Extending the LACO-Wiki Tool for Land Cover Validation to Crop Area Estimation

[1]Juan-Carlos Laso-Bayas, [1]Steffen Fritz, [1]Linda See, [1]Christoph Perger, [1]Christopher Dresel, [2]Javier Gallego, [2]Hervé Kerdiles, [3]Juergen Weichselbaum, [1]Ian McCallum, [1]Inian Moorthy, [1]Florian Kraxner and [1]Michael Obersteiner

[1]International Institute for Applied Systems Analysis (IIASA), Ecosystems Services and Management Program (ESM)
Schlossplatz 1
Laxenburg, Austria
fritz@iiasa.ac.at, lasobaya@iiasa.ac.at, see@iiasa.ac.at, perger@iiasa.ac.at, dresel@iiasa.ac.at, mccallum@iiasa.ac.at, moorthy@iiasa.ac.at, kraxner@iiasa.ac.at, obersteiner@iiasa.ac.at.

[2]Food Security Unit,
Joint Research Centre,
21027 Ispra, Italy
javier.gallego@jrc.ec.europa.eu, herve.kerdiles@jrc.ec.europa.eu.

[3]GeoVille Information Systems GmbH
Sparkassenplatz 2
Innsbruck Austria
weichselbaum@geoville.com.

DOI: 10.1481/icasVII.2016.g45b

ABSTRACT

LACO-Wiki is an open access, online map validation tool that has been being developed by IIASA and GeoVille. Funded by the Austrian Agency for the Promotion of Science and originally as a prototype by the European Space Agency, the application is intended for a variety of stakeholders including map producers and public agencies, such as environment agencies, that need a land cover validation solution for their products. In addition, LACO-Wiki can be used by researchers and students who are interested in learning about land cover map validation and for examining different methods of accuracy assessment. A secondary aim of LACO-Wiki is to become

an open data repository for calibration and validation data as well as a repository for land cover and land use maps, which could become a valuable resource for the remote sensing and research community.

LACO-Wiki provides guidance that takes users through the full validation workflow from uploading the map to be validated, generating a validation sample, initiating a validation session and creating a customized report on accuracy assessment. The system currently allows crowdsourcing of the validation tasks to members of a team or to a wider set of participants where users can generate random or stratified random samples to validate existing maps using publicly available very high resolution imagery (VHR) such as Google Earth and Bing. New functionality is currently being added to allow for the creation of a systematic sample that can be used for applications such as crop area estimation. Following the methodology outlined in Kerdiles et al. (2014), LACO-Wiki will automatically suggest an indicative number of sample points required to calculate the crop area with a user-specified level of confidence. The system is currently being tested on area estimation of wetland rice in Madagascar. In countries where access to VHR ortho imagery and image processing technology are not available or where field work costs (i.e. ground-truthing) are prohibitive, LACO-Wiki could represent a highly reliable, low cost solution specifically for area estimation in categories that can be consistently identified (i.e. stable over time) in VHR images, but also more generally for map validation.

Keywords: LACO-Wiki, Area estimation, map validation, land use, land cover, crowdsourcing

1. Introduction

LACO-Wiki is a web-based platform allowing users to validate land cover and land use maps. It provides users with a sequentially guided interface where they can upload their own land cover and land use maps, generate samples, create validation sessions and obtain customized reports including an accuracy assessment and a confusion matrix. During the validation session, different reference layers can be used including satellite and aerial imagery from Google and Bing, and features from OpenStreetMap. This paper outlines new functionality that is currently being added to LACO-Wiki to allow users to estimate areas of selected classes, in particular cropland classes.

