Improvement of the accuracy on image classification process through incorporation of contextual information

Leonardo Campos de Assis, Carlos Antônio Oliveira Vieira and **Fabyano Fonseca Silva.**
Federal University of Viçosa
Viçosa-MG, Brazil
{leonardo.assis,carlos.vieira,fabyanofonseca}@ufv.br,

Abstract — **The present study used contextual information modelled through Bayesian Inference to improve image classification in urban areas, with objective to produce a vegetation Map of Belo Horizonte Municipal District (capital of Minas Gerais State – Brazil). Contextual information was inputted into the classification process through ancillary data and specialist knowledge about land cover types. Accuracy assessment shows that despite the amount of generated data (one image based on probability values to each vegetative category), the proposed method improved the image classification accuracy, being more efficient than a conventional Gaussian Maximum Likelihood estimator.**

***Keywords:** contextual information; image processing; Bayesian Inference.*

I. Introduction

Remote sensing applied to urban areas land cover mapping through identifying different categories on earth surface has been proved to be a very difficult task, mainly because of the diversity of possible spectral responses. Natural and anthropic land cover types occupy spatial areas in a dynamic, complex and iterative way. Under this perspective, the correct recognition of the land surface is extremely important for territorial planning and management purposes. Considering this context, the EMDAS (Environmental Municipal District Adjunct Secretary) of Belo Horizonte (BH) proposes a partnership project aiming to produce a thematic map of vegetation categories of entire Municipal District, using Quickbird imagery sets. This vegetation map was used to determine the Green Area Index, one component of the environment variable, employed to establish the Belo Horizonte Urban Life Quality Index (BH-ULQI). BH-ULQI is a spatial index emerged in 1996, composed by 81 Spatial Planning Units of BH that can be viewed as a municipality administrative tool to manage resources distribution with equality. It is composed by 11 variables, 29 components and 75 indexes. This article focuses on the description of a method employed to obtain a thematic map using contextual information modelled through Bayesian Inference to improve image classification in urban area of Belo Horizonte Municipal District.

II. Study Area

Belo Horizonte Municipal District is approximately located at coordinates 19°43'S and 20°03'S of latitude and 43°51'W and 44°04'W of longitude, having an area around 331 Km^2, with last population estimation about 2.5 million of people. The city is featured by intensive urban occupation and concentrates technological and industrial development at several activities sectors like commerce and mining.

The region has a dramatic landscape. The uppermost vegetative cover types found are Native Forest (NF), Planted Forest (PF), CErrado (CE), Cerrado Fields (CF) and Planted FIelds (PI). Anthropic categories were defined as Exposed Soil (ES) and Impermeable (urbanized) Area (IA). Under hydrographical view, the category Water Resources (WR) had to be established. The region exhibits a large number of streams, which at major scale are located on São Franscisco River Basin.

III. Method

The main purpose of this paper is to present an image classification method using ancillary data combined with remote sense image in order to obtain higher accuracy into the classification process. Ancillary data have been used to model contextual information about each category, through a Bayesian Inference approach. In order to verify the accuracy of this method, it was performed a comparison with traditional Gaussian Maximum Likelihood estimator method.

A. Database

Essentially, the databases available were of two major data model types: vector and raster. Raster data model consisted of five Quickbird2 images of different dates in one year period. The vector dataset consisted of the informational layers: streams, contours (with 5 meters of vertical equidistance) and municipal district boundary. Pre-processing operations, such as geometric corrections, were performed in all databases in order to provide consistency and the best spatial agreement.

A careful vector editing was performed to the vector dataset such as system transformation to standardize the projection (UTM) reference systems (SAD 69, 23S) and units (meters). The raster database were submitted to pre-processing procedures that initially reduced the radiometry of imagery set from 11 to 8 bits, because of the classification algorithms employed, inappropriate to handle large values intervals.

