Science Bulletin 64 (2019) 370–373

Contents lists available at ScienceDirect

Science Bulletin

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

Science Bulletin
www.scibull.com

Short Communication

Stable classification with limited sample: transferring a 30-m resolution sample set collected in 2015 to mapping 10-m resolution global land cover in 2017

Peng Gong [a,b,*], Han Liu [a], Meinan Zhang [a], Congcong Li [c], Jie Wang [b,d,*], Huabing Huang [d,*], Nicholas Clinton [e,*], Luyan Ji [f], Wenyu Li [a], Yuqi Bai [a], Bin Chen [a], Bing Xu [a], Zhiliang Zhu [g], Cui Yuan [a], Hoi Ping Suen [a], Jing Guo [a], Nan Xu [a], Weijia Li [a], Yuanyuan Zhao [a], Jun Yang [a], Chaoqing Yu [a,b], Xi Wang [a,b], Haohuan Fu [a,h], Le Yu [a], Iryna Dronova [i], Fengming Hui [j], Xiao Cheng [j], Xueli Shi [k], Fengjin Xiao [k], Qiufeng Liu [k], Lianchun Song [k]

[a] Ministry of Education Key Laboratory for Earth System Modeling, Department of Earth System Science, Tsinghua University, Beijing 100084, China
[b] AI for Earth Lab, Cross-Strait Institute, Tsinghua University, Beijing 100084, China
[c] Department of Environmental Science, Policy and Management, University of California, Berkeley, CA 94720, USA
[d] State Key Laboratory of Remote Sensing Science, Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, Beijing 100101, China
[e] Google LLC, Mountain View, CA 94043, USA
[f] Institute of Electronics, Chinese Academy of Sciences, Beijing 100190, China
[g] United States Geological Survey, Reston, VA 20192, USA
[h] National Supercomputing Center in Wuxi, Wuxi 214072, China
[i] Department of Landscape Architecture, University of California, Berkeley, CA 94720, USA
[j] State Key Laboratory of Remote Sensing Science, College of Global Change and Earth System Science, Beijing Normal University, Beijing 100875, China
[k] National Climate Center, China Meteorological Administration, Beijing 100081, China

ARTICLE INFO

Article history:
Received 25 December 2018
Received in revised form 26 February 2019
Accepted 27 February 2019
Available online 2 March 2019

As the world strives to reduce the impact of population growth, urbanization, agricultural expansion, and climate change on food security, energy and water shortage, resource over-exploration, biodiversity loss, environmental pollution, and ultimately human health, timely and higher resolution land cover information is urgently needed to achieve the sustainable development goals of the United Nations. Finer than 100-m resolution land cover mapping of the entire world was not in place until the 2010 s when 30-m resolution Landsat images covering the world were made freely available [1]. However, more and more applications such as crop field mapping, solar energy planning, wildlife management, and urban planning require not only higher resolution but also more frequent global land cover maps. The large volume of data required makes it more labor and computation intensive to develop finer resolution and more frequent maps. The labor intensiveness is not only a restrictive factor to visual interpretation of high resolution images but also a huge burden to training sample collection in automated mapping. Fortunately, with continuing training sample collection and accumulation from our previous efforts [2,3], and both the complete storage and free accessibility of the 10-m resolution Sentinel-2 images and the huge computing capability provided by Google Earth Engine, we are in an advanced position to develop a 10-m resolution global land cover map.

A multi-seasonal sample set including a training set and a validation set has been collected from Landsat 8 images acquired in 2014 and 2015. The training set contains approximately 340,000 sample units of various sizes (from 30 m × 30 m to 500 m × 500 m) located at approximately 93,000 sites worldwide [3]. The validation set contains approximately 140,000 sample units of land cover type in different seasons at over 38,000 locations. Our previous experiments have indicated that a random forest classifier is both computation efficient and optimal in performance when dealing with high dimensionality of data [4]. Preliminary test results indicated that an overall classification accuracy could be achieved at better than 71% using the 2015 training and validation sample sets [3]. Our goal here is to apply the training sample set to Sentinel-2 images acquired in 2017 to produce a 10 m resolution global land cover map with the random forest classifier. The question is whether or not we can directly

* Corresponding authors.
E-mail addresses: penggong@tsinghua.edu.cn (P. Gong), sohuwangjie@163.com (J. Wang), huanghb@radi.ac.cn (H. Huang), nclinton@google.com (N. Clinton).

https://doi.org/10.1016/j.scib.2019.03.002
2095-9273/

apply the 2015 sample sets to 2017 images acquired with a different sensor.

