IGARSS 2021 - 2021 IEEE International Geoscience and Remote Sensing Symposium | 978-1-6654-0369-6/21/$31.00 ©2021 IEEE | DOI: 10.1109/IGARSS47720.2021.9553499

Global land use / land cover with Sentinel 2 and deep learning

Krishna Karra, Caitlin Kontgis, Zoe Statman-Weil, Joseph C. Mazzariello, Mark Mathis, Steven P. Brumby

Impact Observatory, Inc., Washington, D.C.

ABSTRACT

Land use/land cover (LULC) maps are foundational geospatial data products needed by analysts and decision makers across governments, civil society, industry, and finance to monitor global environmental change and measure risk to sustainable livelihoods and development. There is a strong need for high-level, automated geospatial analysis products that turn these pixels into actionable insights for non-geospatial experts. The Sentinel 2 satellites, first launched in mid-2015, are excellent candidates for LULC mapping due to their high spatial, spectral, and temporal resolution. Advances in deep learning and scalable cloud-based compute now provide the analysis capability required to unlock the value in global satellite imagery observations. Based on a novel, very large dataset of over 5 billion human-labeled Sentinel-2 pixels, we developed and deployed a deep learning segmentation model on Sentinel-2 data to create a global LULC map at 10m resolution that achieves state-of-the-art accuracy and enables automated LULC mapping from time series observations.

Index Terms— land use land cover, deep learning, segmentation, Sentinel 2

1. INTRODUCTION

Over the last several decades, human-induced land use/land cover (LULC) change has affected ecosystems across the globe [1]. While there is an unprecedented amount of Earth observation imagery data available to track change, analyses must be automated in order to scale globally. To the best of our knowledge, no time-series product is available that maps global LULC at 10m Sentinel-2 resolution. Fortunately, recent advances in large scale, low cost cloud computing make such research feasible.

We trained a deep learning based segmentation model from scratch that leverages a very large (over 5 billion human-labeled pixel) training dataset, and can be run annually on Sentinel-2 imagery to provide up-to-date information about global LULC. This work demonstrates an improvement over current global LULC maps that are produced at a coarser resolution, e.g. 500m MODIS Land Cover Type product [2], 300m European Space Agency Climate Change Initiative (CCI) Land Cover product [3], 100m Copernicus Global Land Cover [4], or 30m USGS National Land Cover Database [5] and the Land Change Monitoring, Assessment, and Projection (LCMAP) [6].

2. METHODS

2.1. Training Data

To create the LULC classification algorithm, we need a global, geographically-balanced training dataset. For this, we use over 24,000 5km x 5km image chips that were hand-labeled into ten classes (water, trees, grass, flooded vegetation, crops, scrub/shrub, built area, bare ground, snow/ice, and clouds) and collected across 14 major biomes (as defined in [7]) employing a random stratified sampling approach [8]. Annotators used dense markup instead of single pixel labels by drawing vector boundaries around individual feature classes on a scene, shown in Figure 1. Dense labeling enables deep learning algorithms to explore both spatial and spectral features for categorizing images, and permits a much faster recovery of per-pixel labels compared to single pixel annotation.

2.2. Model Development

Using the hand-labeled data described in 2.1, we trained a large UNet model from scratch. This is a convolutional neural network architecture originally developed for biomedical image segmentation that has also proven to be effective at semantic segmentation tasks on satellite images [9-11]. We formulate the segmentation task as a pixel-wise categorical classification problem with ten classes as described in 2.1 and an additional “no data” class for unlabeled pixels. We utilize the categorical cross entropy loss function, using an inverse-log weighting based on the percent proportion of each class (shown in Figure 2) to account for class imbalance in the dataset. The class weight for unlabeled pixels (i.e., no data) is zeroed out, which forces the model to ignore unlabeled regions in the image during training.

