Sampling to validate a global cropland map.

Javier Gallego*[1], Anne Schucknecht[1], François Waldner[2]

[1] Joint Research Centre. Ispra, Italy
[2] Université Catholique de Louvain, Belgium

*Corresponding author: javier.gallego@jrc.ec.europa.eu

Abtract
A global cropland map is being produced with a 300 m resolution in the SIGMA Project of the EU (Stimulating Innovation for Global Monitoring of Agriculture). The map, based on PROBA-V 300 m images of 2015, is not yet available but the collection of reference data is ongoing on a two-tier sample of plots photo-interpreted on publicly available Very High Resolution (VHR) images with the help of two versions of a Geo-wiki tool: a large sample (initially 40000 plots) photo-interpreted by volunteers (crowdsourcing) and a sub-sample of 4000 plots photo-interpreted by experts. A stratification based on a cropland probability map has been used to select a higher rate in the most difficult areas where the cropland probability is between 25% and 75%. Stratified systematic sampling with a variable number of replicates has been used. The sampling scheme has been assessed using as pseudo-truth the cropland probability map in the European Union. The conclusions suggest that the variance efficiency of the sampling scheme due to the improved spatial homogeneity is moderate. A more important gain appears in terms of traceability of the sample.

Keywords
Global cropland map; stratified systematic sample; crowdsourcing for validation.

I INTRODUCTION

The SIGMA project (Stimulating Innovation for Global Monitoring of Agriculture, http://www.geoglam-sigma.info/Pages/default.aspx) is elaborating a global cropland map based on PROBA-V 300 m images (Dierkx et al., 2014). The project is funded by the research framework program FP7 of the European Union. This paper presents the sampling scheme that has been elaborated for the validation of this map. The aim is ensuring at the same time 1) flexibility, 2) homogeneous spatial layout and 3) improved traceability. The solution we chose was a systematic sample with ranked replicates based on 1° latitude-longitude cells.

The concept of cropland for this map corresponds rather to "annual crops in the current year" rather than the standard concept of cropland in agricultural statistics (FAO, 1996). The chosen definition includes a minimum size threshold of 0.25 ha and a minimum width of 30 m. The SIGMA global cropland map is a binary map (cropland/non cropland) with the reference year 2015. The map will be delivered in latitude-longitude coordinates. Therefore the area of each pixel changes with the latitude. For the sample selection the same latitude-longitude coordinates have been used rather than a cartographic projection.

A two-tier sample is applied for the validation: a large sample of around 40,000 units is being photo-interpreted on very high resolution (VHR) images by volunteers that label as crop/non-crop a grid of 9 points inside each pixel. Experts photo-interpret the whole pixel after automatic segmentation for a smaller subsample of around 4000 units. Validators will interpret sub-pixels or polygons as cropland or non cropland based on a very high resolution imagery such as Google Earth and medium resolution time-series. Additionally, there will be the possibility to opt for "unknown/undetermined", if the validator does not know whether the sample is cropland or not.

II SAMPLING SCHEME

The collection of reference data for validation is being carried out before the cropland map to be validated is available. For this reason the map itself is not used for stratification as it is usually recommended (Olofsson et al., 2014). However a stratification has been built on the basis of the IIASA cropland probability map that has been produced by comparing different thematic maps (Fritz et al., 2015). Areas with very high or very low probability are considered easier to classify and need a lower sampling rate. Pixels with an attributed crop probability between 25% and 75% are sampled with a higher rate (table 1). The fact that the sampling plan does not use the SIGMA cropland map for the stratification might make the reference data set more easily usable to validate other cropland maps, as long as they have a compatible geometry.

A Global Agro-Environmental Stratification (GAES) provided by Alterra and FAO (Mücher et al., 2016), which defines strata based on agrosystem characteristics is applied. These zones will be used as post-strata for the computation, but were not used at the sampling stage.

Table 1. Strata definition from the IIASA cropland probability map

Stratum	Cropland probability
1	0
2	1-25%
3	25-75%
4	>75%

The sampling unit is the pixel of the map (300 m). The option of using 3 x 3 pixels windows to limit the impact of location inaccuracy has been discussed and discarded to reduce the photo-interpretation effort. Comparing the accuracy computed from the centre and the peripheral part of the 300 m unit should help quantifying the impact of location errors.

