Overselling overall map accuracy misinforms about research reliability

Guofan Shao · Lina Tang · Jiangfu Liao

Received: 11 August 2019 / Accepted: 25 September 2019 / Published online: 4 October 2019
© The Author(s) 2019

Abstract

Context Image classification is routine in a variety of disciplines, and analysts rely on accuracy metrics to evaluate the resulting maps. The most frequently used accuracy metric in Earth resource remote sensing is overall accuracy. However, the inherent properties of this accuracy metric make it inappropriate as the single metric for map assessment, particularly when a map contains imbalanced categories.

Objectives We discuss four noteworthy problems with overall accuracy. Under circumstances frequently encountered, overall accuracy is misleading or misinterpreted.

Methods Literature review, hypothetical examples, and mathematic equations are used to prove overall accuracy is a poor general indicator of map quality.

Conclusions Any research that involves classification techniques or a map product that is evaluated only

with overall accuracy may be unreliable. It is necessary for map providers to publish the error matrix and its development procedure so that map users can computer whatever metrics as they wish.

Keywords Image processing · Error propagation · Imbalanced classes · Map accuracy · Comprehensive assessment

Digital classification is commonly used to convert image data into map products by assigning each pixel into one of two or more categories. This approach is routine in many different fields, ranging from medical imaging to satellite remote sensing (Celeb et al. 2019; Phiri and Morgenroth 2017; Xiao et al. 2019). The resulting maps are frequently used for decision making and scientific research, so it is vital to verify map quality. One standard way to evaluate maps is to use error matrices (also known as confusion matrices) to compute accuracy metrics or indices, including map-level and category-level accuracies. In the geospatial field, the map-level accuracy metric is called the overall accuracy (*O*) and the category-level accuracy metrics are called the producer's (*P*) and user's (*U*) accuracies (Table 1) (Story and Congalton 1986; Congalton 1991; Foody 2002). The overall accuracy is the percentage of cases (e.g., pixels or sampling units) correctly classified. If only a single accuracy metric is reported for a classification map, it

G. Shao (✉)
Department of Forestry and Natural Resources, Purdue
University, West Lafayette, IN 47907, USA
e-mail: shao@purdue.edu

L. Tang (✉)
Key Laboratory of Urban Environment and Health,
Institute of Urban Environment, Chinese Academy of
Sciences, Xiamen 361021, China
e-mail: lntang@iue.ac.cn

J. Liao
College of Computer Engineering, Jimei University, 185
Yinjiang Road, Xiamen 361021, China

Table 1 A population error matrix and selected accuracy metrics

Map data	Reference data			Map total	Users' accuracy (U)
	$j = 1$	$j = 2$	$j = J$		
$i = 1$	p_{11}	p_{12}	p_{1J}	p_{1+}	p_{11}/p_{1+}
$i = 2$	p_{21}	p_{22}	p_{2J}	p_{2+}	p_{22}/p_{2+}
...					
$i = J$	p_{J1}	p_{J2}	p_{JJ}	p_{J+}	p_{JJ}/p_{J+}
Reference total	p_{+1}	p_{+2}	p_{+J}	1	
Producer's accuracy (P)	p_{11}/p_{+1}	p_{22}/p_{+2}	p_{JJ}/p_{+J}		

p_{ij} represents the proportion of correctly classified sample points for category j in the diagonal of a confusion matrix; p_{ij} is the proportion of sample points at row i and column j , and J is the total number of land-cover types or categories; p_{i+} is the total proportion of category i as derived from the map data; and p_{+j} is the total proportion of category j as obtained from the reference data. Overall accuracy $O = \sum_{j=1}^J p_{jj}$

is likely to be the overall accuracy, as shown in an overwhelming preponderance of published research in the field of earth-surface detection and mapping (Fig. 1). We refer to this phenomenon as “overselling overall accuracy,” meaning that map handlers overemphasize the importance of map-level accuracy.

Overselling overall accuracy should be discouraged because the O metric has inherent properties that lead to misleading perceptions about the soundness of classification methods and map-based research.

Specifically, the inherent properties of the O metric contribute to at least four application problems: (1) The O metric cannot explicitly explain the accuracy or uncertainties of map derivatives, (2) the O values are not comparable between maps with different numbers of categories, (3) the O values are not comparable between maps with different category proportions, and (4) a greater O value may not necessarily represent a more realistic map.