We utilize six bands (red, green, blue, nir, swir1, swir2) of Sentinel 2 L2A surface reflectance corrected imagery. These bands were chosen to balance model complexity with relevant information for the model

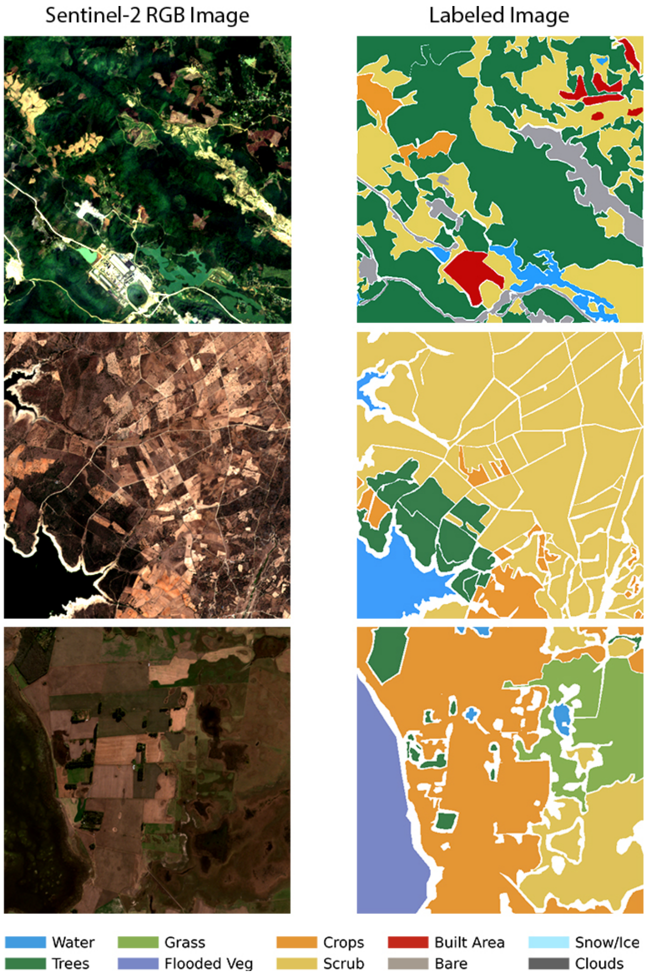


Fig. 1. Three example image (left) and target (right) pairs from the training dataset. Annotators use dense markup by drawing vector boundaries around individual feature classes in a scene.

informed by commonly used techniques from the remote sensing domain. Every band is converted to floating point and scaled between 0 and 1. Data augmentation is applied by randomly flipping the images vertically and horizontally, which has the effect of introducing more geographic pattern realizations. To combat overfitting, we employ the dropout technique during training, randomly turning off 20% of the neurons in the UNet in every batch [12]. The model is trained for 100 epochs to convergence, with a stepped learning rate that drops an order of magnitude after validation loss begins to plateau.

2.3. Model Deployment

Due to cloud cover variability and gaps in data coverage, it is necessary to incorporate multiple observations in order to create a seamless annual LULC map. Our scene selection logic chooses the least cloudy scenes per 100km x 100km Sentinel 2 tile, incorporating more scenes in very cloudy areas and less scenes in mostly cloud-free areas. Each scene from a Sentinel-2 tile is chipped into hundreds of 5km x 5km image chips with 1km overlap. A final LULC map is generated by computing a class weighted mode across all the model predictions, which incorporates the classification, the associated probability, and a custom class weight per pixel. The custom weight per class, akin to a seasonal adjustment, emphasizes ephemeral classes that may only occur a few times per year (e.g. grass) and de-emphasizes classes that are transient (e.g. snow/ice).