1) Orthorrectification

The orthorrectification procedure was performed in each one of the five images that cover the municipal district areas, using the streets intersection and other physical references of the vector database as ground control points (GCP). The ground control points have been collected (where possible) over the entire images in sufficient number to solve the *Rational Polynomial Function* (RPF) model by the *Rational Polynomial Coefficients* (RPC), available at the image metadata files. The criterion to omit or accept a specific GCP was their individual contributions to the total RMS value increased. Orthorrectification procedure has been effectively performed when total RMS value of a specific image was smaller than one pixel.

2) Atmospheric Effects Attenuation

Once the images were from different acquisition dates, the attenuation of the atmospheric effects procedure has been performed. The Antunes-6S model customized software was used (NCAVEO, 2005). Major differences from de original software parameters used were in relation to: sensor calibration, the need to inform the number of the total image pixels and the Top Of Atmosphere (TOA) correction option suppression.

The parameters used in 6S model have been selected according to their availability and Quickbird2 sensor characteristics, once the model data input are flexible. The selected parameters were: Geometric conditions; Atmospheric model; Aerosol model; Horizontal visibility; Altitude of the scene above sea level; Sensor indication (airborne or spaceborne); Bandwidth interval; Reflectance values scale; and, the total number of pixel in the image. Parameter values referred to scene acquisition have been obtained directly from metadata files, climatic conditions of the atmosphere were acquired from a meteorological station approximately located at the area centre and used as reference parameter values for each image in the respective acquisition date. It is necessary to use the Aerosol model in order to inform the visibility parameter, but they were not used because: the visibility parameter was not available at meteorological station data; relative humidity was not greater than 60%, indicating negligible interference of aerosol concentration in the visibility parameter.

B. Ancillary Data

A Digital Elevation Model (DEM) has been generated from the contours, streams and municipal district boundaries vector themes using Topogrid algorithm. Ancillary data were composed by topographic and spectral environmental variables. NDVI was the spectral variable. Topographic variables were generate from the DEM (without spurious depressions), such as Digital Slope Model (DSM), Digital Northness Model (DNM) and Digital Westness Model (DWM). Northness/Southness and Eastness/Westness were determined respectively by cosine and sine operations applied on Digital Aspect Model (DAM) - both derived from DEM (Domaç and Süzen, 2006).

C. Specialist Knowledge and Contextual Information

Specialist knowledge applied to image classification process has been largely mentioned at literature, such as Cayuela *et al.* (2006), Daniels (2006) and Domaç and Süzen (2006). Baltsavias (2004) reviews knowledge representation methods which improve image classification accuracy and mentions some important aspects relative to specialist knowledge, such as: their distinguished types; associated problems in applications; their management and representation; the current and possible employment; and exploration, updating and acquisition. Thus, contextual information can be used for the purpose of knowledge representation. Despite contextual information has been found in several papers as in Stuckens *et al.* (2000), De Jong *et al.* (2001), Laha *et al.* (2006), which the authors use the term in different semantic meanings, depending on the support reference. Ideally, the contextualization of the factors associated with a specific land cover category occurrence needs specialist knowledge about them.

The work hypothesis is that contextual information can help to identify certain vegetation classes that occur associated with specific conditions, characterized by the environment, and such conditions can be modelled through specialist knowledge employing a Bayesian inference model.

D. Bayesian Inference and Image Classification

There are several possibilities of Bayesian paradigm application to improve the results of image classification, like data integration (Melgani and Serpico, 2002; Malpica, Alonso, Sanz, 2007) and specialist knowledge incorporation (Abkar, Mohammed, Mulder, 2000), among other ones.

