with the correlogram.

The correlogram is defined by

$$\rho(d) = \frac{C(d)}{S^2} \quad \text{where} \quad C(d) = \text{cov}(y_j, y_k), \quad d(j, k) = d$$  \hspace{1cm} (4)

estimated by

$$\hat{\rho}(d) = \frac{1}{s - n_d} \sum_{k} (y_j - \bar{y})(y_k - \bar{y})$$  \hspace{1cm} (5)

where $n_d$ is the number of pairs of points at a distance $d$. A tolerance region is usually defined around $d$. It is often accepted that each interval of distances should contain at least 30 observations (Journel and Huijbregts, 1978), however much larger samples are necessary to obtain smooth correlograms with categorical data, such as land cover data, expressed as 0-1 values.

We can compute the intracluster correlation as a weighted average of the correlogram:

$$\hat{\rho}_M = \frac{\sum_{d>0} n_d \hat{\rho}(d)}{n_T}$$  \hspace{1cm} (6)

where $n_T$ is the number of pairs in the same cluster (Gallego et al, 1999). If we accept that the correlogram has a smooth behaviour, it can be estimated from a sample of points, such as LUCAS, that are not specifically adapted to study a clustered sample.

Figure 2 gives a plot of the correlogram of vineyards as represented by CLC and as observed on the ground in the LUCAS sample. For distances between 1.5 km and 18 km some assumptions have to be accepted to interpolate the correlogram based on LUCAS 2001 because it does not provide pairs of points for such distances; this introduces a small uncertainty on $\rho_M$ but the changes are small with different interpolation assumptions. For sites of 10 x 10 km, we get an intra-cluster correlation $\rho_M \approx 0.3$ if we consider CLC data, which leads to an equivalent number of points $Q$ slightly above 3. If we consider field data from LUCAS, the intra-cluster correlation is $\rho_M$ is close to 0.2, leading to a value for $Q$ around 5. For less frequent and more scattered classes, the spatial correlation is lower and the “equivalent number of points” is higher.
Discussion and conclusions.

This paper presents an approach to assess the efficiency of a sampling scheme based on a grid of 10x10 km, suitable to sample VHR images. We mainly focus on the comparison of samples of 10x10 km sites with samples of points, considering only the sampling error, i.e. we implicitly assume that the non-sampling errors (mainly identification errors) are similar for different approaches. This is actually debatable, but we exclude the discussion of this issue from this paper.

The concept of “equivalent number of points” gives an easy-to-understand tool for users to assess the value of each sampled VHR image in terms of statistical accuracy. In the examples presented here the results are not very encouraging for land cover area estimation: if we consider that the average marginal cost per point in LUCAS is around 14 € (Eurostat, 2007) or 22 € if we impute general costs, the added value of a VHR image for land cover area estimation would range between 50 € and 500 € for major land cover types. When we consider land cover change the spatial correlation seems to be much lower and the equivalent number of points can be much higher suggesting a better cost-efficiency. Unfortunately at the moment we do not have suitable ground observations to assess the spatial correlation of land cover change. Correlograms should also behave differently in stratified sampling; this needs to be carefully assessed.

It should be also taken into account that the cost per point in LUCAS is relatively low because the sample size is large. Comparing with smaller surveys might lead to different cost per observation unit. Cost considerations will dramatically change if we consider a different geographic context. For example ground observations in tropical rainforest are extremely expensive; in this case reaching a cost-efficiency threshold would be much easier.

Incidentally it is interesting to notice in the formulas for the “equivalent number of points” that the term $M$ does not explicitly appear, although it has an influence because $\rho_M$ decreases when $M$ increases. This simply confirms once again that questions of the
type “which is the required sampling rate to reach a given sampling accuracy?” are wrongly stated if the sampling units have not been defined. This is obvious to statisticians but must be often repeated to many users.

For land cover change we do not have suitable field data to estimate the correlogram or the intra-cluster correlation, but we have computed that the “equivalent number of points” $Q$ is reasonable high on the basis of the CLC change layer. It is very likely that the value of $Q$ is higher considering point data. With reference to the Geoland2-SATCHMO project, this study strongly suggests that the focus should be on land cover change rather than land cover area estimation, although there is some room for samples of VHR images in areas of difficult access, such as mountains.

References


Cochran W., 1977, Sampling Techniques. New York: John Wiley and Sons