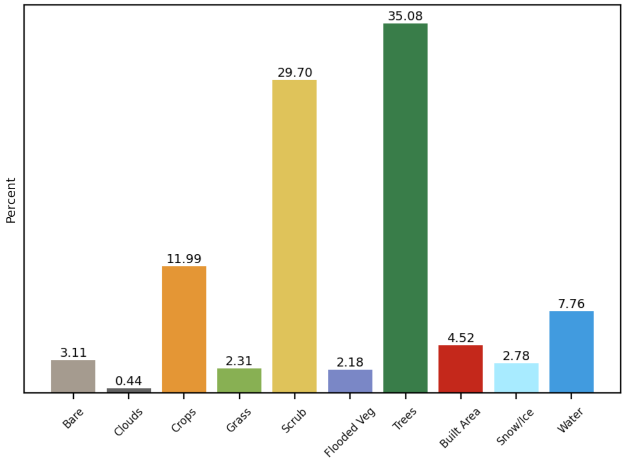
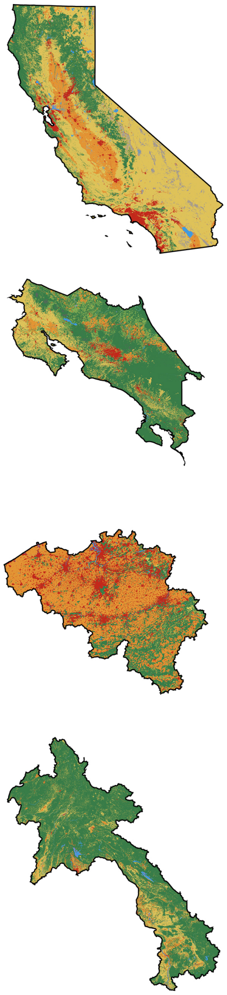


Fig. 2. Class distribution across the training dataset, which spans the globe across 14 major biomes. We see that the classes are unevenly distributed, and we account for this imbalance by weighting under-represented classes higher in the model.

Following this methodology, we generate a global LULC map by processing over 20,000 Sentinel 2 tiles across the entire Earth's land surface. Sentinel 2 surface reflectance corrected imagery was accessed from the Microsoft Planetary Computer. The global run requires approximately 1.2 million core hours of compute time, which we execute in Microsoft Azure Batch, running up to 6400 cores simultaneously. By leveraging cloud computing resources in this fashion, we are able to create a global LULC map at 10 meter resolution that incorporates imagery across the entire year in approximately 7 days.

2.4. Model Evaluation

During training, we set aside 15% of the data for validation at the end of every epoch to ensure we are not overfitting. In addition, we utilized a "gold standard" set of tiles labeled by multiple expert annotators over 409 5km x 5km sample areas, following a similar sampling approach as outlined in 2.1. These validation tiles were entirely excluded from training, and represent labels where there is strict agreement among 3 expert annotators. All error matrices report validation statistics compared to these holdout tiles.



California - Overall Accuracy 85%

Class	Water	Trees	Grass	Flooded Veg	Crops	Scrub	Built Area	Bare	Total	User's	Producer's	Overall	Area [ha]
Water	0.01195	0.00125	0.00004	0.00020	0.00035	0.00042	0.00018	0.00002	0.01440	0.82997	0.89298	0.84975	7903 ± 1081
Trees	0.00017	0.19758	0.00079	0.00097	0.00395	0.01290	0.00219	0.00014	0.21869	0.90349	0.81771		2167043 ± 10057
Grass	0.00004	0.00036	0.00442	0.00041	0.00358	0.00134	0.00122	0.00012	0.01148	0.38473	0.20998		9896 ± 5905
Flooded Veg	0.00001	0.00004	0.00001	0.00099	0.00043	0.00009	0.00013	0.00001	0.00172	0.57557	0.22415		310 ± 2081
Crops	0.00005	0.00119	0.00213	0.00017	0.09875	0.00482	0.00129	0.00053	0.10894	0.90650	0.86840		508039 ± 4280
Scrub	0.00034	0.04103	0.01363	0.00157	0.00627	0.45416	0.02716	0.00228	0.54645	0.83111	0.93078		10934894 ± 14215
Built Area	0.00016	0.00005	0.00001	0.00000	0.00032	0.00155	0.05999	0.00054	0.06263	0.95789	0.64959		237195 ± 8207
Bare	0.00067	0.00012	0.00001	0.00009	0.00006	0.01266	0.00018	0.02191	0.03571	0.61359	0.85758		37417 ± 2917
Total	0.01338	0.24163	0.02103	0.00441	0.11372	0.48794	0.09235	0.02555					

