DETECTING OUTLIERS AND ASSERTING CONSISTENCY IN AGRICULTURE GROUND TRUTH INFORMATION BY USING TEMPORAL VI DATA FROM MODIS

D. Arvor [a], M. Jonathan [b], M. S. P. Meirelles [b,c], V. Dubreuil [a,d]

[a] COSTELUMR 6554 CNRS-LETG, Université Rennes 2, Place du Recteur H. Le Moal, 35043 Rennes Cedex, France - (damien.arvor, vincent.dubreuil)@univ-rennes2.fr
[b] Embrapa Solos, Rua Jardim Botânico, 1024, Rio de Janeiro, RJ – (milton, margareth)@cnps.embrapa.com.br
[c] Universidade do Estado do Rio de Janeiro, UERJ, Departamento de Engenharia de Sistemas e Computação.
[d] Professor visitante, Centro de Desenvolvimento Sustentavel, Universidade de Brasilia (bolsista da CAPES, Brazil)

KEY WORDS: Multi-temporal Image Processing, Land Use, Crop Mapping, Image Understanding, Satellite Remote Sensing

ABSTRACT:

Collecting ground truth data is an important step to be accomplished before performing a supervised classification. However, its quality depends on human, financial and time ressources. It is then important to apply a validation process to assess the reliability of the acquired data. In this study, agricultural infomation was collected in the Brazilian Amazonian State of Mato Grosso in order to map crop expansion based on MODIS EVI temporal profiles. The field work was carried out through interviews for the years 2005-2006 and 2006-2007. This work presents a methodology to validate the training data quality and determine the optimal sample to be used according to the classifier employed. The technique is based on the detection of outlier pixels for each class and is carried out by computing Mahalanobis distances for each pixel. The higher the distance, the further the pixel is from the class centre. Preliminary observations through variation coefficent validate the efficiency of the technique to detect outliers. Then, various subsamples are defined by applying different thresholds to exclude outlier pixels from the classification process. The classification results prove the robustness of the Maximum Likelihood and Spectral Angle Mapper classifiers. Indeed, those classifiers were insensitive to outlier exclusion. On the contrary, the decision tree classifier showed better results when deleting 7.5% of pixels in the training data. The technique managed to detect outliers for all classes. In this study, few outliers were present in the training data, so that the classification quality was not deeply affected by the outliers.

RESUME :

La collecte de données de terrain est une étape importante à accomplir avant d'effectuer une classification supervisée. Cependant, sa qualité dépend des ressources disponibles en temps, argent et personnes. L'application d'un processus de validation pour s'assurer de la fiabilité des données acquises est donc nécessaire. Dans cette étude, des données agricoles ont été obtenues dans l'Etat du Mato Grosso en Amazonie brésilienne afin de cartographier l'expansion des cultures à partir de séries temporelles MODIS EVI. Les données de terrain se basent sur des entretiens réalisés auprès de producteurs. Les informations récoltées concernent les récoltes 2005-2006 et 2006-2007. Cet article présente une méthodologie pour valider la qualité des données d'entraînement et déterminer l'échantillon optimal à utiliser selon le classificateur employé. La technique est basée sur la détection de pixels anomalies pour chaque classe. Pour cela, la distance de Mahalanobis est calculée pour chaque pixel. Plus la distance est grande, plus le pixel se trouve loin du centre de la classe. Ainsi, plusieurs sous-échantillons sont définis en appliquant différents seuils d'exclusion des pixels anomalies du processus de classification. Les résultats de la classification mettent en avant la robustesse des classificateurs Maximum de Vraisemblance et *Spectral Angle Mapper*. En effet, ils se sont avérés insensibles à l'exclusion des pixels anomalies. Au contraire, le classificateur en arbre de décision a donné des meilleurs résultant en éliminant 7.5% des pixels dans les données d'entraînement. La méthode a été efficace pour détecter les anomalies de chaque classe. Dans cette étude, peu d'anomalies ont été repérées dans les données d'entraînement. Ainsi, la qualité de la classification n'a pas été trop affectée par les anomalies.

