Comparison of data fusion methods for generating a forest cover map

Myroslava Lesiv[1,2], Elena Moltchanova[3],
Dmitry Schepaschenko[1,4], Linda See[1], Anatoly Shvidenko[1], Steffen Fritz[1]

[1] lesiv@iiasa.ac.at, International Institute for Applied Systems Analysis (IIASA), Laxenburg, Austria
[2] Lviv Polytechnic National University, Lviv, Ukraine
[3] University of Canterbury, Christchurch, Canterbury, New Zealand
[4] Moscow State Forest University, Mytischi, Russia

Abstract

Various land cover and forest products derived from remote sensing have recently emerged. Previous studies have shown that combining different sources of information (e.g. remote sensing data and ground-based data on the distribution of land cover classes) will result in a global land cover/forest product of higher accuracy than any of the individual input datasets. In this study we compared some of the most commonly used methods to combine available land cover/forest products into a new forest map. These included: nearest neighbor, Naive Bayes Classifier, logistic regression and geographically weighted logistic regression (GWR), and classification and regression trees (CART). In contrast to non-spatial logistic regression, GWR estimates the regression coefficients for each geographically weighted kernel. As inputs we used crowdsourced data on land cover classes at randomly generated locations (obtained through the Geo-wiki project: geo-wiki.org); and extracted forest masks from global land cover/forest datasets. In general, GWR performed slightly better than the nearest neighbor method. However, in areas with high disagreement between input datasets, the results of the prediction by GWR were found to be much more accurate. From this we conclude that GWR provides the best results for the prediction of land cover classes through combining different data sources.

Keywords: data fusion methods, forest cover, remote sensing

Introduction, scope and main objectives

Land cover maps provide useful information on geographical distribution of different land cover types as well as on land cover changes over time. They are therefore widely used as input data in various applications, such as, climate change models, management of natural resources, environmental monitoring, and comprehensive spatial quantification of ecosystems and landscapes, etc (GCOS 2013). Forest land cover maps provide particularly valuable input datasets in modelling forest growth and productivity.

The last few decades have seen an increase in the number of land cover and forest datasets derived from remote sensing products. The overall trend has been towards higher spatial resolution of land cover maps. Previous studies have shown that combining different sources of information (e.g. remote sensing data and ground-based data on the distribution of land cover classes) will result in a global land cover/forest product of higher accuracy than any of the individual input datasets (Song et al. 2013; Fritz et al. 2011;See et al. 2015; Yu, Wang, and Gong 2013).
There are many data fusion methods available for integrating different sources of data (De'ath and Fabricius 2000; Haapanen et al. 2004; Li et al. 2015). Therefore there is always a question as to which

method to apply, e.g. to combine remote sensing land cover datasets, ecological monitoring data and other ground based data. In this study we compared some of the most commonly used methods to combine available land cover/forest products with crowdsourced data (or citizen science) into a new forest map. These included: nearest neighbor, Naive Bayes Classifier, logistic regression and geographically weighted logistic regression (GWR), and regression trees.

The paper presents the results of comparison of the aforementioned methods when used to produce land cover map of higher accuracy.

Methodology / approach

As mentioned above a number of methods have been applied to generate forest cover map by integrating land cover/forest cover maps with crowdsourced data from Geo-Wiki (Fritz et al. 2012). The methods included: nearest neighbor estimator, Naïve Bayes Classifier, logistic regression and GWR, and regression trees. The resulting map has a resolution of 1 km x 1 km for reference year 2000. The comparison study has been done by using the same input datasets as in work presented by Schepaschenko et al. (2015).

The crowdsourced data (or volunteered geographical information) is a valuable source of data for a variety of applications (Fonte et al. 2015). Here, we obtained the crowdsourced data through Geo-Wiki tool (geo-wiki.org) that is widely used for validation, completion and enhancement of land cover products (Comber et al. 2013; Comber et al. 2012; Fritz et al. 2013). The data set was divided into two subsets: training and testing, and it includes land cover information (presence/absence of forest) for 1 km pixels for a subsets of samples around the globe. The GLC2000 (see below) grid was used as the basis for the output map.