For the purpose of sample transfer to data acquired in other years or from different sensors, we wish to know how small a sample set could be sufficient to allow us to achieve a relatively consistent classification result. At the global scale, only a small percentage of the territory in the world would change land cover types due to human activities of land clearing or natural forces such as wildland fires, volcanos, hurricanes, etc. The annual percentage of land cover change can hardly exceed a few percent of the total land area on Earth. Therefore, it is meaningful to find out to what extent when the training sample is so obsolete that it is no longer suitable for transferring the sample to different years or to data obtained from a different sensor. On the other hand, training sample points are collected through image interpretation. The best image interpreter may still make 5%–10% of interpretation errors [2]. How tolerable is a classification algorithm to training sample errors introduced by image interpreters? For any classifier, its sensitivity to a smaller sized training sample or its tolerance to training errors or actual land cover changes from year to year should be determined.

Here we define the concept of a stable classification. We use this concept to approximately determine how much reduction in training sample and how much land cover change or image interpretation error can be acceptable. If the mean accuracy of multiple runs of a classifier trained with a random drawing of a certain percentage of sample points from the total sample is within 1% of what can be achieved with the total sample set, we regard the obtained classification result "stable". The 1% threshold is empirically chosen based on the fact that a loss of overall accuracy in 1% shall not significantly impact the application of a global land cover map. Using a random forest algorithm with 200 trees (as explained later, this is the optimal performer from our experience), we conducted an experiment to find out how "limited" can the training sample be while a stable classification can still be maintained. For the experimentation, we used our 2015 training and validation sample sets. The classification features include 9 Landsat-8 image bands; indices of vegetation (normalized difference vegetation index, enhanced vegetation index), water (modified normalized difference water index), built-up (normalized difference built-up index), and burning (normalized burn ratio); 25%, 50%, 75%, percentile of the annual time-series image spectral values; mean and standard deviations of each of the previous features; elevation, slope, aspect, and hill shadow; and longitude and latitude of the sample location. We designed two sets of experiments. In the first, we gradually reduce the number of training sample points by 1% each time and randomly repeat this process for 1,000 times. In the second, we randomly alter the category of a certain percent of the total sample and used the "noisy" sample to train the random forest classifier. We began to alter the land cover types in 5% increments of the total training sample. We repeated the experiments until 45% of the total sample were altered. For each increment of sample alteration, we randomly alter sample points for 1,000 times. In both sets of experiments we tested the classification accuracy using the validation sample set. The results are presented in Fig. 1. It can be seen that the mean overall accuracy of the sample reduction is very stable until as few as 40% of the training sample are used (72.15% vs. 73.13% obtained with the entire sample). Therefore, it is safe to state that we need only to use approximately 40% of the total sample to keep the classification stable (Fig. 1a). On the other hand, it can be seen from Fig. 1b that when the "error" (altered classes) of the training sample reaches 20%, the mean accuracy is still within 1% from that obtained with un-altered training sample. So the tolerance range of sample error can be set to 20%. These experiments suggest that it is possible to use 60% fewer sample points and even the land cover changed by 20% or the training sample contains 20% errors, we are still able to achieve "stable" classification with the random forest classifier in global land cover mapping. Therefore, we felt safe to transfer the entire training sample in classifying Sentinel-2 images obtained in 2017 because we assumed that the land cover types in the world did not change by more than 5% from 2015 to 2017. It should be possible to produce a stable land cover map based on our circa 2015 training sample set [3].