1. INTRODUCTION

Agricultural statistics are of primary importance in order to monitor crop conditions and their spatio-temporal dynamics. Especially in pioneer frontier regions, it is particularly important to assess the cultivated area corresponding to each crop. Indeed, it offers information on land-use change processes and their environmental impacts. Remote sensing techniques improved agricultural monitoring considerably (Allen *et al*., 2002). For instance, vegetation index data such as those provided by the Moderate Resolution Imaging Spectroradiometer (MODIS) has been proven to be efficient in mapping native vegetation (Anderson, 2005; Jonathan *et al.*, 2007) as well as crops (Doraiswamy *et al.*, 2007; Wardlow *et al.*, 2007). Indeed, these data present high temporal resolutions that are able to capture crop-specific phenological variations over the entire season. However, reliable ground truth data collected on the field are essential to realize efficient supervised classification procedures. Thus, it is fundamental to assert that these training data display high levels of quality and consistency.

In practical cases, collecting ground truth data is a difficult task which aims at collecting the greatest number of consistent data on a wide area depending on the available temporal, human and financial resources. In this study realized in the Brazilian Amazonian State of Mato Grosso, agricultural maps concerning the 2005-2006 and 2006-2007 years were collected through interviews with farmers. This way, a large amount of data could be acquired with low cost. However, the uncertainty due to the insufficient ground controls becomes higher.

This uncertainty can lead to the inclusion of pixels that present greatly disturbed signals. These disturbances can be explained by 4 potential reasons: (i) noisy data due to sensor failure or cloudy situations, (ii) mixed pixels (especialy between crops and forest), (iii) particular crop conditions due to specific events such as drought or disease and (iv) errors in farmer mappings. This last point is the most problematic one. It is difficult to be sure that a farmer who plants more than a thousand hectares will manage to remember which crop he planted in each field 2 or more years ago. Obviously, some farms are already well organized and it is easy to get all the required data. But this way of working is not generalized.

For these reasons, it is important to apply a procedure for criticizing the ground truth data quality before initializing the classification process. It consists in detecting the outliers for each class in order to improve the significance and trustworthiness of the reference data.

Moreover, depending on the classifier to be used, special conditions have to be applied to the training data. For instance, it is generally assumed that the sample size for each class should be of 10 to 100 times higher than the number of bands (Thomas *et al*., 1987). Furthermore, when using a Maximum Likelihood classifier, the training sample is considered to follow a normal distribution. However, remote sensing classes are rarely defined by such a distribution (Sohn and Sanjay Rebello, 2002). On the contrary, a Spectral Angle Mapper (SAM) classifier is not based on a statistical definition of the training samples. It deals with the "spectral form" of the classes (Sohn and Sanjay Rebello, 2002). Thus, it allows working with samples that are not defined by a normal distribution. As such, the SAM classifier is thought to be more robust than statistical based classifiers and more insensitive to variability of VI (Rembold and Maselli, 2006).

The objective of this work is to introduce a method that permits validating the ground truth data collected. This is carried out through the detection of outlier pixels defined by particular temporal VI profiles, which don't correspond to the normal patterns observed for the attributed class.

2. STUDY AREA AND USED DATA

2.1 Study area

The study area is localized in the Brazilian State of Mato Grosso, in the southeastern part of the Amazonian basin (fig. 1). Mato Grosso is known to hold the most active pioneer frontier in the world. Indeed, a huge colonization process has been taking place in this region for the last 30 years (Dubreuil *et al*., 2008). It has led to high rates of deforestation: around 38% (104600 km²) of the Amazonian deforestation between 1992 and 2005 (INPE, 2008). These rates are linked to large expansions of pasture and crop areas : up to 6 millions ha of soybeans in 2005 according to IBGE (2007). The expansion of these new land uses brought many environmental questions concerning the sustentability of such a development. Concerning mechanized agriculture, for instance, increases of the planted areas for soybeans, cotton, corn or sugarcane mean more usage of fertilizers and less biodiverstity, as well as social and economical impacts. As such, it remains important to attempt to adequately map the expansion of these agricultural crops.

