A multitemporal approach for land use mapping using Bayesian Networks

Alexsandro Cândido de Oliveira Silva
Leila Maria Garcia Fonseca
Thales Sehn Körting

National Institute for Space Research (INPE)
Post Office Box: 515 – 12227-010 – São José dos Campos – SP, Brazil
{alexsandro.silva, leila.fonseca, thales.korting}@inpe.br

Abstract: It is possible to trace the phenological profile of targets on the Earth's surface through multitemporal remote sensing data. Different features can be computed from multitemporal data to classify land use classes. In this context, this paper presents a new method to map the land use based on the probabilistic analysis of multitemporal features using Bayesian Networks. Elementary statistical measures were computed from NDVI/MODIS and EVI/MODIS time series of pasture, sugarcane, annual agriculture and other uses classes for 2012/2013 and 2013/2014 crop years in southern Goiás state, Brazil. The model's output is composed by layers representing the occurrence probability of each class over the study area. A thematic map was built from output layers and the classification was evaluated by the Monte Carlo simulation. In our preliminary results, we obtained classification accuracy values within Kappa index range from 0.51 to 0.63. Annual agriculture and other land use classes were more easily distinguished and more confusion happened between pasture and sugarcane classes. Although the accuracy values were not high, the proposed model presented a potential for land use classification and it can be improved.

Palavras-chave: multitemporal data, time series metrics, Bayesian Networks

1. Introduction

The massive amount of data provided by satellites made possible the analysis of objects on the surface through images time series. Multitemporal images can be represented in a sequence of raster data, which are used to extract a sequence of values for each object in different time intervals (KÖRTING et al., 2013).

Moderate Resolution Imaging Spectroradiometer (MODIS) is an important data source due its high temporal resolution allowing the surface analysis in time and space. Studies have used multitemporal sequences of Normalized Difference Vegetation Index (NDVI) and Enhanced Vegetation Index (EVI) data (HUETE et al., 2002) from MODIS to map the land use (XAVIER et al., 2006; ABADE et al., 2015).

Through multitemporal data characterization it is possible to determine what can and what cannot be classified in the data with a minimum of assumptions. Statistical measures computed on multitemporal data can be used as input features to classify land use classes with different phonological patterns. Hüttich et al. (2009) and Arvor et al. (2011) used MODIS times series metrics to map land use through different classifiers like maximum likelihood, decision tree and random forest.

In this context, this paper presents preliminary results of a new method for land use mapping based on probabilistic analysis of multitemporal data features using Bayesian Network model. The enhanced Bayesian Network for Raster Data (e-BayNeRD) method (SILVA et al., 2014) was used to map crops and other land uses in the southern of Goiás state, Brazil. The e-BayNeRD is based on raster data observation and it is able to incorporate experts' knowledge for analysis. In the next section we present a brief description about the theory of Bayesian Networks and the e-BayNeRD method employed in this study.

2. Bayesian Networks

Bayesian Networks are defined in terms of two components: (i) qualitative component – a Directed Acyclic Graph (DAG), in which the nodes represent the variables in the model and the statistical dependence between pair wise variables is indicated by directed arrows; and (ii)

quantitative component – probability functions associated to each variable denoting the strengths of the links in the BN model (AGUILERA et al., 2011).

The prior knowledge of an event is updated taking into account new evidence through the Bayes' theorem (NEAPOLITAN, 2004):

$$P(A = a|B = b) = \frac{P(B = b|A = a)P(A = a)}{P(B = b)} \quad (1)$$

in which $P(A = a)$ is the prior probability of the event A; $P(B = b \vee A = a)$ is the likelihood function and $P(A = a \vee B = b)$ is the posterior probability; and $P(B = b)$ is a normalizing constant. Upper-case letters denote the variables and the same but lower-case letters denote the state or value of the variable. This ability to compute posterior probabilities given new evidence is called inference.

3. The e-BayNeRD algorithm

Figure 1 shows the e-BayNeRD algorighm's workflow (SILVA et al., in press). The method deals with raster data in GeoTIFF format, in which every GeoTIFF data corresponds to a variable (node) in the Bayesian Network model. The variable that represents the studied phenomenon is called *target variable* and its GeoTIFF must contain the reference data for training. Others variables in the model are called *context variables*. After entering all the raster data, the user designs the Bayesian Network graphic model (i.e., the DAG) to define the statistical dependence among all variables.

