ASSESSING THE ACCURACY OF SATELLITE IMAGE CLASSIFICATIONS FOR POLLUTANT LOADINGS ESTIMATION

Lourdes V. Abellera * and Michael K. Stenstrom

Department of Civil and Environmental Engineering, University of California, Los Angeles, California 90095-1593, USA
labeller@ucla.edu

KEY WORDS: Pollutant loadings, Overall accuracy, Weighted overall accuracy, Kappa coefficient, Weighted kappa coefficient, Error matrix, Landsat ETM+

ABSTRACT:

We examined the accuracy measures to evaluate maps created from knowledge-based classifications of remotely sensed data. The automated classifications involved categories that showed different levels of annual loadings of six pollutants. From the classification error matrices that used spectral information and ancillary data, we computed the overall accuracy and the kappa coefficients. These common measures, however, assume that misclassification errors are equally serious. We propose a procedure, directly related to the pollutant loadings, to calculate weights for the cells in the error matrix to reflect the severity of the misclassification errors. With the weights we were able to calculate the weighted overall accuracy and the weighted kappa coefficient. By using the weighted equivalents of the usual measures of accuracy, we find that there is more specificity in the measures of quality of the classifications for the individual pollutants.

1. INTRODUCTION

Calculation of pollutant loadings is necessary to be able to identify areas to be prioritized for the implementation of stormwater best management practices. Since the generation of contaminants is closely related to land use, pollutant loadings are usually computed from land use maps. However, acquiring data to create a land use map is a slow and difficult process. In addition, land use data from public records is not optimized for environmental purposes. Usually, the land use categories are too specific, not relevant, or poorly defined. Therefore, instead of classifying land use from satellite imagery, we identified different levels of pollution directly from the image. We considered six contaminants: total suspended solids (TSS), biochemical oxygen demand (BOD5), total phosphorus (Total P), total Kjeldahl nitrogen (TKN), copper (Cu), and oil and grease (O & G). The study area is Marina del Rey and its vicinity, a 24.7 sq km highly urbanized portion of the Santa Monica Bay watershed in Los Angeles, California. The image used was a 28.5 m resolution Landsat ETM+ acquired on August 11, 2002. To evaluate the quality of the classifications, error matrices were assembled, and overall accuracy and kappa coefficients were computed. However, these measures assume that all misclassification errors are equally serious. We propose a method that weighs the errors, and suggest measures that reflect the accuracy of the classifications with more specificity.

2. METHODOLOGY

Pollutant loadings for the study area were reported by Abellera and Stenstrom (2005) and are shown in Table 1. We designated the pollution levels as high, medium, and low. For example, for TKN, we found that the loadings were similar for single-family, multiple-family, commercial, public, light industrial, and other urban (0.17-0.22 kg/year). Therefore, these types of land use were aggregated to form the high TKN loading category. Since open land had a much lower emission at 0.04 kg/year, we designated this land use as low loading. We did a similar analysis for the rest of the water quality parameters.

Pollutant	Land Use						
	SF	MF	C	P	LI	OU	O
TSS	14.70	15.83	17.31	15.44	17.31	18.01	6.37
BOD5	0.86	1.13	1.35	1.20	1.35	1.29	0.02
Total P	43.08	46.73	41.35	36.88	41.35	53.18	6.76
TKN	0.22	0.18	0.19	0.17	0.19	0.20	0.04
Cu	4.82	7.54	6.92	6.18	6.92	8.58	0.71
O & G	0.15	1.66	2.12	1.89	2.12	1.89	0

SF = Single-Family, MF = Multiple-Family, C = Commercial, P = Public, LI = Light Industrial, OU = Other Urban, O = Open
Loadings are in kg/year, except for Total P and Cu which are in g/year.

Table 1. Annual pollutant loadings

With ERDAS Imagine 8.7, we segregated the imagery to areas that had high, medium, and low loading for each pollutant using

* Corresponding author

knowledge-based classification coupled with standard GIS operations. We applied the ISODATA (Iterative Self-Organizing Data Analysis Technique) procedure (Richards, 1986) on a tasselled cap transformation (Crist and Cicone, 1984) using the greenness, wetness, and haze components computed from the six raw bands of blue, green, red, near infrared, and the two mid-infrared bands. This resulted in the separation of open land from non-open land. Single-family residential was likewise distinguished using the six raw bands. Only the near infrared band was utilized to segregate water. Using only spectral data (first classification), the area where the beach met the ocean showed open land misclassified to non-open land. In the second classification, a buffer distance of five pixels (142.5 m) corrected this error. Neighbourhood analysis was employed in the third classification. To keep its value in the second classification, a pixel should have at least three of its neighbours (in the north, east, west, and south directions) in the same category. Otherwise, it was replaced by the value in the majority image that was processed from a 3 x 3 filter.