With regard to area estimation methods, Craig and Atkinson (2013) review classical and newer techniques, including censuses, described as rather expensive and time consuming, which can therefore only be used every few years in developed countries or even less frequently in developing countries. Some, site-specific studies have analyzed areas where the reference data are effectively covering the whole area, thus becoming a census (Hollister *et al,* 2004). Statistical sampling is more commonly used in order to reduce costs and allow for quite accurate inference on larger populations (Congalton and Green, 2008). List frame and area frame sampling are sound methods to gather information, each one suffering from particular problems, e.g. maintenance of lists is difficult or small area estimation by area sampling has been poorly done (Craig and Atkinson, 2013). Hence, area frames and list frames are sometimes combined to address these issues. Area frame sampling is used in several countries including the United States National Agricultural Statistics Service (NASS). In Europe, the LUCAS survey, carried out by EUROSTAT and the European Commission Joint Research Center (JRC), employs a point system sampling approach for land cover area estimation. The system is called the Land Use/Cover Area Frame Statistical survey (LUCAS), and is used to estimate areas and change over time of land cover more generally, among other uses (Gallego and Delincé, 2010). Additionally, Kerdiles et al. (2014)

estimated areas of cropland in North Korea by using a similar system, with a systematic sample of points stratified by slope that were interpreted using very high resolution satellite imagery (e.g. from Google Earth). From these sample points, the authors at the JRC were able to estimate the area of cropland in hectares for different slope strata along with the standard error and coefficient of variation. Following the methodology outlined in Kerdiles et al. (2014), LACO-Wiki will automatically estimate an indicative number of sample points required to calculate the crop area with a user-specified level of confidence. In the current paper, we will demonstrate how the system works on area estimation of wetland rice in Madagascar. Additionally, and following good practice recommendations (Olofsson *et al*, 2014), LACO-Wiki allows users control and transparency at each step as well as producing several measures of accuracy and uncertainty.

The next section provides an overview of the LACO-Wiki system and the methodology behind the extension of the functionality for area estimation. Some remarks on the openness and applicability of the tool conclude the paper. It should be noted that the most usual metrics for validating a land cover map also apply to area estimation problems. For example, computing the omission error for a given class is equivalent to estimating the area that should have been classified as that class, but has been allocated to a different one.

2. Tool description and methods

LACO-Wiki can be accessed online at http://laco-wiki.net. For first time users, a brief user registration is required via Geo-Wiki; alternatively, users can log in using their Google+ or Facebook accounts. Once inside the tool, a sequence of 4 + 1 steps can be performed. The four initial steps are sequential but can be revisited at any time.

The first step consists of uploading a map in either vector or raster format (Figure 1). The currently accepted formats are shape files and GeoTIFFs in a WGS84 or ETRS89 LAEA projection. Once a map has been uploaded, the legend can be customized for on-screen display. The system allows users to upload ancillary datasets (RGB images in the supported projections) that could be of use in the validation process.

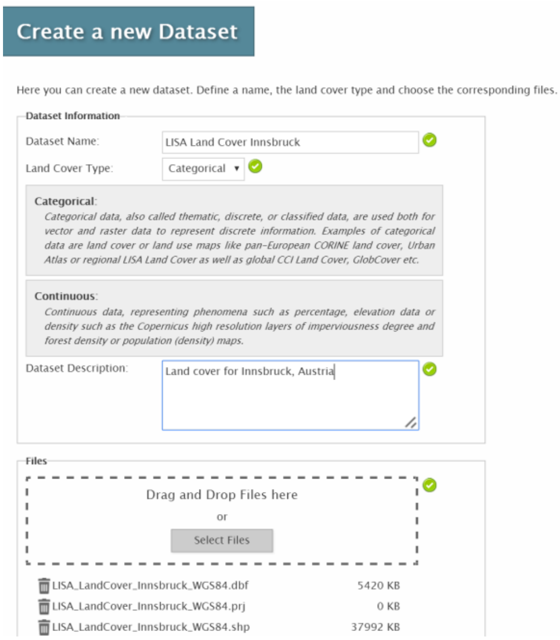

Figure 1. Step 1 – Example showing uploading of a vector dataset

The second step is the generation of a validation sample set (Figure 2). Here users can choose between completely random and stratified random samples for raster data and completely randomized for vector data. In the future, stratified random sampling for vector will be added along with systematic sampling for both types of data sources. The samples can be point, pixel (resolution-dependent) or geometric features (i.e. segments) depending on the type of map that has been uploaded and which sampling method has been chosen. Alternatively, in the near future, a validation sample set created in a previous session or generated in a different tool could be uploaded and used.