The usual approach to sample for the validation of a land cover map is independent random sampling in each of the classes of the map, considered as strata. When field work is necessary, sampling units are grouped into clusters to optimize logistics, but this has little interest when the reference data are obtained by photo-interpretation on publicly available images. An alternative approach is systematic sampling with a random starting point. Systematic sampling provides unbiased estimators of the parameter being assessed and is always more efficient than random sampling if the spatial correlation of the parameter of interest decreases with the distance (Bellhouse, 1988) and in particular for land cover data obtained from remote sensing (Dunn and Harrison, 1993). However systematic sampling has some drawbacks:

- It lacks flexibility if the sample size needs to be modified for practical reasons, adding more units in certain strata or reducing the initially foreseen sample (Stehman, 2009).
- Stratified systematic sampling may result in a loss of the spatial homogeneity of the sample distribution that is the basis of its good performance. This happens in particular when the strata are not defined by a limited number of large compact patches, but by scattered small areas.

- There is no unbiased estimator for the variance. The usual estimator for the variance may strongly overestimate it, although some alternatives have been proposed to reduce the overestimation (Wolter, 1984). We use a bi-dimensional adaptation of one of the formulas Wolter had assessed for unidimensional systematic sampling:

$$\hat{V} = \frac{\sum_{j \neq j'} w_{jj'} \delta_{jj'} \left(y_j - y_{j'}\right)^2}{2n \sum_{j \neq j'} w_{jj'} \delta_{jj'}} \quad (1)$$

where the weight $w_{jj'}$ is an average of the weights w_j and $w_{j'}$; $\delta_{jj'}$ is a decreasing function of the distance between *j* and *j'*, usually with zero value beyond a moderate distance.

An important advantage of systematic sampling appears when the estimates are politically sensitive and their acceptance is problematic or if there is a conflict of interest. For example if the validation is performed by the same team that has produced the map under validation, they can be tempted to eliminate difficult points, in particular if an economic interest is involved. It is more difficult to proof that a random sample has not been manipulated than it is with a systematic sample.

Figure 1 illustrates the irregular spatial layout of a random sample compared with a systematic sample and with a stratified systematic sample with the strata defined above based on the IIASA cropland probability map. In the stratified systematic sample the higher sampling rate for stratum 3 (higher uncertainty of cropland presence) has been applied by applying a grid step of 40 km instead of 70 km used for strata 2 and 4. A simple visual inspection suggests that the regular spatial distribution that gives some advantage to the systematic sampling is lost to a large extent with strongly mixed-up strata.

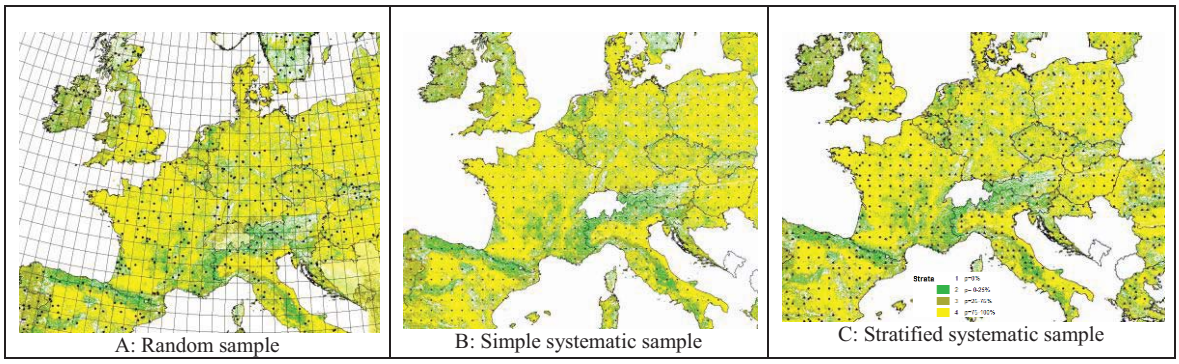


Figure 1: Random and systematic samples in central Europe in an equal-area projection.