2.2 Used Data

Ground truth data : 50 farms were visited and interviews were carried out with farmers. Information about planted crops per field was collected. The mean size of a field is 176 ha. 6 classes were informed: 2 single crop cycles (soybean and cotton) and 4 double-crop cycles (soybean + millet, soybean + sorghum, soybean + corn and soybean + cotton). The detailed collected data are presented in table 1.

Figure 1. Localisation of the Brazilian State of Mato Grosso.

However, tests of separabilty applied to those classes indicated that only 3 classes could be reliably classifed. Thus, the classification is tested only with 3 classes : Soybean + cotton; Cotton and Other Soybean.

	2005-2006	2006-2007
	Hectares	Hectares
Soybean	12112	7798
Soybean+Millet	30066	20832
Soybean+Sorghum	1211	5404
Soybean+Corn	21414	56534
Soybean+Cotton	21149	41549
Cotton	7472	19510
Sum	93424	151627

Table 1. Ground truth data collected.

MODIS Data : The MOD13Q1 products from the MODerate Resolution Imaging Spectroradiometer-TERRA are used in order to study the phenological cycles of crops. Those cycles are analyzed through the calculation of vegetation indices

(NDVI or EVI). In this tudy, the EVI (Enhanced Vegtation Index) has been chosen due to its advantages compared to NDVI. Indeed, it is less affected by atmosphere and soil disturbances. It is also more sensitive than NDVI in areas of high vegetetation activity (Huete *et al*., 1999), such as Mato Grosso.

The EVI is defined as :

$$EVI = \frac{2(NIR - R)}{(L + NIR + C1.R + C2.B)} \quad (1)$$

where *R*, *NIR* and *B* correspond respectively to red, near infra-red and blue bands. *L*, *C1* and *C2* are adjusting parameters to minimise aerosol effects (Huete *et al*., 1999).

The spatial resolution of these data (250 m) is particularly adequate to analyze crops in Mato Grosso. Indeed, the mean area of fields of 176ha allows using such a moderate resolution. The temporal reslution of 16-days (23 images per year) is composed through the Maximum Value Composite method based on daily data (Huete *et al*., 1999). This treatment allows deleting some noise due to cloud effects for instance. However, in tropical regions such as Mato Grosso, cloud effects still remain. A smoothing algorithm was then applied to improve the quality of the EVI profiles. This algorithm is the Weighted Least Squares smoothing algorithm proposed by Swets et al. (1999).

The EVI MODIS data were then acquired, processed and filtered for the referred years so as to build two annual temporal sequences with 23 images each. Moreover, a principal component analysis (PCA) was carried out for each year and the 5 principal components were selected so as to attempt to better capture the main variability factors present within each class.

3. METHODOLOGY

To validate the ground truth data quality and optimize the training sample to be used in the classification process, a methodology that aimed to detect outliers in a multivariate data set was applied. There are a large number of methods in the literature for outlier detection from multivariate data, as reviewed by Ben Gal (2005) and Penny and Jolliffe (2001). Data mining techniques such as clustering are not considered in this study. Indeed, when using clustering, the number of outliers depends on the number of clusters wanted. Moreover, clusters are defined to detect groups of homogeneous pixels, whereas outliers can be represented by isolated pixels.

Thus, the chosen procedure consists in applying a multivariate statistical analysis. The technique is geared towards computing distances between each sample and the remaining pixels of its class. So, it allows identifying which samples are more central and commonplace, as opposed to the ones that present more abnormal behaviour. In order to do that, robust measures of each class's center and covariance matrix are computed, respectively by calculating the median vector of the sample attributes and by computing the minimum covariance determinant (MCD). From this point, Mahalanobis distances are computed for each sample in relation to its class center. The higher the distance, the further the pixel is from the class centre. The Mahalanobis distance is defined by the equation:

$$M_i = \left(\sum_{i=1}^{n} \left(x_i - \overline{x}_n\right)^T V_n^{-1} \left(x_i - \overline{x}_n\right) \right)^{1/2} \quad (2)$$

for $i = 1, \ldots, n$ where n is the sample size, $\overline{x}_n$ is the sample mean vector and V_n is the sample covariance matrix.