The next task was to quantify the quality of the classifications. This was done first by assembling error matrices. We tested 1,040 randomly generated pixels which were 3.4% of the study area. These points were mainly checked from aerial photos and field visit. The land use digital map published by the Southern California Association of Governments (SCAG) was not used fully because there was no one-to-one correspondence between its categories and the classes we have designated. For example, "other urban" areas in the SCAG data have both open land and built-up areas. This illustrates that land use data from public records are often incompatible with environmental objectives.

Overall accuracy is the sum of the correctly classified pixels divided by the total number of test pixels. The kappa coefficient factors in the effect of chance in the classification (Lillesand and Kiefer, 1994). For example, a kappa value of 78% indicates that the classification is 78% better than a classification that resulted from random assignment. Therefore, kappa is lower than the overall accuracy. Both measures were computed using ERDAS Imagine 8.7. To calculate for their weighted equivalents, we need to assign a weight for each cell in the error matrix to reflect the severity of the misclassification error. Let w_{ij} be the weight associated with the i,jth cell in the error matrix. Fleiss *et al.* (1969) state that weights are limited to the interval $0 = w_{ij} = 1$ for $i \neq j$, and that the weight for perfect agreement is 1 (i.e., $w_{ii} = 1$). Naesset (1996) suggested that weights may reflect the loss of utility because of misclassification. If $U_{c,j}$ is the utility when a pixel is correctly classified into dass j and $U_{E,ij}$ is the utility when a pixel belonging to class j is wrongly classified into class i, then the weight is

$$w_{ij} = U_{E,ij}/U_{c,j}$$

The benefit of creating a pollutant loading map is to be able to identify areas generating high levels of pollution. Therefore, initially, we thought of quantifying the value of best management practices (BMPs). However, BMPs can vary depending on the type of pollutants. In addition, it may be difficult to determine the dollar value of the benefit arising from a BMP. Next, we looked at the pollutant loadings. We computed for the average values associated with high, medium, and low loadings. If we put the absolute values of pollutant loadings in the equation above, we may have a value of zero in the denominator. Amounts over- or underestimated from the misclassification errors can also produce zeroes in the denominator.

Cicchetti and Allison (1971) proposed a way of assigning weights specifically for ordinal data. Perfect agreement is assigned a weight of 1, and the worst disagreement is assigned a weight of zero. Weights of other cases of misclassifications are determined linearly. Because our classification is in ordinal scale, this procedure is applicable. However, we have modified it so that the weights were linearly related to the amount of pollutant loadings. The difference between TSS low loading and TSS medium loading, for example, is not the same as the difference between BOD5 low loading and BOD5 medium loading.

To demonstrate how weights are calculated, we take copper as an example. Similar procedures were applied to the other pollutants. Complete agreement is assigned a weight of 1, and the worst disagreement is assigned a weight of zero. (Table 2) When we misclassify an actual water pixel to high loading, we are "putting" high amounts of loading to an area where there is none generated. Misclassifying medium to high loading or vice-versa has a less serious effect because the over- or underestimated amount is smaller than in the worst misclassification case. In Table 2, we need to compute for the weights *a, b, c, d, e*. Note that the severity of error associated with misclassifying low to high loading, for example, is as severe as misclassifying high to low loading, hence they have the same weight.

	Water	Cu, Low	Cu, Medium	Cu, High
Water	1	*a*	*b*	0
Cu, Low	*a*	1	*e*	*c*
Cu, Medium	*b*	*e*	1	*d*
Cu, High	0	*c*	*d*	1

Table 2. Agreement weight matrix for copper with variable weights to be computed

In Table 3, the weights are related to the pollutant loadings. In the second and third columns, when the loading is zero, the weight is one, and when the loading is 7.23, the weight is zero. For the last column, when the loading is 0, the weight is also zero. When the loading is 7.23, the weight is one. By simple ratio and proportion, we computed for the values of *a, b, c,* and *d*. Using these values, the weight *e* was calculated by averaging the weights of its neighbours in the north, east, west, and south directions. Table 4 shows the completed agreement weight matrix for copper. Similar tables were made for TSS, BOD5, Total P, TKN, and O & G. After the weight matrices were

computed, weighted overall accuracy and weighted kappa coefficients (Cohen, 1968) were calculated using STATA 8.2.