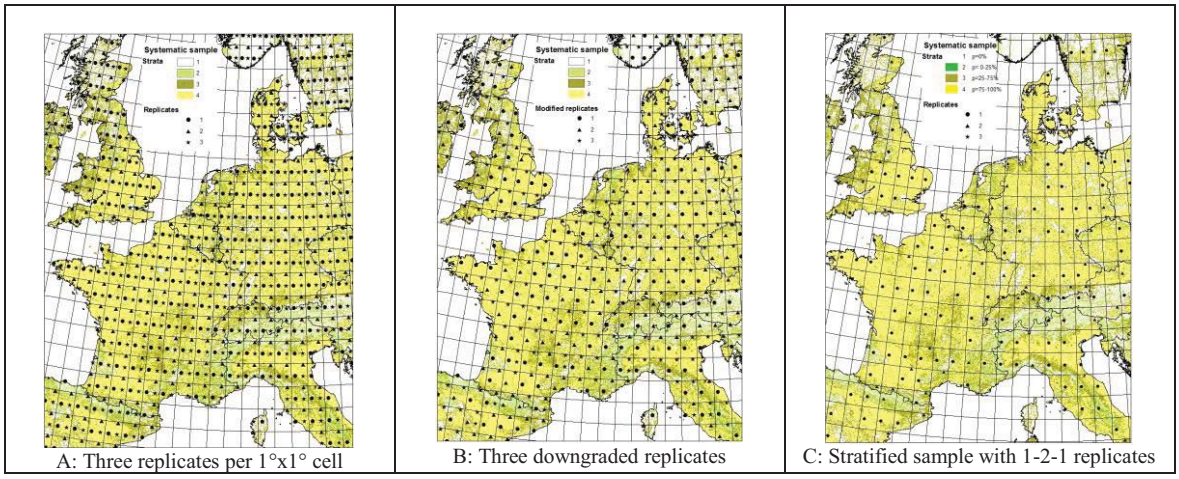
An additional issue with systematic sampling appears if we work with global products, such as the one we are considering, and we want to use geographic coordinates (latitude-longitude) in the sampling process. A straight forward systematic sampling in latitude-longitude is strongly unbalanced because the area of each cell changes with the latitude (111 km x 111 km x cos(lat) assuming a spherical shape).

One possible way to deal with this issue is keeping the systematic sample in latitude-longitude coordinates. This means we have an unequal sampling probability π and we have to apply 1/π as weight to the sample elements (Horvitz-Thompson estimator, Cochran, 1977). However this solution is not very efficient because the higher sampling rate far from the equator is not justified.

Figure 2 illustrates the approach we have used: Figure 2A represents a systematic sample built with three replicates in a 1° x 1° cell. Each replicate is the set of locations with the same relative position in each cell of the grid. Even if the area covered by the illustration is not very large, the higher density of the sample in northern latitudes is visible. This uneven density can be

rebalanced by modifying replicates in such a way that the number of points per replicate for a given latitude band (ring) is proportional to the land area in the corresponding ring (Figure 2B). Since the parallel at latitude α has approximately a length cos(α) compared to the equator, a proportion 1-cos(α) of points belonging to replicate 1 is downgraded to replicate 2, a proportion 2*(1-cos(α)) is downgraded from replicate 2 to 3 and so on. Very quickly we have a proportion rep*(1-cos(α))>1 and points may be downgraded for example from replicate 5 to replicate 8. In the current scheme the selection of points to be downgraded is purely random.

Figure 2B represents a stratified sample based on two modified replicates: stratum 1 is disregarded, 1 replicate is kept for strata 2 and 4 and 2 replicates are kept for stratum 3 (stratified sample 1-2-1). We can see that the geographical homogeneity is broken with some empty rows of cells. Some of the corresponding points have been dropped in the random selection of points to be downgraded from one to another replicate.



A: Three replicates per 1°x1° cell

B: Three downgraded replicates

C: Stratified sample with 1-2-1 replicates

Figure 2: Low density sampling schemes in latitude-longitude coordinates

Figure 3 gives an example of stratified systematic sample in latitude-longitude coordinates with 5 modified replicates for strata 2 and 4 and 10 for stratum 3 (stratified sample 5-10-5). It is similar to the stratified 1-2-1 above, but the irregularities are much smaller because the random subsampling for downgrading replicates happens only in the last replicate.