To start with, Bayes theorem as presented in (1), which $p(y)$ assume the form $\sum_{\theta} p(\theta)p(y|\theta)$ for discrete values and the form $\int p(\theta)p(y|\theta)\partial\theta$ for continuous values, to guarantee exchangeability of θ. One can note that $(p(y))$ data is independent of the hypothesis (θ), conducting to an equivalent form of Bayes theorem (2) composed by the likelihood and the prior density functions.

$$p(\theta|y) = \frac{p(y|\theta).p(\theta)}{p(y)} \quad (1)$$

$$p(\theta|y) \propto p(\theta).p(y|\theta) \quad (2)$$

Where θ means the hypothesis, given the data y. According to Bayes theorem statements, $p(\theta)$ is called prior probability, that is adjusted by specialist knowledge specifications about hyper-parameters (parameters that define the core of prior density function).

E. Probabilistic Model Determination

Probabilistic model election has been made considering that the sample space is for each individual pixel and the desired inference is about the pixel probability of a specific land cover category that needs to be determined. Such features points to a binomial density as likelihood (3) and a *beta* density as prior (4), which lead to a *beta*-binomial form (5) obtained by their multiplication (2), also called posterior density.

$$p(y|\theta,n) \propto \theta^{y}.(1-\theta)^{n-y} \quad (3)$$

$$p(\theta|\alpha,\beta) \propto \theta^{\alpha-1}.(1-\theta)^{\beta-1} \quad (4)$$

$$p(\theta|\alpha,\beta,n,y) \propto \theta^{y+\alpha-1}.(1-\theta)^{n-y+\beta-1} \quad (5)$$

Where α and β (prior hyperparameters) are respect to the position and dispersion of prior density core that values could vary from 0 to 1; n are the number of possible land cover categories (possible events) and y is the effective land cover that have occurred (success event). This guarantees unique values to n and y, respectively 5 and 1, for all the vegetation categories. Although, α and β assumed large variations and have being defined iteratively by simulation, it is possible to test specific values and visually analyzing the result obtained, looking at a more realistic scenario. The model chosen (5) has a convenient form, because it has a proper prior, what is, the densities are of the same family (more details can be consulted at Gelman *et al.* (1995)).

Initial probability values *(θ)* determination has been made considering a *Binomial Logistic Regression* (BLR) model due to its capability to describe statistical relation between independent variables of continuous values (contextual information) and a binary dependent variable (training site). Finally, after BLR solution and replacement on (5), the corrected probability values to all the considered variables under the Bayesian approach conducted to (6).

$$p(\theta|\alpha,y,\beta,n) \propto [L(\theta)]^{y+\alpha-1}.[1-L(\theta)]^{n-y+\beta-1} \quad (6)$$

Where L means application of BLR to this model solution.

F. *Accuracy Assessment*

Evaluation has been performed during several phases of this research; however, only the most relevant ones will be mentioned. *Relative Operation Characteristic* (ROC) procedure has been applied to BLR model in order to fitting quality measurement. ROC values can vary from 0 to 1, where one means a perfect fit and, 0.5 only a fit by chance.

Image classification assessment has been made by comparison between conventional Gaussian Maximum Likelihood Classification (MLC) and the postulated method, named here as Contextual Classification (CC). Evaluation has been performed applying the statistical Z test at parameters Kappa coefficient and its variance, which verifies significant difference between independent Kappa coefficient values. Two statistical hypotheses were tested at 0.05% significance level, nullity (equivalence) and alternative (difference). Same reference test sites were used to evaluate both methods with their respective classified test sites. Sample pixels for training and test sites of all land cover categories were collected randomly in the respective images. Test sites were used after a Mahalanobis distance algorithm operation using an arbitrary threshold value of 50%, with intent to purify the samples from "in-class" large variations.

IV. Results

BLR evaluation through ROC procedure has presented good results to NF (0.92) and CF (0.91) categories, moderate results to CE (0.72) and PI (0.71) categories and just one category, PF (0.63), presented poorer results. Apparently, ROC values are not associated with the number of independent variables fitted to a specific category model.