This Mahalanobis distance was applied to the collected data on the field in year 2005-2006. The distances were calculated for each class based on the 23 EVI MODIS and on the PCA components.

A threshold is then estimated in order to separate acceptable samples from those considered as outliers. Different thresholds are tested from considering 0% to 20% of outliers to be present in the data set.

The training sample without the outliers is then used to classify the pre-defined classes. Different classifiers are tested in order to to evaluate if the impact of outlier detection on classification depends on the used algorithms. The tested classifiers are Maximum Likelihood, Spectral Angle Mapper (Rembold and Maselli, 2006) and Decision Tree C4.5 (Quinlan, 1996). The classification training is based on year 2005-2006 and applied on year 2006-2007 in order to know if the selected data can be used to classify other years.

4. RESULTS

Results showed that low distance measures could be observed for the majority of each class's samples (fig. 2). It indicates that there are few outliers in each class. Visual inspection of the samples with larger distances confirmed that these MODIS pixels generally corresponded to cases with abnormal phenological responses. Variation coefficients analyses (fig. 3) show that the variability in the detected outlier samples is always higher than in the more confident samples. It thus confirms that the detected outliers correspond to particular pixels, which can potentially deteriorate classification quality. Moreover, studying only those more confident pixels allows representing profiles for each class that can be considered as "correct" pixels or nearly "pure" pixels. Thus, figure 4 presents the different MODIS EVI profiles obtained with samples corresponding to lowest and highest Mahalanobis distances. It appears that the outlier pixels do present different profiles that can potentially affect the classification quality.

Three classifiers were tested with different training data. First, the Mahalanobis are computed on EVI profiles or on PCA components. Then, progressive thresholds are considered to detect outliers (0% to 20% of outliers per sample).

Results are significantly different depending on the classifier used (fig. 5). The Spectral Angle Mapper classifier was the most robust one. It allows keeping good Kappa indices (Kappa > 0.8) even if the training sample size is reduced. Outliers, either detected based on the entire EVI profiles or on PCA components, don't deteriorate the classification quality. This is

linked to the assertions of Rembold and Maselli (2006) who state that this classifier is relatively insensitive to interannual variation of vegetation indices due to meteorological variability. Indeed, those particular situations can lead to outlier pixels which won't affect the classification quality.

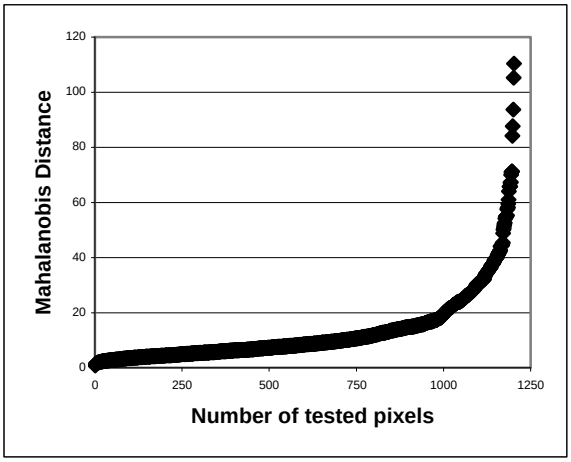

Figure 2. Mahalanobis distances computed for class Cotton.

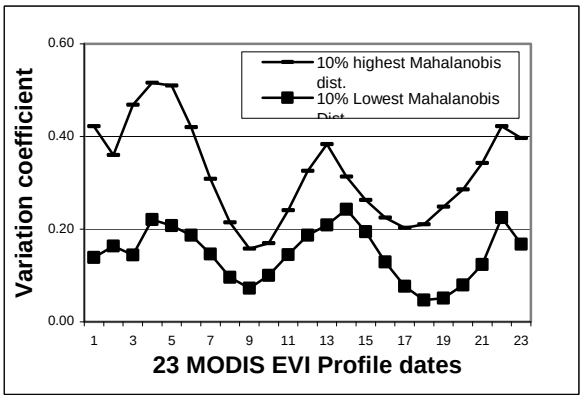

Figure 3. Variation coefficient of EVI profiles for class Soybean + Cotton.