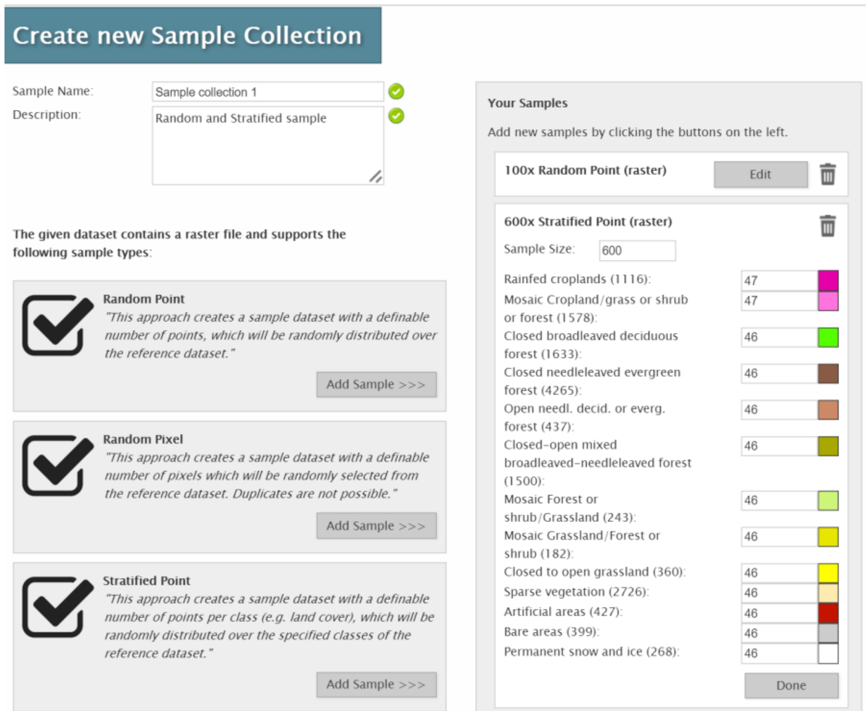

Figure 2. Step 2 – Selection of sampling methods for validation

The third step is the map validation or sample interpretation. With the use of reference data including satellite and aerial imagery from Google and Bing and a feature layer from OpenStreetMap, samples produced during the previous step can be validated. Users can employ the previously defined legends and the validation can be done in two ways: 1) Blind validation, where users interpret the reference data and then decide to which of the available land cover-land use classes the selected sample point, pixel or segment belongs or 2) plausibility validation, where users decide whether the land cover-land use class associated with the point or area of interest is correct or not (Figure 3). In both cases where a segment contains more than one land cover, dominance or majority of one is the deciding factor. In the future a third validation option will be added, called enhanced plausibility, which will allow users to correct the land cover-land use class if they think it is incorrect.