Costa Rica - Overall Accuracy 84%

Class	Water	Trees	Grass	Flooded Veg	Crops	Scrub	Built Area	Bare	Total	User's	Producer's	Overall	Area [ha]
Water	0.00673	0.00071	0.00002	0.00011	0.00019	0.00023	0.00010	0.00001	0.00810	0.82997	0.86189	0.84067	320 ± 116
Trees	0.00046	0.53860	0.00214	0.00265	0.01077	0.03515	0.00597	0.00039	0.59614	0.90349	0.96646		1685738 ± 822
Grass	0.00035	0.00302	0.03742	0.00345	0.03034	0.01132	0.01035	0.00102	0.09726	0.38473	0.82073		22502 ± 768
Flooded Veg	0.00000	0.00002	0.00001	0.00060	0.00026	0.00006	0.00008	0.00001	0.00105	0.57557	0.08073		39 ± 321
Crops	0.00003	0.00072	0.00129	0.00010	0.05960	0.00291	0.00078	0.00032	0.06575	0.90650	0.57555		34545 ± 765
Scrub	0.00012	0.01418	0.00471	0.00054	0.00217	0.15697	0.00939	0.00079	0.18887	0.83111	0.75478		199314 ± 928
Built Area	0.00011	0.00004	0.00000	0.00000	0.00022	0.00104	0.04025	0.00036	0.04202	0.95789	0.60135		14268 ± 615
Bare	0.00002	0.00000	0.00000	0.00000	0.00000	0.00029	0.00000	0.00050	0.00081	0.61359	0.14705		14 ± 193
Total	0.00780	0.55729	0.04559	0.00746	0.10355	0.20797	0.06693	0.00339					

Belgium - Overall Accuracy 90%

Class	Water	Trees	Grass	Flooded Veg	Crops	Scrub	Built Area	Bare	Total	User's	Producer's	Overall	Area [ha]
Water	0.00874	0.00092	0.00003	0.00015	0.00025	0.00030	0.00013	0.00001	0.01053	0.82997	0.88677	0.89616	318 ± 60
Trees	0.00017	0.20311	0.00081	0.00100	0.00406	0.01326	0.00225	0.00015	0.22480	0.90349	0.95925		145945 ± 210
Grass	0.00013	0.00111	0.01377	0.00127	0.01117	0.00417	0.00381	0.00037	0.03580	0.38473	0.55742		2712 ± 239
Flooded Veg	0.00000	0.00001	0.00000	0.00023	0.00010	0.00002	0.00003	0.00000	0.00040	0.57557	0.06605		4 ± 86
Crops	0.00023	0.00546	0.00976	0.00078	0.45253	0.02211	0.00592	0.00241	0.49920	0.90650	0.96413		718437 ± 411
Scrub	0.00001	0.00094	0.00031	0.00004	0.00014	0.01044	0.00062	0.00005	0.01256	0.83111	0.18688		2152 ± 339
Built Area	0.00056	0.00019	0.00002	0.00002	0.00111	0.00534	0.20693	0.00186	0.21602	0.95789	0.94186		145525 ± 258
Bare	0.00001	0.00000	0.00000	0.00000	0.00000	0.00024	0.00000	0.00041	0.00067	0.61359	0.07819		10 ± 128
Total	0.00986	0.21173	0.02471	0.00348	0.46936	0.05588	0.21970	0.00527					