Fig. 1 Histogram comparing the number of publications reporting three types of accuracy based on a topic search from the Web of Science Core Collection. The data cover the top ten journal groups, including remote sensing, imaging science photographic technology, environmental sciences, geoscience multidisciplinary, geography physical, ecology, forestry, water

resources, agriculture interdisciplinary, and geography during 1978–2018. The topic keywords include remote sensing or remotely sensed, classification or map* in addition to overall accuracy, producer's accuracy, and user's accuracy in titles, keywords, and abstracts

The first O -metric problem arises because the O metric is computed using only diagonal elements in an error matrix and thus does not reflect error distributions among categories. A high O value for a map does not assure the reliability of map derivatives. These derivatives include category area, the most common map derivative for remote-sensing applications (Olofsson et al. 2014). For example, the rate of 90% overall accuracy cannot ensure that the area estimate of a category is accurate to within $\pm 10\%$. Consider a case where a 100 ha forest has a carbon density of 100 tons/ha at time 1 and 110 tons/ha at time 2. The forest will sequester 1000 tons of carbon in total between the two times. Given $\pm 10\%$ uncertainties for the estimates of area and carbon density, the maximum possible range of carbon sequestration between time 1 and time 2 is from -3190 to 5210 tons. This example illustrates that relatively high O values for a map do not guarantee uncertainties small enough for map derivatives to avoid big uncertainties in map applications. By contrast, the producer's and user's accuracies are more informative than O for area estimates. Despite a reasonable value for O , errors or uncertainties can be surprisingly high in landscape quantification (Arbia et al. 1998; Shao and Wu 2008). That is because the O metric does not reflect error's spatial distribution.

The second problem with the O metric appears because an increase in the number of mapping categories leads to an increase in the number of spectrally similar pixels and in the number of edge pixels. When a pixel probability of assignment is nearly equal for each of two or more categories, the chances increase for the pixel to be wrongly classified. Because edge pixels are located at the border of adjacent components, they tend to contain a mixture of two categories, which complicates image classification (Sweeney and Evans 2012; Heydari and Mountrakis 2018). Thus, an increase in mapping categories usually lowers O . The balance between the number of categories and classification accuracy is a practical consideration in classification design and implementation. On the application side, if a single category is the focus, such as deforestation or urban sprawl, a low- O map with multiple categories may be aggregated into a high- O map with only two categories. Such an aggregation approach does not, however, increase the accuracy of the target category.

To understand the third problem with the O metric, we include a range of p_{ij} into its formula:

$$O = \sum_{j=1}^J p_{ij} \quad (0 < p_{ij} \leq p_{+j}) \quad (1)$$

A large category means that p_{+j} is large (Olofsson et al. 2014) and that p_{ij} covers a broad range. Thus, a larger category can have a greater role than a smaller category in regulating the value of O . For example, if a dominant category accounts for 90% of the total mapping area (i.e., $p_{+j} = 90\%$) and its p_{ij} value is also 90%, O will be at least 90%. If its p_{ij} value is reduced to 70%, the O value cannot exceed 80% because the other categories have only a 10% share in total, and their p_{ij} values cannot exceed 10%. In contrast, a single small category does not exert such a dominant effect. This dominant control of O by a large category has already attracted attention in predictive analytics and machine learning (Fielding and Bell 1997; He and Garcia 2009) and remote sensing (Stehman and Foody 2019).

When a map contains two unevenly sized or imbalanced categories (e.g., category 1 is the majority category, and category 2 is the minority category), and the classification accuracies between P and U are balanced for each category (i.e., $P_1 = U_1$, $P_2 = U_2$, and $p_{12} = p_{21}$), we obtain

$$\frac{P_1}{P_2} = \frac{\frac{p_{11}}{p_{+1}}}{\frac{p_{22}}{p_{+2}}} = \frac{1 - \frac{p_{21}}{p_{+1}}}{1 - \frac{p_{21}}{p_{+2}}} \quad (2)$$

Because $p_{+1} > p_{+2}$, the accuracy of a large category exceeds that of a small category. In other words, the relative error for the minority category always exceeds that of the majority category. Equation (2) represents a simple case of image-data classification and map assessment. Although multiple unevenly sized categories are more complicated than Eq. (2), they show similar trends. For example, a positive correlation between category areas and accuracies was found in the 1-km-resolution global land-cover data sets (Scepan 1999). Equation (2) also implies that requiring imbalanced categories to be equally accurate may be unrealistic. Thus, O cannot represent the accuracy of every category.

Under the circumstances of Eq. (2), the minimum value of the majority category cannot be less than the difference between the majority and minority

shows that the O metric can be a poor measure of accuracy. In the field of predictive analytics, this phenomenon (i.e., when high-accuracy models do not have greater predictive power than lower-accuracy models) is called the “accuracy paradox” (Thomas 2013; Kim et al. 2017). It is worth it to mention that the value of kappa coefficient for Fig. 2d is 0 and does not cause accuracy paradox in this case. Although kappa coefficient is more conservative than overall accuracy, it has its own problems and its overuse should also be avoided.

Note that these problems exist even though the O metric was computed correctly from an error matrix that represents an estimated population. The inappropriate use of reference sampling data may worsen the problem. For example, O can be unrealistically high if only pure pixels are used to quantify accuracy. Biased pixel sampling can amplify the problems inherent to O . Overselling overall accuracy has misleading effects on research reliability because overall accuracy alone cannot explain the robustness of a classification model on the method-research side or the quality of classification output on the application-research side.