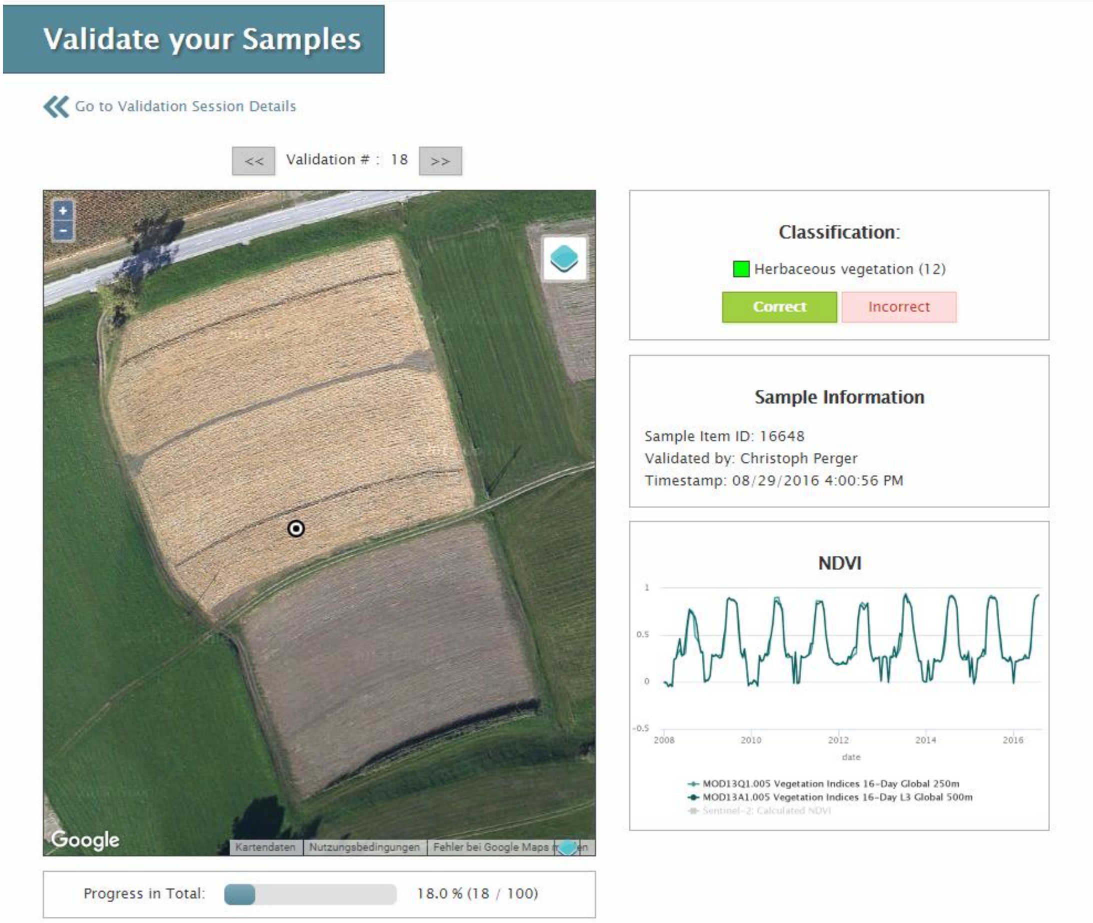

Figure 3. Step 3 – Plausibility validation using reference satellite and aerial imagery

The fourth step is the quality assessment and report generation. Here users can select what accuracy measures they want calculated including user, producer and overall accuracy but also Average Mutual Information (AMI) (Finn, 1993), quantity and allocation disagreement (Pontius and Millones, 2011), kappa (Congalton and Green, 2009) and Portmanteau accuracy (Comber *et al*., 2012). An excel report is created that includes all this information. It is also possible to download the confusion matrix and the raw data.

More recently, we have been extending the LACO-Wiki tool by adding an additional piece of information and an iterative process as shown in Figure 4.

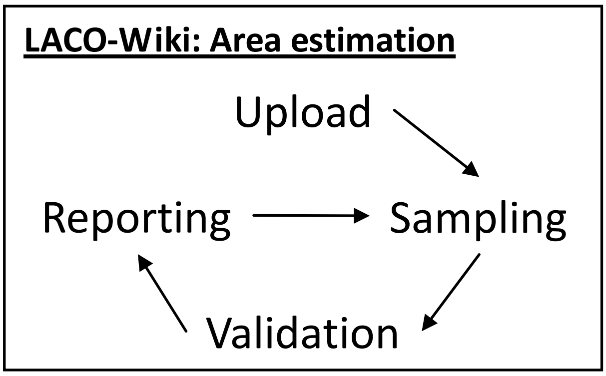

Figure 4. LACO-Wiki iterative process allowing re-sampling after obtaining area estimates reports

