Accounting for training data error in machine learning applied to Earth observations – Supplemental Information

Arthur Elmes*[1,2], Hamed Alemohammad[3], Ryan Avery[4], Kelly Caylor[4,5], J. Ronald Eastman[1], Lewis Fishgold[6], Mark A. Friedl[7], Meha Jain[8], Divyani Kohli[9], Juan Carlos Laso Bayas[10], Dalton Lunga[11], Jessica L. McCarty[12], Robert Gilmore Pontius Jr[1], Andrew B. Reinmann[13, 14], John Rogan[1], Lei Song[1], Hristiana Stoynova[13, 14], Su Ye[1], Zhuang-Fang Yi[15] and Lyndon Estes[1]

Supplementary Materials

Map Accuracy Reporting Practices

To understand how TD errors can impact map accuracy, it is necessary to first review current practices and standards for measuring and reporting final map accuracy, which are well established in the EO literature[36,37,42,54,75]. While the emphasis of this paper is specifically on TD, as opposed to map reference data, it is necessary to review procedures for accuracy assessment. Sampling protocols for accuracy assessment are more stringent than those for the collection of TD[54], but because both training and map reference data are often collected as part of a single campaign or using the same methods[e.g. 52], the stricter set of procedures should be followed for both. We therefore summarize several important features and best practices for error analysis.

Error analysis compares a mapped variable to a corresponding map reference variable. Map reference data used for accuracy assessment are collected according to sampling and response designs that specify, respectively, the probabilities of inclusion for each location, and the protocol for creating the labeled map reference data[54,224]. Map reference data and TD may both be collected as part of a single larger sample[1], provided there is strict separation between the two datasets. Sampling design, whether simple random, stratified random, or systematic is dependent on application and *a priori* knowledge of the study area, and should be probability-based, such that the inclusion probability of each sample relates to the likelihood of that sample unit being included[36,54,225]. If the observations do not have equal probability of selection, then it is essential to convert the sample data to a confusion matrix (i.e. a square contingency table) that reflects an unbiased estimate for the entire population using methods summarized in Stehman and Foody[54].

Map accuracy is typically assessed using a metric or metrics designed to provide information regarding the correspondence of mapped and reference data. The objective of these metrics is to provide insights into the product's expected best use cases and potential shortcomings. Accuracy metrics vary according to whether the mapped variable is categorical or continuous, with each type of variable having its own foundation for error analysis[79–83]. The confusion matrix is the foundation for categorical variables. Conventionally, the table's rows provide mapped categories and the columns show the matching reference categories, with the diagonal entries showing agreement between the two. The confusion matrix is used to calculate user's accuracy (i.e. the complement of commission error), producer's accuracy (i.e. the complement of omission error), and overall accuracy (i.e. the complement of proportion error)[38]. More details on the interpretation of these values and other aspects of the error matrix are provided in several existing publications[34,36,54,79,226–228].

Several other accuracy measures are also calculated from the error matrix. Most prominent among these is the Kappa Index of Agreement[229], which is widely used in the remote sensing and species distribution modelling literature. However, Kappa varies with class prevalence[84] and can

[1] It is often advantageous to have a separate train sample design, however, as these may be more purposive and targeted to classes of interest[54].

be easily misinterpreted, thus its use is no longer recommended[37]. More recently, a number of additional metrics have started to be more commonly used in EO accuracy analysis, in part due to contributions from other disciplines, such as computer science. Due to differing conventions and objectives within these disciplines, the metrics and terminology relating to error and accuracy are often quite different. To help resolve this confusion, we summarize these metrics and their meanings in Table S1.