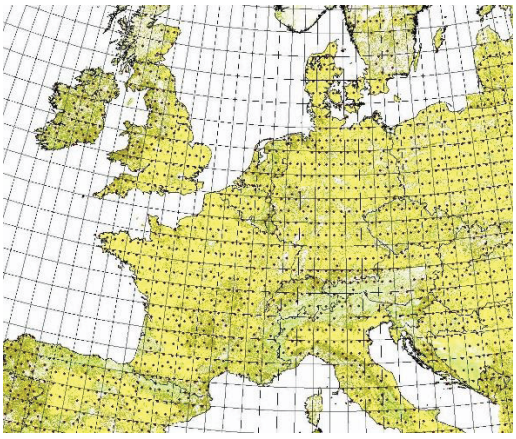


Figure 3: A higher density stratified sample with a systematic of replicates.

III SOME SIMULATIONS

At the moment of writing this paper, the SIGMA cropland map is not yet available and the reference data collection is still ongoing. For this reason we limit the scope of this paper to assessing the behaviour of the sampling schemes described above with two sets of data: the IIASA cropland probability map that we have used for stratification in SIGMA is used as pseudo-truth to be estimated. Considering a target on which we have an exhaustive knowledge allows us to make as many simulations as we wish and to distinguish the real variance of a sampling approach from the estimated variance computed with an empirical formula. We are aware however that the spatial structure of this map is not necessarily similar to the spatial structure of the inaccuracies of the SIGMA cropland map that are the target of our study. For the stratified versions we have chosen CORINE Land Cover (CLC) as data source (EEA 2007). For this reason, we have limited the simulation to the European Union (EU). The CLC land cover classes have been grouped into 3 strata: arable land, other agricultural classes and non-agricultural, that have been sampled with rates proportional to 3-2-1.

Table 2 summarises some of the results obtained. Simple random sampling (srs) is the benchmark with which the other standard errors are compared. Sampling in latitude-longitude coordinates, as if it were an equal area projection, introduces a bias of 7%. Therefore we do not even compute variances or relative efficiency, since such a bias is a sufficient reason to reject the method whatever the variance of the estimates. The same bias appears both with random and with systematic sampling: it has nothing to do with the sampling approach itself, but with the use of a reference system that is not equal-area. The other sampling schemes confirmed in the simulation to be unbiased. Simple systematic sampling has a relative efficiency of 1.22 (i.e. the variance is divided by 1.22 with the same sample size). This standard error could be computed in our example only because it was a simulation and we had an exhaustive knowledge of the sampled population. We can notice that using for systematic sampling the usual variance estimator for random sampling has a positive bias (overestimation) of 23%, leading to an "apparent efficiency" of 0.99. The estimator (Equation 1) based on local variance has a slightly positive bias and reports an estimated efficiency of 1.19. It should be noticed that the sample size in some approaches is not the same as the reference sample size of 1000 units we have chosen for the comparison. This highlights the rigidity of the usual systematic sampling towards the sample size.

Table 2. Comparison of mean and std. error of cropland probability for some sampling schemes.

Sampling method	Mean	n	Std error	relative efficiency
Simple random Sample (srs) equal area	53.1	1000	1.306	
Srs lat-long	49.45	****	****	****
Systematic latlong	49.45	****	****	****
Systematic equal-area	53.1	889	1.256	1.22
Syst. Eq-area usual variance estimator	53.1	889	1.391	0.99
Syst. Eq-area local variance	53.1	889	1.268	1.19
Systematic latlong balanced replicates	53.1	1000	1.213	1.16
Systematic latlong balanced replicates	53.1	100	4.07	1.03
Stratif random	53.1	1000	1.103	1.40
Stratified syst. equal area independent	53.1	959	1.056	1.59
Stratified systematic latlong balanced	53.1	1000	1.049	1.55

The systematic sampling schemes that we have labelled in the table as "balanced replicates" refers to the process described above to reduce the number of units in each replicate by a factor cos(latitude) (Figure 2B and 2C). If we consider a relatively large sample (n=1000 corresponding to a bit less than 3 replicates), we get a reasonably good efficiency of 1.16. In exchange a small sample of 100 units (a portion of one replicate) does not significantly improve the efficiency of a random sample of the same size. This happens because the subsample inside the replicate is random. At the same time the traceability of systematic sampling is lost to a large extent.