Laos - Overall Accuracy 89%

Class	Water	Trees	Grass	Flooded Veg	Crops	Scrub	Built Area	Bare	Total	User's	Producer's	Overall	Area [ha]
Water	0.01200	0.00126	0.00004	0.00020	0.00035	0.00042	0.00018	0.00002	0.01446	0.82997	0.94038	0.88756	4245 ± 404
Trees	0.00056	0.66768	0.00266	0.00328	0.01335	0.04358	0.00741	0.00049	0.73900	0.90349	0.97715		11624007 ± 4140
Grass	0.00001	0.00009	0.00114	0.00011	0.00093	0.00035	0.00032	0.00003	0.00297	0.38473	0.12239		638 ± 1334
Flooded Veg	0.00001	0.00004	0.00002	0.00110	0.00048	0.00011	0.00015	0.00002	0.00191	0.57557	0.20806		233 ± 918
Crops	0.00002	0.00052	0.00094	0.00008	0.04338	0.00212	0.00057	0.00023	0.04785	0.90650	0.71548		66790 ± 1852
Scrub	0.00011	0.01369	0.00455	0.00052	0.00209	0.15150	0.00906	0.00076	0.18229	0.83111	0.76278		833464 ± 3957
Built Area	0.00003	0.00001	0.00000	0.00000	0.00005	0.00026	0.01026	0.00009	0.01071	0.95789	0.36714		6886 ± 1973
Bare	0.00002	0.00000	0.00000	0.00000	0.00000	0.00029	0.00000	0.00050	0.00081	0.61359	0.23376		39 ± 559
Total	0.01276	0.68329	0.00934	0.00530	0.06063	0.19862	0.02794	0.00213					

Fig. 3: Error matrices of estimated area proportions for California, Costa Rica, Laos & Belgium (top to bottom). Across all regions, we note that water, trees, crops and built area are highly accurate while grass, flooded vegetation and bare ground need further improvement.

3. RESULTS AND DISCUSSION

We achieve an overall accuracy of 85% across all ten classes on the holdout validation tiles. A more detailed accuracy assessment per class over four regions (California, Costa Rica, Belgium and Laos) appears in Figure 3, reported as an error matrix of estimated area proportions for each region along with user's and producer's accuracies. From the error matrix over each region, we also compute the area of each class with 95% confidence error bounds [13]. These four regions were chosen as initial areas for evaluation to capture a mixture of various geographies and biomes. We note that across all regions, water, trees, crops, and built area perform particularly well with user's accuracies above 80%, while other classes, notably grass, flooded vegetation, and bare ground, need further improvement. Both grass and bare ground show considerable confusion with shrub / scrub. We do not report accuracies for the cloud and snow/ice categories as these classes are de-weighted during the class weighted mode during model deployment.

Expectedly, the highest performing classes tend to be those that are most commonly occurring in the dataset, shown in Figure 2. Rare classes in the training set reflect the fact that these classes are also rare in the world at large. For lower performing classes, the confusers make intuitive sense. For example, the biggest confuser of grass is crops, which is attributable to pastures and fallow fields that have similar patterns as cropland but are not actively being farmed. Confusion between grass and scrub is intuitive since the transition between those two classes is difficult to define and difficult to distinguish in 10m satellite imagery. For grass / built area confusion, we are aggressively classifying built areas and include features like lawns and parks. This is a question of land use vs land cover - though the land cover might be grass, functionally they are built areas.

For flooded vegetation, the main confuser is crops. Globally, many crops are grown in river deltas, so the transition between these two classes is often subtle. Similarly, the main confuser for bare is scrub, and these two classes overlap and gradually transition between one another. The fact that the main confuser for bare is scrub/shrub, and scrub/shrub is also a confuser for grassland, suggests the model may over-classify scrub/shrub. We see that the distribution of classes in the validation dataset is not evenly balanced. This reflects the real world ambiguity of observations at the Sentinel 2 scale.

4. CONCLUSIONS AND NEXT STEPS

Our results show that, with a robust training dataset and a deep learning model, it is possible to create a globally consistent LULC map at 10m resolution. Our model achieves an overall accuracy of 85% across ten classes, and given that the main confusers make intuitive sense, we are confident that the global map is scientifically defensible and useful. There remain several promising avenues for future improvements. For example, including Sentinel-1 radiometrically-corrected ground range detected (GRD) data could help with all classes, specifically in teasing out flooded vegetation versus croplands and bare versus scrub/shrub. Also, adding in time-series features like measures of vegetation health over a year could differentiate grasslands vs crops vs scrub/shrub.