The iterative process allows for two-step sampling. The first report produced will include an initial estimation of the area A_c corresponding to each class c (e.g. overall cropland) computed by simply multiplying the total map area A with the proportion p_c of points classified as c.

$$A_c = A\, p_c \qquad \text{with variance} \qquad V(A_c) = A^2 \frac{p_c(1-p_c)}{n-1} \tag{1}$$

If we have a stratification, the estimates are:

$$A_c = \sum_h A_{ch}\, p_{ch} \quad \text{with variance} \qquad V(A_c) = A_h^2 \sum_h \frac{p_{ch}(1-p_{ch})}{n_h - 1} \tag{2}$$

and a coefficient of variation $CV(A_c) = sqrt(V(A_c))/A_c$ usually expressed in percentage.

In general, when a stratification is used for the first time, the available information is not enough to apply the usual Neyman rule for optimal allocation of a sample among strata (Cochran, 1977). A good strategy in this case is allocating half of the foreseen sample in a "reasonable" but subjective way, and applying a more formal optimal allocation to the rest of the sample at a later stage. The "reasonable" allocation applied by LACO Wiki will be based on coarse indications provided by the user on the possible values of p_{ch}. The user is asked about the possible proportion of class c in stratum h suggesting some values: (0, 10%, 20%, 50%, and 80%) and giving the option to provide a user defined value. The user is warned to choose "0" only if they are sure that p_{ch} is negligible. LACO-Wiki will propose a sample per stratum:

$$n_h^* = k^* A_h\, \dot{p}_{ch}(1-\dot{p}_{ch}) \qquad \text{where} \qquad k^* = n^* \frac{A_h\, \dot{p}_{ch}(1-\dot{p}_{ch})}{\sum_h A_h\, \dot{p}_{ch}(1-\dot{p}_{ch})} \tag{3}$$

This first lot allows better tuning the proportions per stratum, obtaining $\ddot{p}_{ch} = n_{ch}^* / n_h^*$ where n_{ch}^* represents the number of points interpreted as belonging to class c in stratum h in the initial sample.

Now we can make a better allocation of the overall sample of size n, provisionally provided by the user:

$$n_h = k A_h\, \ddot{p}_{ch}(1-\ddot{p}_{ch}) \qquad \text{where} \qquad k = n \frac{A_h\, \ddot{p}_{ch}(1-\ddot{p}_{ch})}{\sum_h A_h\, \ddot{p}_{ch}(1-\ddot{p}_{ch})} \tag{4}$$

The additional sample in each stratum will be $n_h - n_h^*$. If the initial selected values for $\dot{p}_{ch}$ were not adequate (i.e. not reflecting actual field conditions), it may happen that $n_h - n_h^* < 0$ for a given stratum. This will lead to an overall sample size slightly larger than foreseen and the user will be warned. The user will be also informed of the foreseen coefficient of variation CV computed by applying formulae (2) with the proportions $\ddot{p}_{ch}$ computed with the initial sample. The user will be asked if the foreseen accuracy and sample size are satisfactory. Otherwise a new value of n will be requested and the foreseen CV computed again until an acceptable trade-off is reached. A sequential approach to tuning the sample size and the accuracy is also possible (Carfagna and Marzialetti, 2009), but not implemented at the moment in LACO-Wiki.

For the classes where more samples are required, the system will randomly allocate them to each stratum and steps 3 and 4 would proceed as described earlier, producing a new validation and a report of accuracy and area estimates. If a pre-selected sample has been loaded in step 2, LACO-Wiki will then use this sample and the classification/validation done in step 3 to estimate the area. The results of the area estimation step will be displayed in the report produced in step 4. The system is iterative in the sense that if users are not satisfied with the area estimates obtained or their variability at any stage in the process, they could decide to run a sampling campaign again using the updated variance estimates.