In the next step, the user needs to convert continuous context variables into categorical ones. The range of observed values for each variable is divided into intervals according to the lower and upper limits chosen by the user. Each interval defines one category; therefore, a variable will have n categories if user discretizes it into n intervals. The limits of the intervals for each variable should be appropriately chosen to describe as best as possible the variable according to the phenomenon studied.

In the e-BayNeRD algorithm, probability functions are computed based on pixel counting according to the dependence relations among variables and their categories. Prior probability is assigned to those variables without parents, whereas conditional probability is assigned to descendant ones. After compute the probability functions associated with each variable, e-BayNeRD is able to calculate the probability of target presence given the values observed in the context variables. When the probability is computed for each pixel in the study area, the output, called Probability Image, is formed.

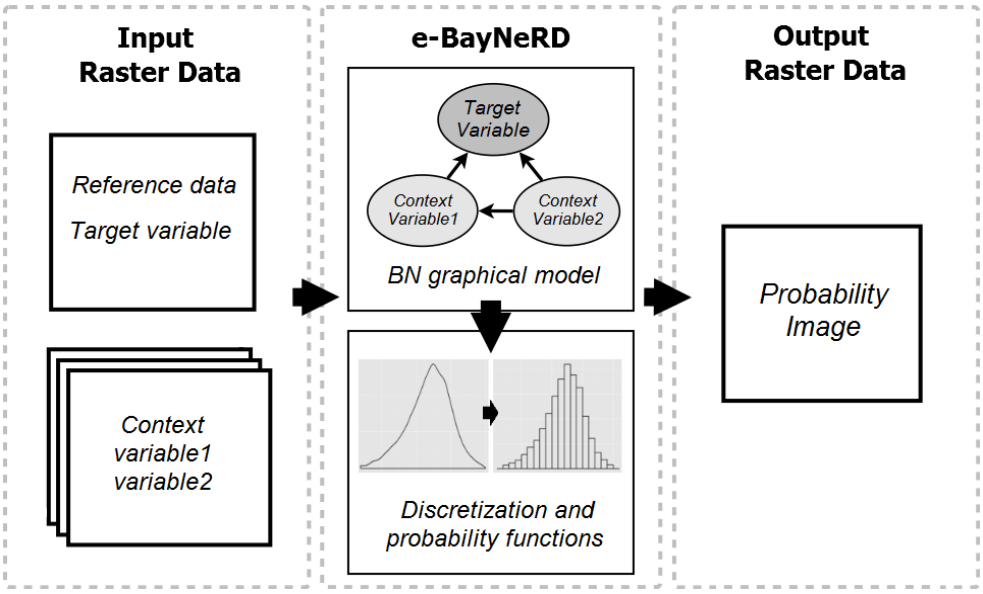

Figure 1. e-BayNeRD's workflow; adapted from Silva et al. (in press).

4. Study Area

The Central-West region of Brazil, which previously had an economy centered on cattle ranching and grains, has witnessed an intense sugarcane expansion, mainly in the central-south region of Goiás State (SILVA; MIZIARA, 2011; SHIKIDA, 2013). In the study area selected for this work we can find pasture, sugarcane and annual agriculture areas, mostly soybeans. It covers four municipalities located in the southern of Goiás State: Santo Antônio da Barra, Santa Helena de Goiás, Acreúna and Turvelândia, as presented in the Figure 2.