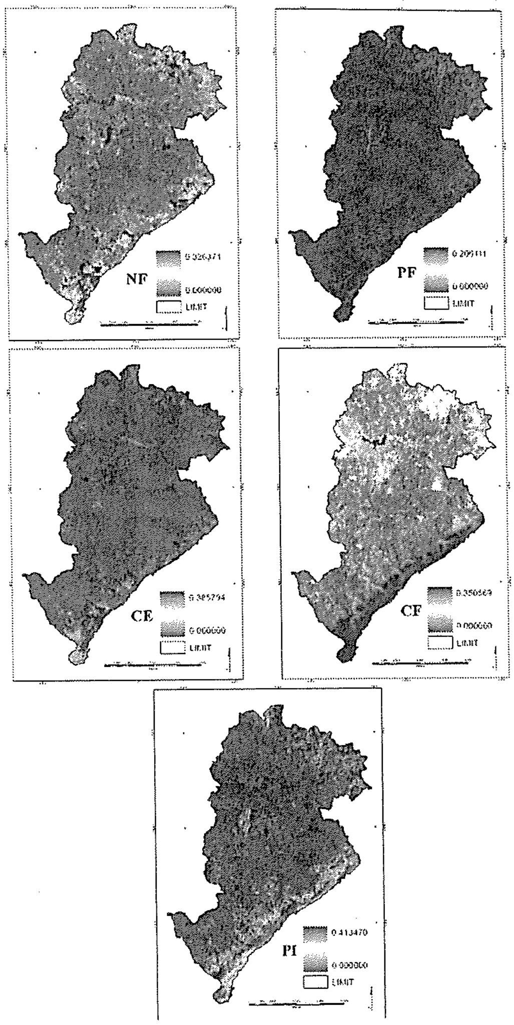

Figure 1. Images based on probability values for vegetative land cover categories (NF, PF, CE, CF and PI) at Belo Horizonte Municipal District boundary. Blue colours means low values and red colours high ones.

Hyperparameter values have been obtained in an onerous procedure, by simulating and iteratively checking the intermediate results, until reach proper values. Difficulties were greater at the beginning stage, where the model sensibility had not been verified yet, but, after a few iterations the process becomes more efficient. Table I presents hyperparameter values used, from which image based on probability values were generated for each category, as showed in Fig. 1. It can be verified the homogeneous PF interval distribution, expressing low efficiency to this specific category distinction and, therefore, to entire process. It probably occurred due to the poor BLR fitting result, what means that, despite the *beta*-binomial model flexibility, the BLR deserves special care on fitting assessment quality.

TABLE I. TABLE OF HYPERPARAMETER VALUES PER CATEGORY

Categories	Hyperparameter Values	
	α (position)	*β (dispersion)*
NF	0.34	0.20
PF	0.30	0.70
CE	0.26	0.60
CF	0.31	0.20
PI	0.18	0.22

Images based on probability values were used as prior probability values to MLC method performing the Contextual Classification; and the MLC reference method was employed with equal prior probability values.

Classification accuracy assessment is presented on Table II, where Kappa coefficients, their variances, and statistical Z values for each method can be observed. One can note that CC Kappa coefficient value was superior to Gaussian MLC traditional method. Considering the two-sided statistical Z test, evaluated at 95% confidence level, Z value is greater than 1.96 (computed Z value 1835.098), so nullity hypothesis must be rejected. This means that there is significant difference between the classification methods. Then, the CC method has been used to generate the Vegetation Land Cover Map of Belo Horizonte, presented in Fig. 2.

TABLE II. TABLE OF CLASSIFICATION ASSESSMENT PARAMETERS

Assessment parameter and Classification Method	Classification Method	
	MLC	*CC*
Kappa	0.85	0.92
Kappa variance	0.00000013	0.000000076

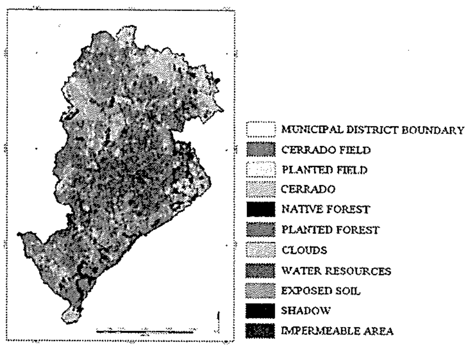

Figure 2. Vegetation map of Belo Horizonte Municipal District produced by Contextual Classification method.