Direct crowdsourcing or team support is implemented in the system since all validation tasks can be shared with several users in order to distribute the work and achieve faster results.

3. Discussion and applicability

The system is currently being evaluated by Madagascan experts and local experts at IIASA as well as local students. An overall agreement between these interpreters will be done and presented at the conference to demonstrate and compare current capabilities of the system. An increase in the accuracy of the area estimated for wetland rice is expected.

This system can be applied to different land cover classes. The main conditions for its applicability are:

- The targeted land cover class can be reliably identified on VHR images. If we think of crops, this may only be true in the case of paddy rice and some permanent crops, such as olive trees or vineyards, but this strongly depends on the complexity of the agricultural landscape. It can be also applicable in many cases to the overall area of arable land, of permanent crops or cropland.
- The dates of the images in Google Earth and Bing are recent enough to accept that few changes have happened between the image date and the reference date. In any case, the estimation obtained is some type of average of the yearly areas of the different dates of the VHR images considered.

LACO-Wiki was designed as a very straight-forward tool allowing users to easily validate their land cover-land use map and estimate the per-class area. As an online independent platform, it allows users from anywhere in the world to achieve fast map validation results. The calculations shown in the tool and the reports produced are standardized and therefore allow for comparison against different methodologies or even between different samples generated using LACO-Wiki.

The essential advantage of the tool is that validation and area estimation can be done entirely without ground-truth information. This is particularly valuable where field campaigns might be prohibitively expensive or where the areas are inaccessible. Additionally, if local knowledge exists, it can be incorporated as an extra layer in the system and visualized simultaneously with the satellite imagery. This would allow the user to produce a more refined interpretation of what land cover or land use exists at a given location.

Acknowledgments

This work was supported by the EU FP7 funded ERC grant Crowdland (No. 617754).

REFERENCES

Carfagna, E., & Marzialetti, J. (2009). Sequential design in quality control and validation of land cover databases. *Applied Stochastic Models in Business and Industry, 25*(2), 195-205.

Craig, M., Atkinson, D. (2013). A Literature Review of Crop Area Estimation. FAO - Food and Agriculture Organization of the United Nations, Rome, Italy.

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

Comber, A., Fisher, P.F., Brunsdon, C. and Khmag, A. (2012). Spatial analysis of remote sensing image classification accuracy. *Remote Sensing of Environment*, 127, pp.237–246.

Congalton, R.G. and Green, K. (2009). Assessing the Accuracy of Remotely Sensed Data: Principles and Practices. Boca Raton FL, USA: CRC Press/Taylor & Francis.

Finn, J.T. (1993). Use of the average mutual information index in evaluating classification error and consistency. *International Journal of Geographical Information Systems*, 58, 349–366.

Gallego, F.J., Delincé J. (2010) The European Land Use and Cover Area-frame statistical Survey (LUCAS), in *Agricultural Survey Methods*, ed: R. Benedetti, M. Bee, G. Espa, F. Piersimoni, Ch. 10. pp. -168, John Wiley & sons.

Hollister, J. W., Gonzalez, M. L., Paul, J. F., August, P. V., & Copeland, J. L. (2004). Assessing the accuracy of National Land Cover Dataset area estimates at multiple spatial extents. *Photogrammetric Engineering & Remote Sensing*, 70(4), 405-414.

Kerdiles H., Spyratos S. and Gallego F.J. (2014) Assessment of cropland area on sloping land in DPRK, *Earth and Environmental Science*, 17**:** 012043.

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.

Pontius, R.G. and Millones, M. (2011). Death to Kappa: birth of quantity disagreement and allocation disagreement for accuracy assessment. *International Journal of Remote Sensing*, 32 (15), 4407–4429.