ARTICLE IN PRESS

ISPRS Journal of Photogrammetry and Remote Sensing xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

journal homepage: www.elsevier.com/locate/isprsjprs

Building a hybrid land cover map with crowdsourcing and geographically weighted regression

Linda See [a,*], Dmitry Schepaschenko [a], Myroslava Lesiv [b,c], Ian McCallum [a], Steffen Fritz [a], Alexis Comber [d], Christoph Perger [a], Christian Schill [e], Yuanyuan Zhao [f], Victor Maus [g], Muhammad Athar Siraj [h], Franziska Albrecht [i], Anna Cipriani [j,k], Mar'yana Vakolyuk [a,r], Alfredo Garcia [l], Ahmed H. Rabia [m], Kuleswar Singha [n], Abel Alan Marcarini [o], Teja Kattenborn [p], Rubul Hazarika [s], Maria Schepaschenko [q], Marijn van der Velde [a], Florian Kraxner [a], Michael Obersteiner [a]

[a] International Institute of Applied Systems Analysis (IIASA), Ecosystem Services and Management Program, Schlossplatz 1, Laxenburg A-2361, Austria
[b] Lviv Polytechnic National University, Lviv, Ukraine
[c] Systems Research Institute of Polish Academy of Sciences, Newelska 6, 01-447 Warsaw, Poland
[d] Department of Geography, University of Leicester, Leicester, UK
[e] Felis, University of Freiburg, Germany
[f] Ministry of Education Key Laboratory for Earth System Modelling, Centre for Earth System Science, Tsinghua University, Beijing 100084, China
[g] National Institute for Space Research (INPE), São José dos Campos, Brazil
[h] Institute of Geographical Information Systems, National University of Sciences and Technology (IGIS-NUST), Islamabad, Pakistan
[i] University of Vienna, Department of Geography and Regional Research, Vienna, Austria
[j] Dipartimento di Scienze Chimiche e Geologiche, Università degli Studi di Modena e Reggio Emilia, Largo S. Eufemia 19, 41121 Modena, Italy
[k] Lamont Doherty Earth Observatory of Columbia University, Palisades, NY 10964, USA
[l] National Institute of Agricultural Technology (INTA), Buenos Aires, Argentina
[m] Faculty of Agriculture, Damanhour University, Egypt
[n] Gauhati University, Assam, India
[o] AgroParis Tech, Nancy, France
[p] Chair of Remote Sensing and Landscape Information Systems, Albert-Ludwigs-University Freiburg, Tennenbacher Str. 4, 79085 Freiburg, Germany
[q] Russian Institute of Continuous Education in Forestry, Pushkino, Russia
[r] Scientific Centre for Aerospace Research of the Earth, Institute of Geological Science of the National Academy of Sciences of Ukraine, Kyiv, Ukraine
[s] B. Borooah College, Guwahati, Assam, India

A R T I C L E I N F O

Article history:
Available online xxxx

Keywords:
Land cover
Validation
Crowdsourcing
Map integration
Global land cover
Geographically weighted regression

A B S T R A C T

Land cover is of fundamental importance to many environmental applications and serves as critical baseline information for many large scale models e.g. in developing future scenarios of land use and climate change. Although there is an ongoing movement towards the development of higher resolution global land cover maps, medium resolution land cover products (e.g. GLC2000 and MODIS) are still very useful for modelling and assessment purposes. However, the current land cover products are not accurate enough for many applications so we need to develop approaches that can take existing land covers maps and produce a better overall product in a hybrid approach. This paper uses geographically weighted regression (GWR) and crowdsourced validation data from Geo-Wiki to create two hybrid global land cover maps that use medium resolution land cover products as an input. Two different methods were used: (a) the GWR was used to determine the best land cover product at each location; (b) the GWR was only used to determine the best land cover at those locations where all three land cover maps disagree, using the agreement of the land cover maps to determine land cover at the other cells. The results show that the hybrid land cover map developed using the first method resulted in a lower overall disagreement than the individual global land cover maps. The hybrid map produced by the second method was also better when compared to the GLC2000 and GlobCover but worse or similar in performance to the MODIS land cover product depending upon the metrics considered. The reason for this may be due to the use of the GLC2000 in the development of GlobCover, which may have resulted in areas where both maps agree with one another but not with MODIS, and where MODIS may in fact better represent land

* Corresponding author. Tel.: +43 2236 807 423.
E-mail address: see@iiasa.ac.at (L. See).

http://dx.doi.org/10.1016/j.isprsjprs.2014.06.016

Please cite this article in press as: See, L., et al. Building a hybrid land cover map with crowdsourcing and geographically weighted regression. ISPRS J. Photogram. Remote Sensing (2014), http://dx.doi.org/10.1016/j.isprsjprs.2014.06.016

cover in those situations. These results serve to demonstrate that spatial analysis methods can be used to improve medium resolution global land cover information with existing products.