For lower-performing classes (e.g. grass, flooded vegetation), additional collection of hand-labeled training data to provide more examples of these classes across geographies could result in accuracy improvements. We also plan to experiment with model architectures, class weighting and additional data augmentation techniques to improve model performance and generalization. We will extend this work to Tier 2 land cover classes to better understand specific land uses in different regions (e.g. plantation vs. natural forest, residential vs. commercial built area).

5. ACKNOWLEDGEMENTS

The authors wish to acknowledge the support of Esri Inc., and the Microsoft AI for Earth/Planetary Computer program. This work makes use of the National Geographic Society Dynamic World training dataset, produced for the Dynamic World Project by National Geographic Society in partnership with Google and the World Resources Institute.

6. REFERENCES

[1] X.P. Song, M.C. Hansen, S.V. Stehman, P.V. Potapov, A. Tyukavina, E.F. Vermote, and J.R. Townshend, “Global land change from 1982 to 2016,” Nature, pp. 639-643, 2018.

[2] S.P. Abercrombie and M.A. Friedl, "Improving the consistency of multitemporal land cover maps using a hidden Markov model," IEEE Transactions on Geoscience and Remote Sensing, pp. 703-713, 2015.

[3] A. Mousivand and J.J. Arsanjani, "Insights on the historical and emerging global land cover changes: The case of ESA-CCI-LC datasets," Applied Geography pp. 82-92, 2019.

[4] M. Buchhorn, M. Lesiv, N.E. Tsendbazar, M. Herold, L. Bertels, and B. Smets, “Copernicus Global Land Cover Layers—Collection 2,” Remote Sensing, 2020.

[5] J. Wickham, S.V. Stehman, and C.G. Homer, "Spatial patterns of the United States National Land Cover Dataset (NLCD) land-cover change thematic accuracy (2001–2011)," International Journal of Remote Sensing, pp. 1729-1743, 2018.

[6] J. Brown, H. Tollerud, C. Barber, Q. Zhou, J.L. Dwyer, J. Vogelmann, T. Loveland, C. Woodcock, S.V. Stehman, Z. Zhu, and B. Pengra, “Lessons learned implementing an operational continuous US national land change monitoring capability: The LCMAP approach,” Remote Sensing of Environment, 2019.

[7] E. Dinerstein, D. Olson, A. Joshi, C. Vynne, N. Burgess, E. Wikramanayake, N. Hahn, S. Palminteri, P. Hedao, R. Noss, and others, “An Ecoregion-Based Approach to Protecting Half the Terrestrial Realm,” BioScience, 2017.

[8] C. F. Brown, S. P. Brumby, B. Guzder-Williams, T. Birch, S. B. Hyde, J. Mazzariello, W. Czerwinski, V. J. Pasquarella, R. HaerteI, S. llyushchenko, K. Schwehr, M. Weisse, F. Stolle, C. Hanson, O. Guinan, R. Moore, and A.M. Tait, “Dynamic World: Near real-time global 10m land use land cover mapping,” Nature Scientific Data, In review.

[9] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation," International Conference on Medical Image Computing and Computer-assisted Intervention, Springer, Cham, pp. 234-241, 2015.

[10] P. Zhang, Y. Ke, Z. Zhang, M. Wang, P. Li, and S. Zhang, “Urban land use and land cover classification using novel deep learning models based on high spatial resolution satellite imagery,” Sensors, 3717, 2018.

[11] J. McGlinchy, B. Muller, M. Joseph, and J. Diaz, “Application of UNet fully convolutional neural network to impervious surface segmentation in urban environment from high resolution satellite imagery,” IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium, IEEE, pp. 3915-3918, 2019.

[12] S.Wager, S. Wang and P. Liang, “Dropout Training as Adaptive Regularization,” Advances in Neural Information Processing Systems (NIPS) 2013.

[13] P. Olofsson, G. Foody, S. Stehman and C. Woodcock, “Making better use of accuracy data in land change studies: Estimating accuracy and area and quantifying uncertainty using stratified estimation”, Remote Sensing of Environment, Volume 129, 2013.