Since its launch in 2015, Sentinel-2 acquires data in 13 spectral bands including four 10-m resolution visible and near infrared bands, six 20-m resolution red-edge and middle infrared spectral bands, and three additional bands measuring atmospheric conditions. We used all but the atmospheric bands. After tests and adjustment, Sentinel-2 acquired more images covering the world in 2017. Therefore, we used 2017 data in mapping 10-m global land cover. To extend the samples for use in 10-m resolution Sentinel-2 images, we used the center of each sample location to match the nearest locations of the Sentinel data to extract and construct spectral features. Elevation data from Shuttle Radar Topographic Mission (SRTM) were also used as ancillary data (https://doi.org/10.1029/2005RG000183). The input features include the spectral values of the greenest time in each year, the 0, 25, 50, 75 and 100 percentile of time series, and indices of vegetation, water, building and snow as above mentioned in classify-

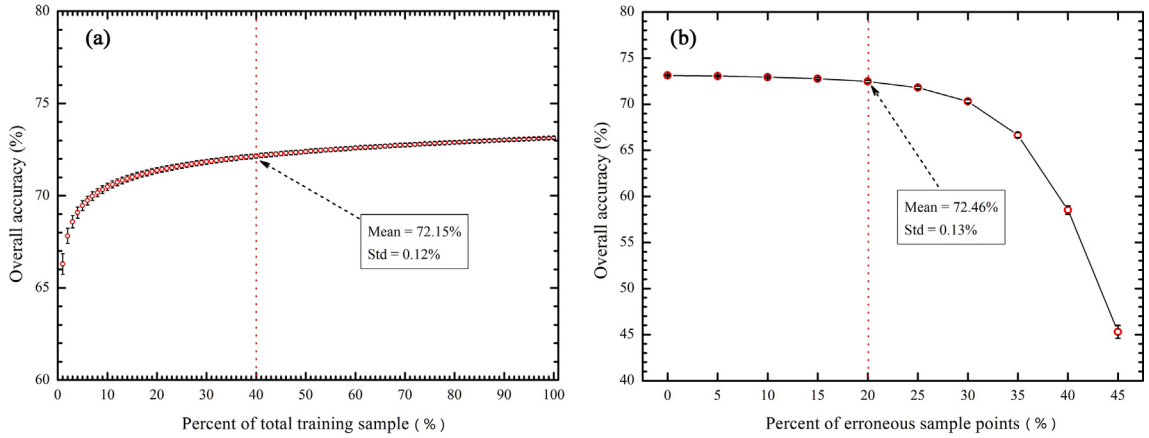

Fig. 1. Sample robustness to size reduction and errors in sample. (a) As sample size increases, the accuracy quickly reaches a plateau. (b) As the impurity percentage of sample increases the accuracy decreases. In both cases, the 1,000 times random drawing of sample points produced very stable overall classification accuracies with most standard deviations much lower than 0.5%.

Table 1
Confusion matrix for the 2017 global land cover map, FROM-GLC10, obtained from Sentinel-2 data.[a)]

Classification	CR	FR	GR	SR	WE	WB	TU	IA	BL	SI	Total	PA (%)
Cropland	1864	262	629	205	2	4	0	33	53	0	3052	61.07
Forest	304	7951	628	455	5	9	48	17	25	1	9443	84.20
Grassland	441	502	4378	632	15	15	111	66	625	4	6789	64.49
Shrubland	203	680	1083	2444	8	7	33	10	346	0	4814	50.77
Wetland	25	22	89	13	30	66	13	1	33	4	296	10.14
Water	14	10	23	2	15	1238	10	1	19	9	1341	92.32
Tundra	0	45	164	5	1	8	1089	0	75	14	1401	77.73
Impervious	36	5	34	6	0	0	0	231	8	0	320	72.19
Barren	317	12	416	128	9	31	43	15	5507	33	6511	84.58
Snow/ice	2	37	38	5	2	55	88	0	74	743	1044	71.17
Total	3206	9526	7482	3895	87	1433	1435	374	6765	808	35011	
UA(%)	58.14	83.47	58.51	62.75	34.48	86.39	75.89	61.76	81.40	91.96		72.76

[a)] Overall accuracy = 72.76%, CR = cropland, FR = forest, GR = grassland, SR = shrubland, WE = wetland, WB = water body, TU = tundra, IA = impervious area, BL = bare land, SI = snow/ice, UA = user's accuracy and PA = producer's accuracy.