A special and increasingly used type of categorical map is derived from Object-Based Image Analysis (OBIA), in which the output map is classified into polygons representing discrete objects[90]. At present there is no commonly accepted standard for reporting the accuracy of such maps in the remote sensing literature[62], since the optimal set of metrics for polygon accuracy assessment depends on the intended use of the categorical map. For example, edge similarity metrics are useful for assessing the segmentation of individual agricultural fields, whereas area based metrics will fail where multiple objects are frequently mapped as a single object[62]. The Jaccard Index, also called Intersection over Union, is a commonly used benchmark in the computer vision and segmentation literature for evaluating polygon-to-polygon classification accuracies, and has the advantage of being straightforward to calculate and interpret[e.g 230,231–233]. This and other similar area-based metrics can be used in a remote sensing context, and thus may help to strengthen communication between EO and computer vision researchers. However, we caution that for many mapping goals, these metrics should be complemented by others that account for shape and edge similarity. Perhaps due to these complexities, many existing studies have assessed the accuracy of object-based maps using per-pixel accuracy assessments, which itself is problematic because it involves comparing fundamentally different spatial units[62].

The scatter plot, showing the mapped variable on the y-axis and the reference variable on the x-axis, is the foundation of error analysis for continuous variables. Since any point falling off the 1:1 line indicates deviation from a measurement of the true value, a visual assessment of the plot is an intuitive first step for assessing error in the mapped variable. Several metrics are commonly used to quantify disagreement between mapped and reference variables, including mean deviation, Root Mean Square Error (RMSE; a.k.a. Root Mean Square Deviation, RMSD), and Mean Absolute Deviation (MAD). The use of RMSE may be inappropriate, since it combines MAD with the variation among the deviations, and is frequently misinterpreted as the measurement of average error[85–87]. The Receiver Operating Characteristic (ROC) and the Total Operating Characteristic (TOC) enable analysis of a continuous mapped variable relative to a binary reference variable, for example presence or absence[81,88,89]. The area under this curve (AUC) of an ROC/TOC plot is often used as a single measure of overall accuracy that summarizes numerous thresholds for the continuous variable[89].

Most of the metrics reported above (Table 1) provide useful information for users about map reliability. However, the usefulness of that information depends on the map reference data having higher accuracy than the mapped data, which is an assumption that is often unexamined[31,178]. This tendency is illustrated by Ye et al.[62], who reviewed 209 journal articles focused on object-based image analysis and found that one third gave incomplete information about the sample design and size of their map reference data, let alone any mention of error within the sample. Errors in map reference data can bias the map accuracy assessment[44,149], as well as estimates derived from the confusion matrix, such as land cover class proportions and their standard errors[43]. To correct for such biases caused by map reference error, one can use published procedures for estimating map reference data accuracy[44] and to calculate variance measures for area estimates[43]. These approaches depend on quantifying errors in the map reference data. For the common case of image-interpreted map reference data, this can be achieved by having multiple interpreters create reference polygons and labels for the same locations, and then calculating the level of agreement in their categorical labels[31,54,149,153]. Additionally, knowledge of this uncertainty can be quantitatively incorporated into continuous estimates based on the image interpreted data[43]. *In situ* observations can similarly be used to assess the accuracy of image-interpreted map reference samples[57], although their availability is often limited by cost considerations.

Table 1. List of peer-reviewed publications retrieved using Google Scholar search algorithm results. Search performed January, 2019, with terms land cover and land use, including permutations of spelling and punctuation. Twenty-seven articles kept after initial screening for relevance.