Input data

The following global land cover datasets were utilised in the analysis:

- Global Land Cover Project 2000 (GLC2000, http://bioval.jrc.ec.europa.eu/products/glc2000/products.php):
  The GLC 2000 product was generated by the Global Vegetation Monitoring Unit of the Joint Research Centre (JRC) of the European Commission in collaboration with a network of international partners. The GLC 2000 is a consistent global harmonized land cover database for the environmental reference year 2000 at a spatial resolution of 1 km. The GLC2000 was produced using the VEGA 2000 dataset with 14 months of pre-processed daily global data acquired by the VEGETATION instrument on board the SPOT 4 satellite.

- Global Land Cover by National Mapping Organisations (GLCNO, http://www.iscgm.org/gm/glcnmo.html) 2003:
  The GLCNMO was produced by the Global Mapping Project organized by the International Steering Committee for Global Mapping (ISCGM). The product was generated in a raster format with a resolution of 1 km. It is organized into twenty land cover classes that are standardized by the Land Cover Classification System. As input data, 16-day composite Moderate Resolution Imaging Spectroradiometer (MODIS) data at a 1 km resolution for the year 2003 was used (Tateishi et al. 2008).

- Global Land Cover Product 2005-2006 (GlobCover, http://due.esrin.esa.int/globcover/):
  The GlobCover dataset was produced by the European Space Agency (ESA) which started in 2005 in collaboration with the JRC, EEA (European Environment Agency), FAO, UNEP (United Nations Environment Program), the GOFC-GOLD (Global Forest Cover – Global Land Dynamics) initiative and the International Geosphere-Biosphere Programme(IGBP). The GlobCover product for the

reference year 2005/06 has a spatial resolution of 300 m. A detailed description is provided in (Defourny et al. 2006).

- Landsat-based continuous fields of tree cover 2000 (Landsat VCF)
  Landsat VCF was produced at global level as 30-m resolution continuous fields of tree cover (Sexton et al. 2013). The product for the years 2000 is available from the Global Land Cover Facility (GLCF) website (www.landcover.org).

- Modis land cover product
  The MODIS global land cover type product (MCD12Q1) was obtained through the online Data Pool at the NASA Land Processes Distributed Active Center (LP DAAC), United States Geological Survey (USGS)/Earth Resources Observation and Science (EROS) Centre (https://lpdaac.usgs.gov/data_access). The product was generated at a spatial resolution of 500 m at annual and biannual time-steps. A detailed description of the dataset is given by (Friedl et al. 2010).

- MODIS Vegetation Continuous Fields (VCF) 2000
  The Terra MODIS Vegetation Continuous Fields (VCF) product is derived from the MODIS sensor, on-board the Terra and Aqua satellites at a spatial resolution of 250 m (DiMiceli et al. 2011).

- Landsat-based tree cover 2000 by (Hansen et al. 2013) (Hansen's TC)
  A global forest cover change product for the years 2000–2012 with a spatial resolution of 30 m has been published by (Hansen et al. 2013). The product is based on Landsat imagery and has three components: forest cover 2000, forest gain 2000–2012 and forest loss per year.
- Regional maps:
  In our study we also used the regional maps that contain forest information. To account for regional and local specifications in forest cover, we aggregated a number of regional land cover and land use maps. These maps include: Congo Basin forest types map (OFAC; http://www.observatoire-comifac. net/) that covers eight countries in Central Africa, i.e. Cameroon, Congo, Gabon, Burundi, Central African Republic, Equatorial Guinea, Democratic Republic of Congo and Rwanda (Verhegghen et al. 2012); Brazil PRODES forest mask 2000 (Kempeneers et al. 2012); Land Use of Australia 2005–2006 (http://data.daff.gov.au/brs/data/warehouse/pe_abares99001806/GuidelinesLandUseMappingLowRes2011.pdf); Pan-European Forest/Non-Forest Map 2000 (http://glcf.umiacs.umd.edu/data/landsat/) (Kempeneers et al. 2012); the National Land Cover Database 2006 (NLCD 2006) for theUnited States (available at http://www.mrlc.gov/nlcd06_data.php); Land cover of Russia 2005 (D Schepaschenko et al. 2011); Forest mask for European Russia 2000 (Potapov, Turubanova, and Hansen 2011).