V. CONCLUSIONS

Bayesian inference resources applied to contextual information modelled shows efficiency to urban area image classification process, enabling human specialist knowledge introduction into the classification process. Aggregation of environmental and spectral variables extended benefits to classification procedure, in spite of computer time consuming to produce the images based on probability values. The postulated CC method improved classification results on urban area, a very well known difficult scenario to remote sensing.

ACKNOWLEDGEMENTS

Authors thank to the EMDAS (Environmental Municipal District Adjunct Secretary) of Belo Horizonte for raster imagery supplied and the PRODABEL Cia. for vector dataset provided. The authors also thank the Fundação de Amparo a Pesquisa no Estado de Minas Gerais (FAPEMIG) for their support

REFERENCES

Abkar, A. A., Mohammed, A. S. and Mulder, N. J. (2000). Likelihood-based image segmentation and classification: a framework for the integration of expert knowledge in image classification procedures. *International Journal of Applied Earth Observation and Geoinformation*. 2(2), 104-119.

Baltsavias, E. P. (2004). Object extraction and revision by image analysis using existing geodata and knowledge: current status and steps towards operational systems. *ISPRS Journal of Photogrammetry and Remote Sensing*. 58, 129-151.

Cayuela, L, Golicher, J. D., Salas Rey, J. and Rey Benayas, J. M. (2006). Classification of a complex landscape using Dempster-Shafer theory of evidence. *International Journal of Remote Sensing*. 27(10), 1951-1971

Daniels, A. E. (2006). Incorporating domain knowledge and spatial relationships into land cover classifications: a rule based approach. *International Journal of Remote Sensing*. 27(14), 2949-2975.

De Jong, S. M., Hornstra, T. and Maas, H. (2001). An integrated spatial and spectral approach to the classification of Mediterranean land cover types: the SSC method. *International Journal of Applied Earth Observation and Geoinformation*. V.3, n.2. p.176-183.

Domac, A. and Süzen, M. L. (2006). Integration of environmental variables with satellite images in regional scale vegetation classification. *International Journal of Remote Sensing*. 27(7), 1329-1350.

Gelman, A., Carlin, J.B., Stern, H. S. and Rubin, D. B. (1995). *Bayesian data analysis*, 1st ed., London: Chapman and Hall.

Laha, A., Pal, N. R., and Das, J. (2006). Land cover classification using fuzzy rules and aggregation of contextual information through evidence theory. *IEEE Transactions on Geoscience and Remote Sensing*. 44(6), 1633-1642.

Malpica, J. A., Alonso, M. C., and Sanz, M. A. (2007). Dempster-Shafer theory in geographic information systems: a survey. *Expert Systems with Applications*. V.32. p.47-55.

Melgani, F.; and Serpico, S.B. (2002). A statistical approach to the fusion of spectral and spatio-temporal contextual information for the classification of remote sensing images. *Pattern Recognition Letters*. 23, 1053-1061.

NCAVEO Network for Calibration and Validation of Earth Observation. (2005). *Atmospheric correction*. Example 1. Atmospheric correction of a Landsat TM image. Available at: <http://www.ncaveo.ac.uk/special_topics/atmospheric_correction/example1/?link=andtemplate=print.inc>.Access in: May 5th, 2008.

Stuckens, J., Coppin, P. R. and Bauer, M. E. (2000). Integrating contextual information with per-pixel classification for improved land cover classification. *Remote Sensing of Environment*. 71, 282-296.