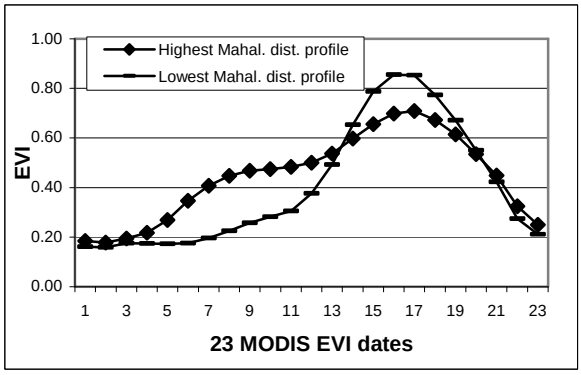

Figure 4. Comparison of temporal MODIS- EVI profiles obtained for the Cotton class obtained by the 10% highest (outliers pixels) and 10% lowest (more reliable pixels) Mahalanobis distance pixels.

The Maximum Likelihood classifier also gave better results when keeping the original data sample. However, the classification quality decreases progressively when excluding outliers. Thus, this classifier seems less robust than the SAM classifier. Furthermore, it seems that a too pure ground truth sample can actually deteriorate the classification quality for this classifier. This is due to the fact that the statistics computed for each class when integrating outliers allow an easier detection of extreme pixels in a class when initiating the classification process. Those extreme pixels can be defined as pixels assigned to a given class but whose EVI profiles don't match that class. On the contrary, the decision tree classifier is well dependent on the input selected data. Indeed, the binary separating thresholds computed at each node of the tree are strongly linked to the input data. Thus, when integrating too many outliers, the thresholds are biased. Then, the classifier will not distinguish if a pixel is an outlier or an extreme pixel for a class. In this study, better Kappa indices are obtained when deleting 7.5% of outliers computed based on the 23 dates entire MODIS EVI profiles. The Kappa index increased from 0.73 (when no outliers are deleted) to 0.80 (when those outliers are deleted).

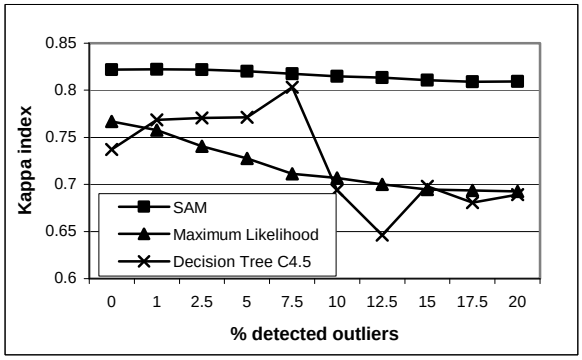

Figure 5. Comparison of tested classifiers depending on outliers detected.

5. CONCLUSION

When performing a supervised classification, it is essential to acquire reliable ground truth data. Constraints on these data can be different depending on the classifier used. A validation step is necessary to ensure the reliability of the collected data and eventually determine an optimal subsample, which improves the classification result. This was carried out through the detection of outliers in each crop class. A methodology based on the computation of Mahalanobis distances has proven to be efficient in performing this task based on annual MODIS EVI profiles. Tested classifiers showed different results when excluding the detected outliers of the training data. Maximum Likelihood and Spectral Angle Mapper classifiers were more robust than the decision tree C4.5. Indeed, the inclusion of outliers did not deteriorate the classification result for the more robust classifiers. The Maximum Likelihood algorithm even led to worst results when excluding outliers, showing that too pure training samples are not ideal for this type of classifier. On the contrary, the decision tree obtained better results when excluding 7.5% of the pixels. Thus, the technique allowed for the validation of the ground truth data acquired through interviews in order to perform crop maps.