The global land cover datasets were resampled to a 1 km grid using a nearest neighbor technique and applying aggregation rules (see for details Schepaschenko et al. 2015)

Methods overview

For building forest map, we estimate probability of forest presence in each grid cell by applying different methods. The overall idea is to benefit from correlation between global land cover datasets and crowdsourced data. We assume that crowdsourced data from Geo-Wiki is ground-truth geographical information about forest absence/presence (dependent variable), and land cover datasets are independent variables.
Overview of methods compared:

- Nearest Neighbour:
  Nearest Neighbour is one of the simplest methods used in a variety of applications. It uses the mean (for continuous) or median (for categorical) of the variable of interest over the predefined neighbourhood as the estimator. Despite its simplicity in application it usually provides good final results. For example, (Meng et al. 2007) used k-nearest neighbour technique to provide forest inventory with remote sensing data.

- Naïve Bayes Classifier:
  Naïve Bayes approach uses the inverse probability formula to produce the likeliest category estimate. It is commonly empolyed for classification tasks, including land cover classification.

- Logistic regression and geographically weighted logistic regression:
  Ordinary logistic regression was used to generate a global equation to predict the probability of presence of forest cover at global level. GWR estimates model parameters at each geographical location by using a distance weighted kernel, so that the observations closer to the studied location have more influence on the parameter estimates than the observations further away (Fotheringham et al 2002).

- Classification and Regression Trees (CART)(Song et al. 2013):
  Classification trees partition the independent variable space into regions for each of which a single likeliest outcome class is selected.

All of the above methods were implemented using the R environment for statistical computing (R Core Team 2014).

Results

Once a forest map has been generated by each of the methods listed above, we calculated the apparent error rate, sensitivity and specificity (forest/non-forest) for both, the training and the testing datasets. The results are summarised in Table 1 and 2. In the tables' abbreviations NN, RT, BN, GLM, GWR correspond to the methods names Nearest Neighbour, CART, Naive Bayes algorithm, logistic regression and GWR.

The lowest apparent error during methods validation for training data was obtained by GWR. In terms of validation by testing data, GWR performs only slightly worse than Naïve Bayes and GLM. Sensitivity and specificity analysis shows that that forest areas are identified with more precision by GWR than by other methods.

Table 1: Training data. Total apparent error rate, specificity and sensitivity

Method	Apparent error rate	Sensitivity	Specificity
NN	0.119	0.931	0.822
RT	0.127	0.908	0.831
BN	0.132	0.852	0.866
GLM	0.127	0.922	0.815
GWR	0.102	0.938	0.842

Table 2: Testing data. Total apparent error rate, specificity and sensitivity

Method	Apparent error rate	Sensitivity	Specificity
NN	0.105	0.961	0.841
RT	0.103	0.964	0.832
BN	0.092	0.961	0.811
GLM	0.096	0.953	0.857
GWR	0.100	0.961	0.840

We have also considered the relationship between the apparent error rate and the level of agreement between the data sources. In order to do this, we have defined a pixel-specific agreement score as the number of sources, out of 9, indicating forest in that pixel. The scores thus vary from 0 to 9. Both, 0 and 9, indicate high level of agreement with respect to the absence of presence of forest respectively. Figure 1 illustrates how the apparent error rate varies with the agreement score and the statistical method used. We can clearly see that in highly contentious areas (agreement score between 2 and 5). GWR performs the best.