Authors	Title	Journal
Zhong, B.; Ma, P.; Nie, A.; Yang, A.; Yao, Y.; Lü, W.; Zhang, H.; Liu, Q.	Land cover mapping using time series HJ-1/CCD data	*Sci. China Earth Sci.* 2014 *57*, 1790–1799.
Pacifici, F.; Chini, M.; Emery, W.J.	A neural network approach using multi-scale textural metrics from very high-resolution panchromatic imagery for urban land-use classification	*Remote Sens. Environ.* **2009**, *113*, 1276–1292.
Abbas, I.I.; Muazu, K.M.; Ukoje, J.A.;	Mapping Land Use - land Cover and Change Detection in Kafur Local Government , Katsina , Nigeria (1995 - 2008) Using Remote Sensing and Gis	*Research journal of environmental and Earth Sciences* **2010**, *2*, 6–12.
Sano, E.E.; Rosa, R.; Brito, J.L.S.; Ferreira, L.G.	Land cover mapping of the tropical savanna region in Brazil	*Environ. Monit. Assess.* **2010**, *166*, 113–124.
Hu, T.; Yang, J.; Li, X.; Gong, P.	Mapping urban land use by using landsat images and open social data	*Remote Sensing* **2016**, *8*, 151.
Galletti, C.S.; Myint, S.W.	Land-use mapping in a mixed urban-agricultural arid landscape using object-based image analysis: A case study from Maricopa, Arizona	*Remote Sensing* **2014**, *6*, 6089–6110.
Hu, Q.; Wu, W.; Xia, T.; Yu, Q.; Yang, P.; Li, Z.; Song, Q.	Exploring the use of google earth imagery and object-based methods in land use/cover mapping	*Remote Sensing* **2013**, *5*, 6026–6042.
Al-Bakri, J.T.; Ajlouni, M.; Abu-Zanat, M.	Incorporating Land Use Mapping and Participation in Jordan	*Mt. Res. Dev.* **2008**, *28*, 49–57.
Mallinis, G.; Emmanoloudis, D.; Giannakopoulos, V.; Maris, F.; Koutsias, N.	Mapping and interpreting historical land cover/land use changes in a Natura 2000 site using earth observational data: The case of Nestos delta, Greece	*Appl. Geogr.* **2011**, *31*, 312–320.
Liu, J.; Kuang, W.; Zhang, Z.; Xu, X.; Qin, Y.; Ning, J.; Zhou, W.; Zhang, S.; Li, R.; Yan, C.; et al.	Spatiotemporal characteristics, patterns and causes of land use changes in China since the late 1980s	*J. Geogr. Sci.* **2014**, *24*, 195–210.
Yadav, P.K.; Kapoor, M.; Sarma, K.	Land Use Land Cover Mapping, Change Detection and Conflict Analysis of Nagzira-Navegaon Corridor, Central India Using Geospatial Technology	*International Journal of Remote Sensing and GIS* **2012**, *1*.
da C. Freitas, C.; d. S. Soler, L.; Sant'Anna, S.J.S.; Dutra, L.V.; dos Santos, J.R.; Mura, J.C.; Correia, A.H.	Land Use and Land Cover Mapping in the Brazilian Amazon Using Polarimetric Airborne P-Band SAR Data."	*IEEE Trans. Geosci. Remote Sens.* **2008**, *46*, 2956–2970.
Dewan, A.M.; Yamaguchi, Y.	Land use and land cover change in Greater Dhaka, Bangladesh: Using remote sensing to promote sustainable urbanization	*Appl. Geogr.* **2009**, *29*, 390–401.
Castañeda, C.; Ducrot, D.	Land cover mapping of wetland areas in an agricultural landscape using SAR and Landsat imagery	*J. Environ. Manage.* **2009**, *90*, 2270–2277.
Griffiths, P.; van der Linden, S.; Kuemmerle, T.; Hostert, P.	A Pixel-Based Landsat Compositing Algorithm for Large Area Land Cover Mapping	*IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **2013**, *6*, 2088–2101.
Ge, Y.	Sub-pixel land-cover mapping with improved fraction images upon multiple-point simulation	*Int. J. Appl. Earth Obs. Geoinf.* **2013**, *22*, 115–126.