	Loading	Weights	Weights
Water	0	1	0
Cu, Low	0.71	*a*	*c*
Cu, Medium	4.82	*b*	*d*
Cu, High	7.23	0	1

Table 3. Relationship of copper loadings to weights

	Water	Cu, Low	Cu, Medium	Cu, High
Water	1.00	0.90	0.33	0
Cu, Low	0.90	1.00	0.61	0.10
Cu, Medium	0.33	0.61	1.00	0.67
Cu, High	0	0.10	0.67	1.00

Table 4. Agreement weight matrix for copper

3. RESULTS AND DISCUSSION

Tables 5-7 summarize the results of all the calculations. The addition of the buffer zone improved the classification. However, the effect of the neighbourhood analysis is hard to tell. In some cases the accuracy increased, but in other cases, the accuracy decreased, or remained the same.

Pollutant	Overall Accuracy	Weighted Overall Accuracy	Kappa	Weighted Kappa
TSS	92.3	95.2	86.0	86.9
BOD5	92.3	92.4	86.0	84.2
Total P	92.3	93.4	86.0	85.0
TKN	92.3	93.8	86.0	85.4
Cu	85.5	90.8	79.2	78.8
O & G	87.1	87.6	78.0	73.6

Table 5. Accuracy measures for classification with spectral data (in percent)

Pollutant	Overall Accuracy	Weighted Overall Accuracy	Kappa	Weighted Kappa
TSS	92.8	95.4	86.9	87.7
BOD5	92.8	92.9	86.9	85.2
Total P	92.8	93.8	86.9	85.9
TKN	92.8	94.2	86.9	86.3
Cu	86.0	91.2	79.9	79.8
O & G	87.6	88.1	78.9	74.5

Table 6. Accuracy measures for classification with spectral data and buffer zone (in percent)

Pollutant	Overall Accuracy	Weighted Overall Accuracy	Kappa	Weighted Kappa
TSS	92.8	95.3	86.8	87.2
BOD5	92.8	92.9	86.8	85.1
Total P	92.8	93.7	86.8	85.7
TKN	92.8	94.1	86.8	86.1
Cu	86.1	91.0	80.0	79.3
O & G	87.0	87.5	77.9	73.3

Table 7. Accuracy measures for classification with spectral data, buffer zone, and neighbourhood information (in percent)

Overall accuracy values and kappa coefficients were the same for TSS, BOD5, Total P, and TKN for each group of classifications. This was because there were only two states for these pollutants, low loading and high loading, which basically meant separating open land from non-open land. This qualitative assignment of pollution levels did not take into account the difference in magnitudes between pollution levels. With the weighted equivalents of the overall accuracy and kappa coefficient, we observed that these pollutants showed different values, indicating the fact, for example, that among TSS, BOD5, Total P, and TKN, TSS loading classification was the best classified. We also observed that weighted overall accuracy was always higher than overall accuracy. But weighted kappa coefficient could be smaller or larger than kappa coefficient. Naesset (1996) states that these values depend on the dataset and the weights applied.

4. CONCLUSIONS

The weighted equivalents of the overall accuracy and the kappa coefficient provide a new way to look at accuracy measures for assessing the quality of maps made from automated classification of remotely sensed data. This becomes more important especially when classifying ordinal data. Since levels of pollution are only designated as high, medium, and low, these more specific accuracy measures will give better information to users and serve as a guide in designing best management practices.

5. ACKNOWLEDGEMENT

This paper was written in connection with the main author's Ph.D. dissertation, which was funded by the Philippine government through the Department of Science and Technology. She is also grateful to the staff of the UCLA ATS Visualization Portal and Modeling Lab, especially to Yafang Su, and to the staff of the UCLA ATS Statistical Consulting Group.

6. REFERENCES

Abellera, L.V. and Stenstrom, M.K., 2005. Estimation of pollutant loadings from remotely-sensed data with knowledge-based systems and GIS. In: *Proceedings of the 8th Map India Annual International Conference*, New Delhi, India.

Cicchetti, D.V. and Allison, T., 1971. A new procedure for assessing reliability of scoring EEG sleep recordings. *American Journal of EEG Technology*, 11, pp. 101-109.

Cohen, J., 1968. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70, pp. 213-220.

Crist, E.P. and Cicone, R.C., 1984. A physically-based transformation of Thematic Mapper data - The TM tasseled cap. *IEEE Transactions on Geoscience and Remote Sensing*, 22(3), pp. 256-263.

Fleiss, J.L., Cohen, J., and Everitt, B.S., 1969. Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin*, 72, pp. 323-327.

Lillesand, T.M. and Kiefer, R.W., 1994. *Remote Sensing and Image Interpretation*, *3rd Edition*. John Wiley & Sons, Inc., U.S.A., pp. 616-617.

Naesset, E., 1996. Use of the weighted Kappa coefficient in classification error assessment of thematic maps. *International Journal of Geographical Information Systems*, 10(5), pp. 591-603.

Richards, J.A., 1986. *Remote Sensing Digital Image Analysis: An Introduction*. Springer-Verlag, Germany, pp. 190-193.