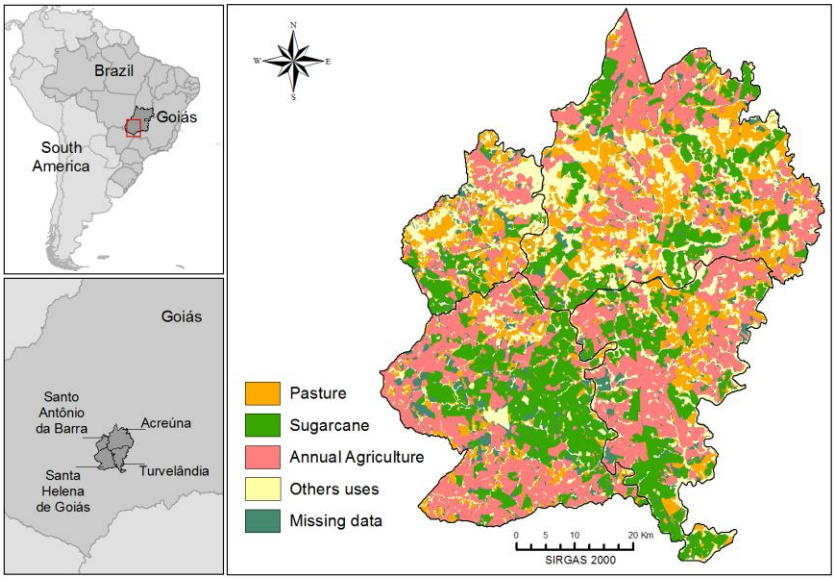

Figure 2. Study area and land use classes.

5. Variables selected to compose the Bayesian Network model

5.1 Target variable

Figure 2 shows the land use of the study area, which comprises pasture, sugarcane and annual agriculture areas for 2013/2014 crop year according to the Image Processing and Geoprocessing Laboratory – LAPIG (http://maps.lapig.iesa.ufg.br/). The class other uses include urban areas, water bodies, native vegetation and areas under environmental protection laws. Areas labeled as missing data means that no observations were made.

The GeoTIFF data representing the *target variable Land Use* as reference data for training is composed by 4 classes: (i) *pasture*; (ii) *sugarcane*; (iii) *annual agriculture*; and (iv) *other uses*. About 70% of the pixels in each class was randomly selected to compose the reference data for training. The remaining 30% was used for accuracy assessment.

5.2 Context variables

Context variables are statistical measures calculated from the NDVI and EVI time series for each pixel inside the study area. In the next sections, we clarify how the context variables were computed.

5.2.1 Temporal image data set

NDVI and EVI images derived from MODIS onboard Terra and Aqua satellites are available with spatial resolution of 250m and 16-day composite of cloud-free (HUETE et al., 2002). There is an interval of 8-day between the products from MODIS/Terra and MODIS/Aqua. Data from both platforms were combined to create NDVI and EVI time series with twice as many observations.

We fitted the original time series using a cubic smoothing spline (R CORE TEAM, 2016) to reduce the noise, and also to preserve the patterns in the original NDVI and EVI series. Figure 3 illustrates an example of the temporal EVI profile for *pasture*, *sugarcane* and *annual*

agriculture classes, in which the original data is represented by black line and the smoothed data by red line. Two periods of NDVI and EVI images were considered: (i) in the 2013/2014 crop year with images from June 2013 to December 2014; and (ii) in the previous 2012/2013 crop year with images from June 2012 to December 2013.