Because artificial intelligence is becoming increasingly popular for image recognition, classification, and mapping in various disciplines (e.g., Grekousis 2019), consistent standards of accuracy quantification and analysis are needed to assure users that the technique is adequately dependable. Although overall accuracy has problems, it cannot be ignored. Instead, overall accuracy should be used together with multiple accuracy metrics to promote a comprehensive approach to quantify map accuracy (Lasko et al. 2005; Liu et al. 2007). Many researchers urge map producers to provide complete information on map development and assessment (Congalton et al. 2014; Congalton and Green 2019; Stehman and Foody 2019). This type of transparency is vital because it would allow map stakeholders to compute any accuracy metrics rather than relying solely on those given by the map providers.

Acknowledgements This research was supported by the USDA National Institute of Food and Agriculture McIntire Stennis Project (IND011523MS), the National Natural Science Foundation of China (41471137), and the National Key R&D Program of China (2016YFC0502902), and the Natural Science Foundation of Fujian Province (2017J01468). The authors thank Drs. Keith E. Woeste and Russell G. Congalton, and anonymous

reviewers for their constructive comments and suggestions that have helped improve the manuscript.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

- Arbia G, Griffith D, Haining R (1998) Error propagation modelling in raster GIS: overlay operations. *Int J Geogr Inf Sci* 12:145–167
- Celeb ME, Codella N, Halpern A (2019) Dermoscopy image analysis: overview and future directions. *IEEE J Biomed Health Inf* 23:474–478
- Congalton RG (1991) A review of assessing the accuracy of classifications of remotely sensed data. *Remote Sens Environ* 37:35–46
- Congalton RG, Green G (2019) *Assessing the Accuracy of Remotely Sensed Data: Principles and Practices*, 3rd edn. CRC Press, Boca Raton
- Congalton RG, Gu J, Yadav K, Thenkabail P, Ozdogan M (2014) Global land cover mapping: a review and uncertainty analysis. *Remote Sensing* 6:12070–12093
- Fielding AH, Bell JF (1997) A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environ Conserv* 24:38–49
- Foody GM (2002) Status of land cover classification accuracy assessment. *Remote Sens Environ* 80:185–201
- Grekousis G (2019) Artificial neural networks and deep learning in urban geography: a systematic review and meta-analysis. *Comput Environ Urban Syst* 74:244–256
- Hammond TO, Verbyla DL (1996) Optimistic bias in classification accuracy assessment. *Int J Remote Sens* 17:1261–1266
- He H, Garcia EA (2009) Learning from Imbalanced Data. *IEEE Trans Knowl Data Eng* 21:1263–1284
- Heydari SS, Mountrakis G (2018) Effect of classifier selection, reference sample size, reference class distribution and scene heterogeneity in per-pixel classification accuracy using 26 Landsat sites. *Remote Sens Environ* 204:648–658
- Kim JK, Han YS, Lee JS (2017) Particle swarm optimization-deep belief network-based rare class prediction model for highly class imbalance problem. *Concurr Comput* 29:e4128
- Lasko TA, Bhagwat JG, Zou KH, Ohno-Machado L (2005) The use of receiver operating characteristic curves in biomedical informatics. *J Biomed Inform* 38:404–415
- Liu C, Frazier P, Kumar L (2007) Comparative assessment of the measures of thematic classification accuracy. *Remote Sens Environ* 107:606–616
- Olofsson P, Foody GM, Herold M, Stehman SV, Woodcock CE, Wulder MA (2014) Good practices for estimating area and

- assessing accuracy of land change. *Remote Sens Environ* 148:42–57
- Phiri D, Morgenroth J (2017) Developments in Landsat land cover classification methods: a review. *Remote Sens* 9:967
- Scepan J (1999) Thematic validation of high-resolution global land-cover data sets. *Photogramm Eng Remote Sens* 65:1051–1060
- Shao GF, Wu JG (2008) On the accuracy of landscape pattern analysis using remote sensing data. *Landscape Ecol* 23:505–511
- Stehman SV, Foody GM (2019) Key issues in rigorous accuracy assessment of land cover products. *Remote Sens Environ* 231:111199
- Story M, Congalton R (1986) Accuracy assessment: a user's perspective. *Photogramm Eng Remote Sens* 52:397–399
- Sweeney SP, Evans TP (2012) An edge-oriented approach to thematic map error assessment. *Geocarto Int* 27:31–56
- Thomas C (2013) Improving intrusion detection for imbalanced network traffic. *Secur Commun Netw* 6:309–324
- Xiao FY, Gao GY, Shen Q, Wang XF, Ma Y, Lu YH, Fu BJ (2019) Spatio-temporal characteristics and driving forces of landscape structure changes in the middle reach of the Heihe River Basin from 1990 to 2015. *Landscape Ecol* 34:755–770
- Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.