Gong, P.; Wang, J.; Yu, L.; Zhao, Y.; Zhao, Y.; Liang, L.; Niu, Z.; Huang, X.; Fu, H.; Liu, S.; 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*, 2607–2654.
Ghorbani, A.; Pakravan, M.	Land Use Mapping Using Visual vs. Digital Image Interpretation of TM and Google Earth Derived Imagery in Shrivan-Darasi Watershed (Northwest of Iran)	*Int. J. Remote Sens.* **2013**, *34*, 2607–2654.
Friedl, M.A.; Sulla-Menashe, D.; Tan, B.; Schneider, A.; Ramankutty, N.; Sibley, A.; Huang, X.	MODIS Collection 5 global land cover: Algorithm refinements and characterization of new datasets	*Remote Sens. Environ.* **2010**, *114*, 168–182.
Deng, J.S.; Wang, K.; Hong, Y.; Qi, J.G.	Spatio-temporal dynamics and evolution of land use change and landscape pattern in response to rapid urbanization	*Landsc. Urban Plan.* **2009**, *92*, 187–198.
Otukei, J.R.; Blaschke, T.	Land cover change assessment using decision trees, support vector machines and maximum likelihood classification algorithms	*Int. J. Appl. Earth Obs. Geoinf.* **2010**, *12*, S27–S31.
Malinverni, E.S.; Tassetti, A.N.; Mancini, A.; Zingaretti, P.; Frontoni, E.; Bernardini, A.	Hybrid object-based approach for land use/land cover mapping using high spatial resolution imagery	*Int. J. Geogr. Inf. Sci.* **2011**, *25*, 1025–1043.
Rozenstein, O.; Karnieli, A.	Comparison of methods for land-use classification incorporating remote sensing and GIS inputs	*Appl. Geogr.* **2011**, *31*, 533–544.
Jawak, S.D.; Luis, A.J.	Improved land cover mapping using high resolution multiangle 8-band WorldView-2 satellite remote sensing data	*JARS* **2013**, *7*, 073573.
Ran, Y.H.; Li, X.; Lu, L.; Li, Z.Y.	Large-scale land cover mapping with the integration of multi- source information based on the Dempster – Shafer theory	*Int. J. Geogr. Inf. Sci.* **2012**, *26*, 169–191.
Clark, M.L.; Aide, T.M.; Grau, H.R.; Riner, G.	A scalable approach to mapping annual land cover at 250 m using MODIS time series data: A case study in the Dry Chaco ecoregion of South America	*Remote Sens. Environ.* **2010**, *114*, 2816–2832.
Berberoglu, S.; Akin, A.	Assessing different remote sensing techniques to detect land use/cover changes in the eastern Mediterranean	*Int. J. Appl. Earth Obs. Geoinf.* **2009**, *11*, 46–53.

Table 2. Summary of commonly used error metrics.

Term	Information Content/Typical Usage	Description
Overall Accuracy	Summary metric combining all class accuracies into a single number	Proportion of correctly classified cases divided by the total of all classified cases
User's Accuracy (a.k.a. Precision)	Metric of the intensity of true positives given the classified category in which the true positives were 'found'. The intensity complement of commission error.	Proportion of correctly classified cases relative to the total number of cases classified into the given category
Kappa Index of Agreement	Single metric for overall accuracy	Used to measure the agreement between mapped and reference categories of a dataset while attempting to correct for agreement that occurs by chance.

Producer's Accuracy (a.k.a. Sensitivity, Recall)	Metric indicating the intensity of true positives given the reference category) The intensity complement of omission error.	True positive rate; ratio of correctly classified cases of a given class to the total true cases of that class
Specificity	Metric for commission error; indicates how well the model avoids false positives	True negative rate; ratio of correctly classified negatives to the sum of true negatives and false positives
True Skill Statistic [78]	Metric that combines sensitivity and specificity while accounting for class prevalence	Sensitivity + Specificity - 1
F1 [79,80]	Combined metric of commission and omission error	Equally weighted harmonic mean of precision and recall
Bias (Mean Bias Error)	Quantifies the average difference between predicted and reference variables	The average error, representing the systematic over- or under-prediction of a continuous variable
Root Mean Square Error/Deviation	Measures a combination of the average error and the variability within the distribution of errors	A potentially misleading metric used to measure disagreement between predicted and reference continuous variables
Mean Absolute Deviation	Measures how far points are from Y=X line	Recommended metric to measure disagreement between predicted and reference continuous variables
Jaccard Index, also called Intersection over Union	Between two discrete/crisp datasets, reports the area of intersection divided by the area of union.	Most commonly used metric to indicate accuracy of object-based classification, which is also called semantic segmentation.