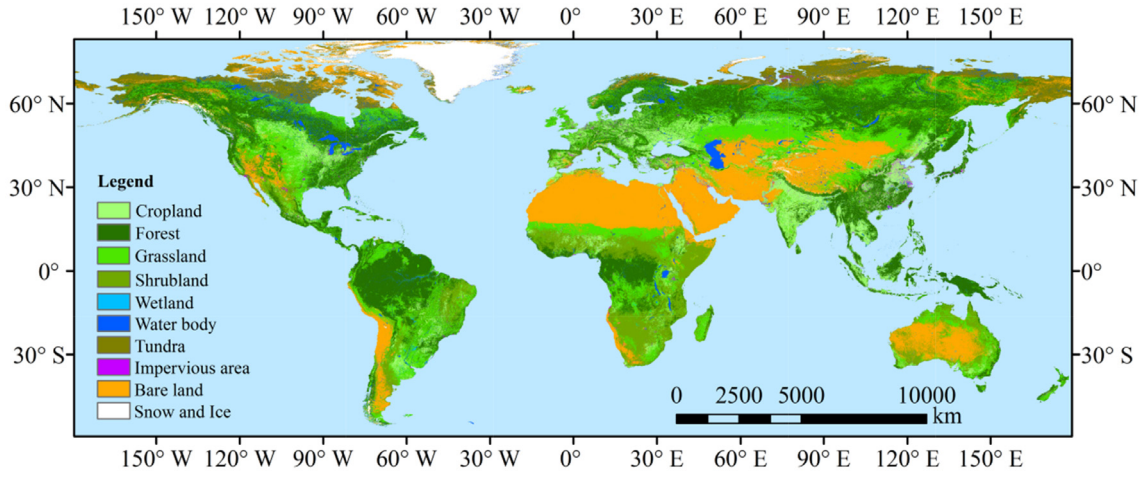

Fig. 2. Global land cover map, FROM-GLC10, based on 10-m resolution Sentinel-2 data acquired in 2017.

ing Landsa-8 data. Slope and aspect data extracted from the SRTM elevation data are also included in the feature set. In addition, the geographical coordinates including the longitude and latitude are also used. The 10 classes in the first level classification scheme used in Gong et al. [1] has been used except the cloud category (Table 1). A random forest classifier had been used with 200 trees. Other parameters in the algorithm were optimized through a grid searching method. This algorithm has its merit for robustness, high efficiency and high accuracy outcome in processing high dimensional data [5,6]. It had been extensively tested in our previous global land cover mapping efforts [4,7].

Using the freely accessible data and the programming environment within Google Earth Engine, we obtained the 10-m resolution global land cover map based on Sentinel-2 data (Fig. 2). We named it FROM-GLC10. The overall accuracy validated against the circa 2015 validation sample [3] is 72.76% (Table 1). From Fig. 2, it can be seen that the stamping effect caused by use of single date images for land cover classification does not exist in this map. In addition, the cloud impact is small even at the most frequent cloud occurrence regions in the low latitudes, particularly in West Africa and the eastern edge of Tibet Plateau. This proves the value of the use of all images acquired in a year in a time-series to derive land cover types over large areas. From Table 1, it can be seen that land cover types occupying a large proportion of the Earth surface, or categories with relatively pure spectral properties can usually be well classified with relatively high producer's and user's accuracies (e.g., Forest, Barren lands and Water Bodies). As usual, the wetland category is most difficult to map automatically because it could be of any surface cover type as long as it is developed in wet areas. Cropland, grassland, shrubland and impervious area are classified at medium level of accuracy. Their accuracies are considerably higher than those first mapped with single-date images [1].

Using the circa 2015 30-m resolution global sample set, we also produced a global land cover map based on all images acquired by Landsat-8 in 2017. Because this is a sample transfer in the same resolution as in 2015, here we named it FROM-GLC30. Checked against the 2015 validation sample set, the overall accuracy for the 2017 FROM-GLC30 is 72.43%. We made a comparison between FROM-GLC10 and FROM-GLC30 (Fig. S1 online). It can be seen that the most noticeable differences are that FROM-GLC10 shows more spatial detail than FROM-GLC30 (e.g., A, E, G, H in Fig. S1 online). In some cases, FROM-GLC10 can better distinguish forest from shrub or grassland classes (B and C in Fig. S1 online). In mountain shadow areas, and high latitude swamp areas, FROM-GLC10 can reduce water mis-classification over shaded areas or better map water bodies (D and I in Fig. S1 online). FROM-GLC10 also performs well in coastal areas (F in Fig. S1 online) and is better at highlighting aquaculture activities (G in Fig. S1 online).