1. Introduction

Spatially explicit information about land cover is of fundamental importance for many applications including nature protection and biodiversity, forest and water management, urban and transport planning, natural hazard prevention and mitigation, and the evaluation of agricultural policies. The importance of global land cover is recognized through its status as an essential climate variable (GCOS, 2013), where this information serves as a critical input to the monitoring of climate change. Global land cover forms a key input to large scale economic land use models (e.g. Havlík et al., 2011), which are used to determine important quantities such as the amount of land available for agricultural expansion, afforestation projects and biofuel production or whether reducing emissions from deforestation and forest degradation (REDD) are the most cost-effective solutions. A critical gap in accurate land cover and land use, which is needed to monitor ecosystem services and change over time, has also been highlighted recently by Tallis et al. (2012).

A number of different coarse to medium resolution global land cover products exist, e.g. the GLC2000 (Fritz et al., 2003), MODIS (Friedl et al., 2010) and GlobCover (Bicheron et al., 2008). These products, which vary from 1 km to 300 m resolution at the equator, have been developed using data from different satellite sensors and using different classification algorithms with varying degrees of automation. Although the published accuracies of these products vary between 68.5% and 74.8%, recent studies have shown that when these maps are compared, there are significant amounts of spatial disagreement across different land cover types, in particular in the cropland and forest domains even when taking semantic differences in the legend definitions into account (Fritz and See, 2008; Fritz et al., 2011a). Research has also shown that model outcomes can vary significantly when different land cover products are used in the same modelling exercise (Quaife et al., 2008; Seebach et al., 2012) while Fritz et al. (2012a) have demonstrated the value associated with reducing the uncertainty in land cover with regards to the cost of different climate mitigation options.

With the opening up of the Landsat archive (Wulder et al., 2012), one of the most recent trends in global land cover mapping has been to produce higher resolution products, i.e. at 30 m (Gong et al., 2013; Yu et al., 2013b), with others currently in the pipeline by groups in China and the USA. The accuracies of these recently produced 30 m products range from 63.7% to 66.0%. The technology and algorithms for classifying Landsat will undoubtedly improve in the future, and there will be new higher resolution sensors coming online soon where the data will be freely available (e.g. Sentinel II). Moreover, there are other multi-temporal and/or multi-sensor classification efforts ongoing (Lu et al., 2011; Roy et al., 2010). Despite this relatively positive outlook for land cover mapping in the future, there is still an urgent need for better land cover maps at the present time. Medium resolution products are also still extremely useful from a modelling and assessment point of view where the issue is not one of needing to improve the resolution for many applications but simply improving the accuracy.

One method which can be used to address this issue of accuracy is to merge existing land cover maps to create an integrated or hybrid product where the resulting accuracy should be higher than the accuracies of the individual products. Data fusion and soft computing are domains which are based on the integration of data from a variety of sources (e.g. from different sensors, models or approaches) so this idea is not new in itself. For example, Jung et al. (2006) developed a fuzzy agreement scoring method to determine the synergies between global land cover products for modelling the carbon cycle while Fritz et al. (2011b) employed this synergy concept in combination with expert knowledge to rank land cover products in order to combine them into a single cropland map of Africa. Iwao et al., 2011 integrated the GLC2000, MODIS and the older University of Maryland (UMD) land cover product using a simple majority voting approach and validated the resulting map with data from the Degrees of Confluence project. However, the resulting improvements in accuracy were not statistically significant. More recently, Yu et al. (2013a) used a decision tree to combine two 30 m cropland products with a 250 m cropland probability layer to produce a global cropland mask. All of these approaches have demonstrated the increased accuracy that has resulted from the integration of existing products.

What these types of integration methods need are much larger amounts of data for training and validation. One potential source is Geo-Wiki, which is a visualisation, crowdsourcing and validation tool developed to help improve global land cover maps (Fritz et al., 2012b, 2009) where crowdsourcing is the use of the volunteers (which can also be experts) to help collect and analyse data (Howe, 2006; Heipke, 2010). Using Google Earth, volunteers are asked to indicate the land cover types that are visible from the images displayed in Geo-Wiki. Samples have been collected through a number of different Geo-Wiki campaigns that have run over the past few years (Perger et al., 2012; See et al., 2014a) and then used in subsequent research, e.g. Fritz et al. (2013a) used data from the first campaign to downgrade estimates of land availability for biofuels. This database, which represents a valuable source of data for both training and validation of land cover, continues to grow, with more than 4.5 million samples collected recently on the presence of cropland using a game version of Geo-Wiki (See et al., 2014b). Initial attempts were undertaken by Comber et al. (2013) to map the areas of highest correspondence between the Geo-Wiki crowdsourced data and the GLC2000, MODIS and GlobCover for one land cover type, i.e. tree cover, for a section of western Africa. The authors employed crowdsourced data from the first Geo-Wiki competition (Perger et al., 2012) and geographically weighted regression (GWR), which is a spatial extension to linear regression in which the coefficients of the regression equation are able to vary across space, which captures any effects due to location (Brunsdon et al., 1998).

The aim of this paper is to extend the research of Comber et al. (2013) in a number of ways. Firstly, we apply the method globally to all land cover types using larger amounts of crowdsourced data from four campaigns, including one campaign focussed on data collection in areas where all global land cover maps disagree with one another. Secondly, we implement a second approach in which the method is only applied to those areas where there is complete disagreement between land cover products, taking agreement between two or more land cover products as the land cover type at all other locations. Finally, we use an independent crowdsourced dataset to validate the products, using the sampling scheme of the global validation dataset created by Zhao et al. (2014) as the basis