The stratification defined from CLC provides a more substantial improvement (efficiency=1.40). Combining stratification with systematic sampling roughly collects the advantages of both aspects in terms of efficiency. Independent systematic samples in different strata with different sampling steps in each stratum (figure 1C) gives a good efficiency in spite of the visually unpleasant pairs of points in different strata that are too close to each other. The type of scheme chosen for SIGMA (stratified systematic samplinglat-long balanced, illustrated in figure 3) behaves slightly worse, but still significantly better than stratified random.

IV SOME PROVISIONAL CONCLUSIONS.

The better performance of systematic sampling compared to random sampling is confirmed even if the relative efficiency is not particularly high. A relative efficiency above 1.5 for systematic sampling is usually not realistic. The rigidity of systematic sampling in terms of sample size can be removed by using a system based on a pattern of replicates. This type of system slightly reduces the efficiency compared to the pure systematic sampling. The usual variance formula for random sampling give a pessimistic bias if applied to systematic sampling. Better estimations are computed on the basis of local variance. The main advantage of systematic sampling is the traceability, i.e. the possibility to proof that no irregularities have been introduced in the sampling process.

Spatial sampling in a geographic reference that is not an equal-area projection, such as latitude-longitude can introduce a significant bias. This can be avoided with different adjustments, both in random and systematic sampling, but such adjustments in systematic sampling can strongly degrade its advantages, both in terms of variance and in terms of traceability.

The conclusions reported in this paper are based on an example of pseudo-truth, even if they are consistent with the conclusions of the wide literature on systematic sampling (see Bellhouse, 1988, for some references). Simulations with a pseudo-truth behaving in a more similar way to the errors of a land cover map would be useful, as would be, of course, the calculation of accuracy indicators on the SIGMA cropland map.

References

Bellhouse D.R., (1988). Systematic sampling, *Handbook of Statistics*, vol. 6, ed. P.R. Krisnaiah, C.R. Rao, pp. 125-146, North-Holland, Amsterdam

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

Dierckx, W., Sterckx, S., Benhadj, I., Livens, S., et al. (2014). PROBA-V mission for global vegetation monitoring: Standard products and image quality. *International Journal of Remote Sensing, 35*(7),

Dunn R., Harrison A.R. (1993). Two-dimensional systematic sampling of land use. *Journal of the Royal Statistical Society Series C: Applied Statistics*, vol. 42 n. 4, pp. 585-601.

EEA (2007) *CLC2006 Technical guidelines*; EEA technical report 17/2007. http://www.eea.europa.eu/publications/technical_report_2007_17

FAO (1996). *Conducting Agricultural Censuses and Surveys*. FAO Statistical Development Series n. 6. FAO Publication, Rome.

Fritz, S., See, L., Mccallum, I., You, L., Bun, A., Moltchanova, E., Obersteiner, M. (2015). Mapping global cropland and field size. *Global Change Biology, 21*(5), 1980-1992.

Mücher, C.A., de Simone, L., Kramer, H., de Wit, A., Roupioz, L., Hazeu, G., Boogaard, H., Schuiling, R., Fritz, S., Latham, (2016). Global Agro-Environmental Stratification (GAES), SIGMA report D31.1.

Olofsson, P., Foody, G.M., Herold, M., Stehman, S.V., Woodcock, C.E., Wulder, M.A., (2014). Good practices for estimating area and assessing accuracy of land change. *Remote Sensing of Environment* 148, 42 - 57

Stehman, S. V. (2009). Sampling designs for accuracy assessment of land cover. *International Journal of Remote Sensing, 30*(20), 5243-5272

Wagner, J. E., & Stehman, S. V. (2015). Optimizing sample size allocation to strata for estimating area and map accuracy. *Remote Sensing of Environment, 168*, 126-133.

Wolter K.M., (1984). An investigation of some estimators of variance for systematic sampling. *Journal of the American Statistical Association*, Vol. 79 No 388, pp. 781-790.