In summary, we show that, with a random forest classifier and the same set of image classification features, less than 1% overall accuracy is lost when less than 40% of our total global training sample set are used. In the meantime, less than 1% overall accuracy loss can be obtained even when 20% of the global training sample points are in error. This indicates that when a sample size is substantially smaller than its current size or 1 in every 5 sample points is in error a stable classification can still be achieved. Similar efforts

can be found in urban classification [8], crop field classification [9], and land cover mapping at the continental scale [10] but have never been tested at the global scale. The experiments presented in Fig. 1 suggest that a theory on "stable classification with limited sample" exists at the global scale although sample size and percentage of sample error may vary from sensor to sensor and classifier to classifier. Using this theory we successfully transferred our global training sample set developed in 2015 at 30-m resolution to classify 10-m resolution images acquired in 2017 with a sensor on board a different satellite. We examined through the 10-m resolution map, FROM-GLC10, and compared it with our 2017 30-m global land cover map, FROM-GLC30. We found while the results are comparable the 10-m map did provide more spatial details. Although an overall accuracy comparable to the 30-m resolution data was achieved, the actual accuracy of the 10-m resolution map can only be properly assessed with test samples collected from the 10-m resolution data. The map can be freely downloadable from http://data.ess.tsinghua.edu.cn.

Conflict of interest

The authors declare that they have no conflict of interest.

Acknowledgments

The idea of stable classification with limited sample was brought into being through insightful discussions with Professor Guanhua Xu. This work was partially supported by the National Key Research and Development Program of China (2016YFA0600103), a donation made by Delos Living LLC, and the Cyrus Tang Foundation.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.scib.2019.03.002.

References

[1] Gong P, Wang J, Yu L, et al. Finer resolution observation and monitoring of global land cover: first mapping results with Landsat TM and ETM+ data. Int J Remote Sens 2013;34:2607–54.
[2] Zhao Y, Gong P, Yu L, et al. Towards a common validation sample set for global land-cover mapping. Int J Remote Sens 2014;35:4795–814.
[3] Li C, Gong P, Wang J, et al. The first all-season sample set for mapping global land cover with Landsat-8 data. Sci Bull 2017;62:508–15.
[4] Yu L, Wang J, Li X, et al. A multi-resolution global land cover dataset through multisource data aggregation. Sci China Earth Sci 2014;57:2317–29.
[5] Breiman L. Random forests. Mach Learn 2001;45:5–32.
[6] Banfield RE, Hall LO, Bowyer KW, et al. A comparison of decision tree ensemble creation techniques. IEEE Trans Pattern Anal 2007;1:173–80.
[7] Wang J, Zhao Y, Li C, et al. Mapping global land cover in 2001 and 2010 with spatial-temporal consistency at 250 m resolution. ISPRS J Photogramm 2015;103:38–47.
[8] Li C, Wang J, Wang L, et al. Comparison of classification algorithms and training sample sizes in urban land classification with Landsat Thematic Mapper imagery. Remote Sens 2014;4:964–83.
[9] Li C, Gong P, Wang J, et al. An all-season sample database for improving land-cover mapping of Africa with two classification schemes. Int J Remote Sens 2016;37:4623–47.
[10] Pelletier C, Valero S, Inglada J, et al. Effect of training class label noise on classification performances for land cover mapping with satellited image time series. Remote Sens 2017;9:173.

Peng Gong is currently a professor and Chair of the Department of Earth System Science, Dean of School of Sciences at Tsinghua University. His major research interests include mapping and monitoring of global environmental change using satellite and ground based sensors, and modeling of environmentally related infectious diseases.

Jie Wang is an assistant professor in the Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences and Chief Engineer in the AI for Earth Lab, Tsinghua Cross-Strait Institute. His major research interest is machine learning for land monitoring.

Huabing Huang is an associate professor in the Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences. His major research interests include global vegetation height and biomass monitoring and training sample transfer in land cover mapping.

Nick Clinton is on the Earth Engine developer relations team. He received a bachelors, masters and Ph.D. from the Department of Environmental Science, Policy and Management at UC Berkeley. From 2012–2015, he was on the faculty of the Center for Earth System Science at Tsinghua University, in Beijing, China. He joined Google in 2015.