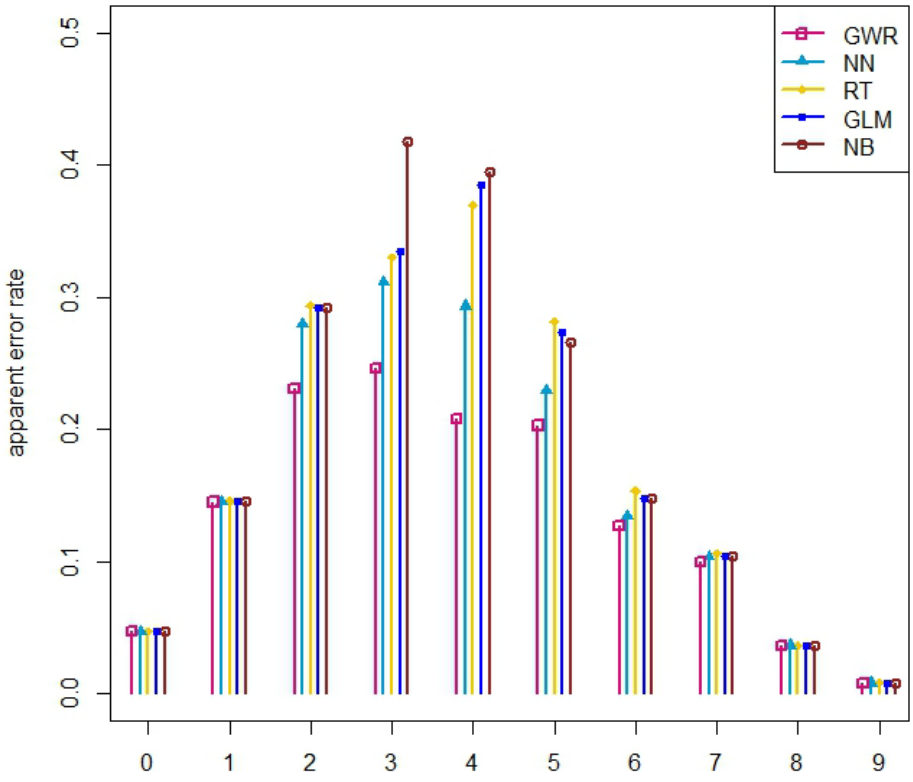

Fig. 1: Apparent error rate for agreement scores by methods (%, training data)

For the testing dataset, the overall performance of the GWR is somewhat worse than for the training dataset, due to the higher overall agreement among the input sources.

Discussion

In general, GWR performed slightly better than other methods. However, in areas with high disagreements between input datasets, the results of the prediction by GWR were found to be much more accurate. From this we conclude that GWR provides the best results for the prediction of land cover classes through combining different data sources. This gain in

accuracy has a trade-off in that it is more computationally intensive than the other methods tested.

The results of the study are significant for building land cover maps of different land cover types. As new land cover products appear, it is always possible to build a hybrid land cover map by applying one of the data fusion methods.

Conclusions / outlook

This paper presents the comparison of performance of selected method in predicting forest cover by integrating land cover datasets and crowdsourced data for Geo-Wiki. Of the methods tested, GWR proved to predict the presence/absence of forest the best. This was especially so in areas with high disagreement among the input data sources. The nearest neighbour method was found to be the second best in terms of prediction accuracy. Because we have found GWR to be substantially more demanding in terms of computing resources, we recommend that both, GWR and NN be considered when producing a land cover map.

Acknowledgements

The work was supported by Marie Curie individual grant FP7-MC-IIF: SIFCAS project No. 627481.