Triple Collocation RMSE

TC-based RMSE estimates at each pixel were used to compute *a priori* probability (P_i) of selecting a particular dataset:

$$P_i = \frac{\frac{1}{\sigma_{\varepsilon_i}^2}}{\sum_{i=1}^{3} \frac{1}{\sigma_{\varepsilon_i}^2}} \quad \text{(eq. S1)}$$

P_i is the probability of selecting measurement system i, $\sigma_{\varepsilon i}$ is the standard deviation of the random error in measurement system i.

Figure S1 depicts how X_T (the training time series for a pixel) is formed by sampling from X_1, X_2, and X_3 over time.

Table 3. Quantitative results of comparing each of the three models trained for the road detection case in Kumasi, Ghana to the validation labels. This region (shown in Figure 9) included 5,406,942 road pixels and 50,627,010 background pixels.

	F1	**IOU**	**Precision**	**Recall**
Khartoum Model				
Average	0.6659	0.5723	0.7758	0.6267
Road	0.3780	0.2330	0.6250	0.2709
Background	0.9538	0.9116	0.9266	0.9862
Kumasi Model				
Average	0.8004	0.6955	0.7693	0.8450
Road	0.6458	0.4769	0.5662	0.7513
Background	0.9552	0.9142	0.9725	0.9386
Khartoum Model retrained in Kumasi				
Average	0.7869	0.6830	0.7965	0.7780
Road	0.6135	0.4425	0.6363	0.5921
Background	0.9603	0.9236	0.9568	0.9639

Table 4. Template and procedure for documenting training data. Note that the 'values' column is intentionally left blank, as this is merely an example. We would expect a fully filled out table to be several pages in length due to the technical nature of the metadata explanation.

Metadata Category	**Value**
Training data set name	
How data were created (technical details, to include number of analysts, whether *in situ* or image interpretation, samples of field sheets, copies of materials used to educate analysts, date of data creation, etc.)	
Funding source	
Purpose	
LULC definitions	
Time period	
Spatial extent	
Spatial resolution (image, field, quadrat, point location)	
Image ID (sensor specific unique identification information)	