ACKNOWLEDGMENTS

This work is helped by the CNPq (through the *ENVIAIR* project linked to the CNPq-INRIA relations), the IAI (Inter American Institute) for Global Change Research through the CRN2 project:"*Land use change in the Rio de la Prata Basin : linking biophysical and human factors to predict trends, assess impacts and support viable strategies for the future"*, the ANR (Agence Nationale de la Recherche) through the *DURAMAZ* project : "*Sustainable development in the Brazilian Amazonia*" and the European Union through the *SENSOR-TTC project : "Land use change, biofuels and rural development in the La Plata Basin"*.

REFERENCES

Allen, R., Hanuschak, G., Craig, M., 2002. History of Remote Sensing from Crop acreage in USDA's National Agriculture Statistics Service. http://www.usda.gov/nass/nassinfo/remotehistory.htm (accessed 9 Apr. 2008).

Anderson, L. O., 2005. Classificação e monitoramento da cobertura vegetal do estado do Mato Grosso utilizando dados multitemporais do sensor MODIS, Master thesis, Instituo Nacional de Pesquisas Espaciais, 249 p.

Ben Gal, I., 2005. *Data Mining and Knowledge Discovery Handbook*. Springer U.S., pp. 131-146.

Doraiswamy, P. C., Stern, A. J., Akhmedov, B., 2007. Crop Classification in the U.S. Corn Belt Using MODIS Imagery, In: *International Geoscience and Remote Sensing Symposium*, Barcelona, Spain. 4 p.

Dubeuil, V., Laques, A.-E., Nédélec, V., Arvor, D., 2008. Paysages et fronts pionniers amazoniens sous le regard des satellites : l'exemple du Mato Grosso, *EspaceGeo.* 21 p.

Huete, A., Justice, C., Van Leeuwen, W., 1999. *Modis vegetation index (MOD13) Algorithm theorical basis document, version 3*, Univerisity of Arizona, 129 p.

IBGE, Instituto Brasileiro de Geografia e Estatistica. 2007. Produção agrícola municipal. Culturas temporárias e permanentes. 1990-2006. http://www.sidra.ibge.gov.br (accessed 10 Oct. 2007).

INPE, Instituto Nacional de Pesquisas Espaciais. 2008. Projeto PRODES. http://www.obt.inpe.br/prodes/prodes_1988_2007.htm (accessed 9 Apr. 2008)

Jonathan, M., Meirelles, M. S. P., da Costa Coutinho, H. L., Berroir, J., Herlin, I., 2007. Aperfeiçoamento do monitoramento do uso e cobertura do solo com dados MODIS a partir da utilização de um diagrama de transição de estados. In: *Anais XIII Simpósio Brasileiro de Sensoriamento Remoto*, Florianopolis, Brazil. pp. 5839-5845.

Penny, K. I., Jolliffe, I. T., 2001. A comparison of multivariate outlier detection methods for clinical laboratory safety, *The Statistician*, 50, pp. 295-308.

Quinlan, J., 1996. Improved Use of Continuous Attributes in C4.5, *Journal of Artificial Intelligence Research*, 4, pp. 77-90.

Rembold, F., Maselli, F., 2006. Estimation of Inter-annual Crop Area Variation by the Application of Spectral Angle Mapping to Low Resolution Multitemporal NDVI Images, *Photogrammetric Engineering & Remote Sensing*, 72(1), pp. 55-62.

Sohn, Y., Sanjay Rebello, N., 2002. Supervised and unsupervised Spectral Angle Classifiers, *Photogrammetric Engineering & Remote Sensing*, 72(1), pp.1271-1280.

Swets, D. L., Reed, B. C., Rowland, J. R., Marko, S. E., 1999. A weighted least-squares approach to temporal smoothing of NDVI. In: *Proceedings of the 1999 ASPRS Annual Conference, From Image to Information*, Portland, Oregon, pp. 526-536.

Thomas, I. L., Ching, N. P., Benning, V. M., D'Aguanno, J. A., 1987. A review of multi-channel indices of class separability, *International Journal of Remote Sensing*, 8(3), pp. 331-350.

Wardlow, B. D., Egbert, S. L., Kastens, J. H., 2007. Analysis of time-series MODIS 250 m vegetation index data for crop classification in the U.S. Central Great Plains, *Remote Sensing of Environment*, 108, pp. 290-310.