References

1. Comber, Alexis, Peter Fisher, Chris Brunsdon, and Abdulhakim Khmag. 2012. "Spatial Analysis of Remote Sensing Image Classification Accuracy." *Remote Sensing of Environment* 127 (December): 237–46. doi:10.1016/j.rse.2012.09.005.
2. Comber, Alexis, Linda See, Steffen Fritz, Marijn Van der Velde, Christoph Perger, and Giles Foody. 2013. "Using Control Data to Determine the Reliability of Volunteered Geographic Information about Land Cover." *International Journal of Applied Earth Observation and Geoinformation* 23 (August): 37–48. doi:10.1016/j.jag.2012.11.002.
3. De'ath, Glenn, and Katharina E. Fabricius. 2000. "Classification and Regression Trees: A Powerful Yet Simple Technique for Ecological Data Analysis." *Ecology* 81 (11): 3178–92. doi:10.2307/177409.
4. Defourny, Pierre, Christelle Vancustem, P. Bicheron, C. Brockmann, F. Nino, L. Schouten, and M. Leroy. 2006. "GLOBCOVER: A 300m Global Land Cover Product for 2005 Using ENVISAT MERIS Time Series." In *Proceedings of the ISPRS Commission VII Mid-Term Symposium: Remote Sensing: From Pixels to Processes. Enscede NL*. http://metalib-a.lib.ucl.ac.uk/V?func=find-ej-1.
5. DiMiceli, C.M., M.L. Carroll, R.A. Sohlberg, C. Huang, M. C. Hansen, and J. R. G. Townshend. 2011. "*Annual Global Automated MODIS Vegetation Continuous Fields (MOD44B) at 250 M Spatial Resolution for Data Years Beginning Day 65, 2000 - 2010, Collection 5 Percent Tree Cover.*" University of Maryland, College Park, MD, USA. http://glcf.umd.edu/data/vcf/.
6. Fonte, Cidália C., Lucy Bastin, Linda See, Giles Foody, and Favio Lupia. 2015. "Usability of VGI for Validation of Land Cover Maps." *International Journal of Geographical Information Science*, March, 1–23. doi:10.1080/13658816.2015.1018266.
7. Fotheringham, A. Stewart, Chris Brunsdon, and Martin Charlton. 2002. *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*. Chichester: John Wiley & Sons.
8. Friedl, Mark A., Damien Sulla-Menashe, Bin Tan, Annemarie Schneider, Navin Ramankutty, Adam Sibley, and Xiaoman Huang. 2010. "MODIS Collection 5 Global Land Cover: Algorithm Refinements and Characterization of New Datasets." *Remote Sensing of Environment* 114 (1): 168–82. doi:10.1016/j.rse.2009.08.016.
9. Fritz, Steffen, Ian McCallum, Christian Schill, Christoph Perger, Linda See, Dmitry Schepaschenko, Marijn van der Velde, Florian Kraxner, and Michael Obersteiner. 2012. "Geo-Wiki: An Online Platform for Improving Global Land Cover." *Environmental Modelling & Software* 31: 110–23. doi:10.1016/j.envsoft.2011.11.015.
10. Fritz, Steffen, Linda See, Liangzhi You, Chris Justice, Inbal Becker-Reshef, Lieven Bydekerke, Renato Cumani, et al. 2013. "The Need for Improved Maps of Global Cropland." *Eos, Transactions American Geophysical Union* 94 (3): 31–32. doi:10.1002/2013EO030006.
11. GCOS. 2013. *"GCOS Essential Climate Variables."* http://www.wmo.int/pages/prog/gcos/index.php?name=EssentialClimateVariables.