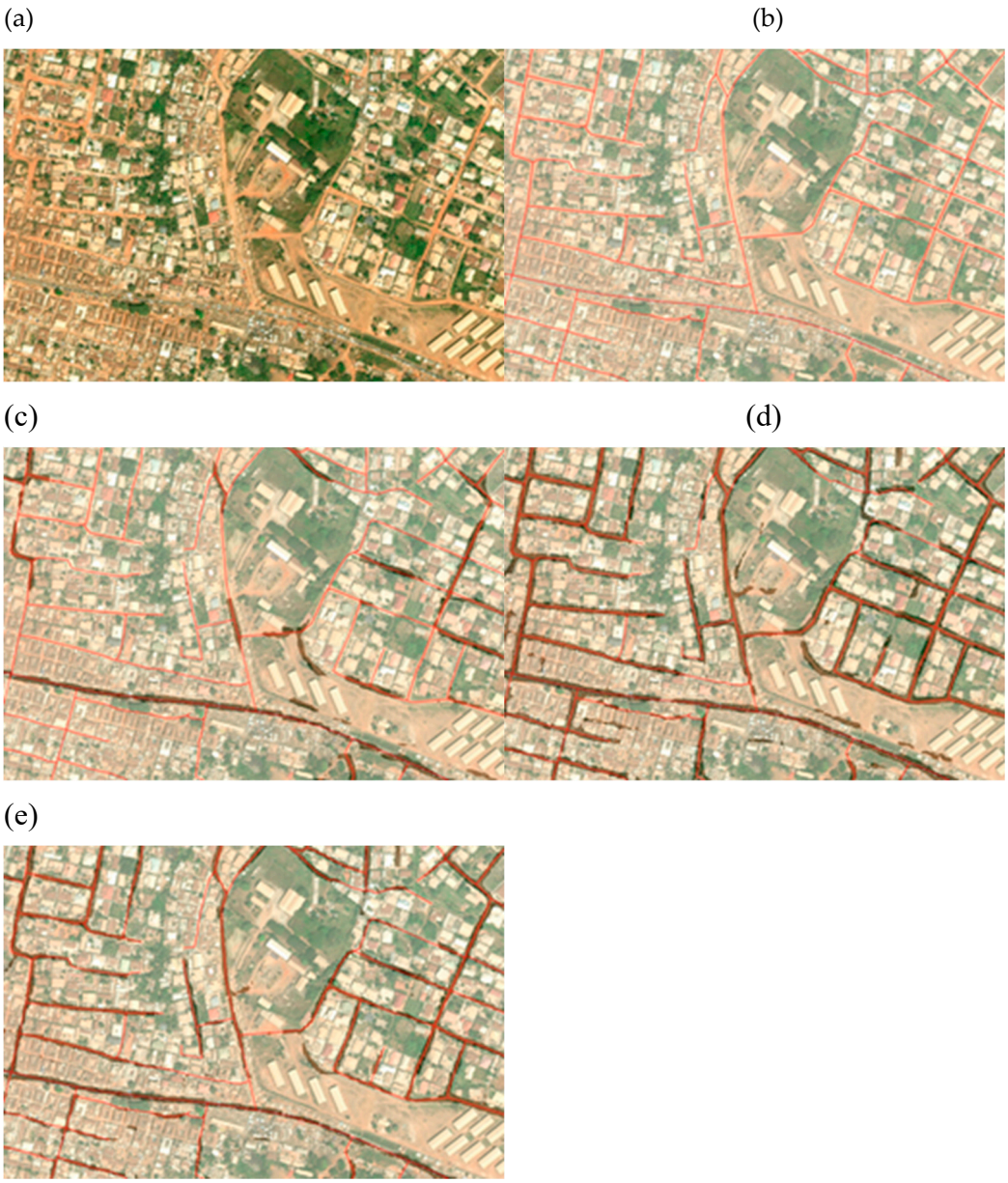

Figure 1. Sample prediction results in Kumasi, Ghana. (a) Input imagery. (b) Predictions from the Las Vegas model. (c) Predictions from the Khartoum model. (d) Prediction from the Kumasi model. (e) Predictions from the Khartoum Model retrained in Kumasi. [In panel b-e model predictions are in shaded color overlaid with validation labels in red on top of imagery.].

Figure S1 shows a qualitative comparison of different model outputs along with the validation labels over a sample area of Figure 4.

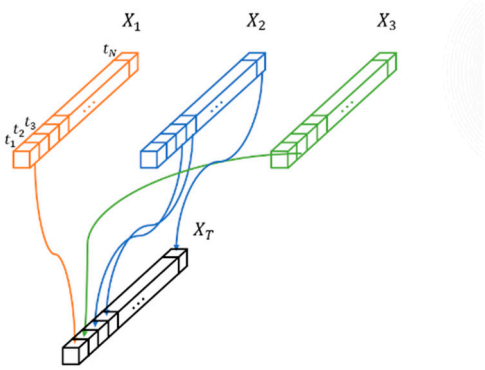

Figure 2. Schematic of product selection using the Triple Collocation approach.