12. Haapanen, Reija, Alan R Ek, Marvin E Bauer, and Andrew O Finley. 2004. “Delineation of Forest/nonforest Land Use Classes Using Nearest Neighbor Methods.” *Remote Sensing of Environment* 89 (3): 265–71. doi:10.1016/j.rse.2003.10.002.
13. Hansen, M. C., P. V. Potapov, R. Moore, M. Hancher, S. A. Turubanova, A. Tyukavina, D. Thau, et al. 2013. “High-Resolution Global Maps of 21st-Century Forest Cover Change.” *Science* 342 (6160): 850–53. doi:10.1126/science.1244693.
14. Kempeneers, P., F. Sedano, A. Pekkarinen, L. Seebach, P. Strobl, and J. San-Miguel-Ayanz. 2012. “Pan-European Forest Maps Derived from Optical Satellite Imagery.” *Earthzine*. July 25. http://www.earthzine.org/2012/07/25/pan-european-forest-maps-derived-from-optical-satellite-imagery/.
15. Li, Weidong, Chuanrong Zhang, MichaelR. Willig, DipakK. Dey, Guiling Wang, and Liangzhi You. 2015. “Bayesian Markov Chain Random Field Cosimulation for Improving Land Cover Classification Accuracy.” *Mathematical Geosciences* 47 (2): 123–48. doi:10.1007/s11004-014-9553-y.
16. Meng, Qingmin, Chris J. Cieszewski, Marguerite Madden, and Bruce E. Borders. 2007. “K Nearest Neighbor Method for Forest Inventory Using Remote Sensing Data.” *GIScience & Remote Sensing* 44 (2): 149–65. doi:10.2747/1548-1603.44.2.149.
17. Potapov, Peter, Svetlana Turubanova, and Matthew C. Hansen. 2011. “Regional-Scale Boreal Forest Cover and Change Mapping Using Landsat Data Composites for European Russia.” *Remote Sensing of Environment* 115 (2): 548–61. doi:10.1016/j.rse.2010.10.001.
18. R Core Team. 2014. *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria.Vienna, Austria. URL http://www.R-project.org/ment for statistical computing.
19. Schepaschenko, D, I McCallum, A Shvidenko, S Fritz, F Kraxner, and M Obersteiner. 2011. “A New Hybrid Land Cover Dataset for Russia: A Methodology for Integrating Statistics, Remote Sensing and in-Situ Information.” *Journal of Land Use Science* 6(4): 245–59. doi:10.1080/1747423X.2010.511681.
20. Schepaschenko, Dmitry, Linda See, Myroslava Lesiv, Ian McCallum, Steffen Fritz, Carl Salk, Elena Moltchanova, et al. 2015. “Development of a Global Hybrid Forest Mask through the Synergy of Remote Sensing, Crowdsourcing and FAO Statistics.” *Remote Sensing of Environment*. doi:10.1016/j.rse.2015.02.011.
21. See, Linda, Dmitry Schepaschenko, Myroslava Lesiv, Ian McCallum, Steffen Fritz, Alexis Comber, Christoph Perger, et al. 2015. “Building a Hybrid Land Cover Map with Crowdsourcing and Geographically Weighted Regression.” *ISPRS Journal of Photogrammetry and Remote Sensing.* Accessed February 10. doi:10.1016/j.isprsjprs.2014.06.016.
22. Sexton, Joseph O., Xiao-Peng Song, Min Feng, Praveen Noojipady, Anupam Anand, Chengquan Huang, Do-Hyung Kim, et al. 2013. “Global, 30-M Resolution Continuous Fields of Tree Cover: Landsat-Based Rescaling of MODIS Vegetation Continuous Fields with Lidar-Based Estimates of Error.” International Journal of Digital Earth 6 (5): 427–48. doi:10.1080/17538947.2013.786146.
23. Song, Xiao-Peng, Chengquan Huang, Min Feng, Joseph O. Sexton, Saurabh Channan, and John R. Townshend. 2013. “Integrating Global Land Cover Products for Improved Forest Cover Characterization: An Application in North America.” *International Journal of Digital Earth*, December, 1–16. doi:10.1080/17538947.2013.856959.
24. Tateishi, R., Bayaer, Ghar, M.A., Al-Bilbisi, H., Tsendayush, J., Shalaby, A., Kasimu, Alimujiang, et al. 2008. “A New Global Land Cover Map, GLCNMO.” The International Archives of the Photogrammetry, *Remote Sensing and Spatial Information Sciences* XXXVII (B7): 1369–72.

25. Verhegghen, A., P. Mayaux, C. De Wasseige, and P. Defourny. 2012. “Mapping Congo Basin Vegetation Types from 300 M and 1 Km Multi-Sensor Time Series for Carbon Stocks and Forest Areas Estimation.” *Biogeosciences* 9 (12): 5061–79. doi:10.5194/bg-9-5061-2012.
26. Yu, Le, Jie Wang, and Peng Gong. 2013. “Improving 30 M Global Land-Cover Map FROM-GLC with Time Series MODIS and Auxiliary Data Sets: A Segmentation-Based Approach.” *International Journal of Remote Sensing* 34 (16): 5851–67. doi:10.1080/01431161.2013.798055.