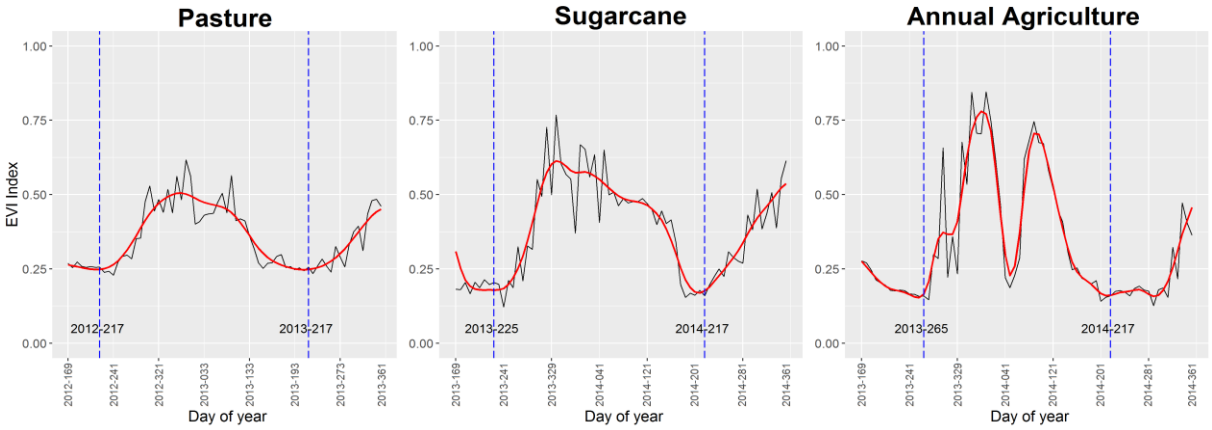

Figure 3. Examples of original (black line) and smoothed (red line) time series with the cycle defined by dotted blue lines.

It is important to define the initial and the final points of the growing season for each land use. In agriculture applications, for example, the cycle may be longer or shorter depending on the planting and harvesting dates. Taking this into account, we developed a script to find the local minimum from the time series peaks within the specified periods, as shown in Figure 3 (dotted blue lines). Hence, time series features can be computed considering a well defined cycle.

5.2.2 Time series features

From each cycle we traced simple statistical measures such as mean, standard deviation, amplitude, and sum (HÜTTICH et al., 2009; KÖRTING et al., 2013). However, only the metrics EVI time series amplitude, NDVI time series amplitude and standard deviation were chosen to compose the Bayesian Network model because they were more relevant to differentiate the four land use classes in the study area.

Figure 4 shows the histogram and boxplot of the chosen metrics for each class in both 2012/2013 and 2013/2014 crop year. In general, the metrics values are lower for *other uses* class followed by *pasture*, *sugarcane* and *annual agriculture* classes. The overlapping areas among the histograms mean that some classes are represented by the same features, which can lead to confusion in the classification.

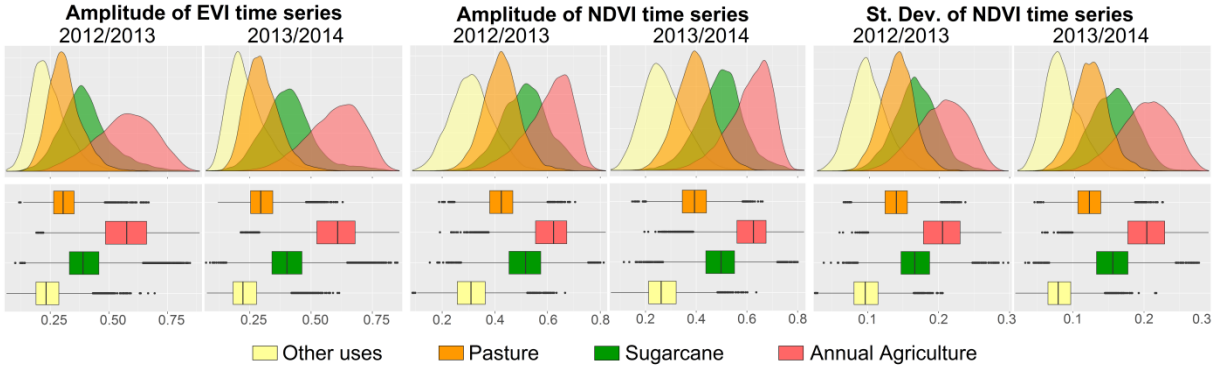

Figure 4. Histogram and boxplot of the chosen metrics (amplitude of EVI and NDVI; and standard deviation of NDVI) for each land use class in the study area.

The chosen metrics were computed from the NDVI and EVI time series for each pixel inside the study area. Therefore, six raster data were built to represent the EVI time series amplitude, NDVI time series amplitude, and standard deviation for both 2012/2013 and 2013/2014 crop year. These raster data were selected as context variables to compose the Bayesian Network model.

5.3 Bayesian Network graphical model

We built the Bayesian Network graphical model by setting up directed arrows between a pairwise of variables (*target* and *context*) to establish statistical dependencies among them. *Land Use* variable data refer to the 2013/2014 crop year. Given that there were no considerable changes in the land cover from the previous crop year to the next one, we settled context variables related to the 2012/2013 crop year as parents of the *Land Use* variable. Considering that the presence of the target variable classes leads to specific NDVI and EVI values, we set context variables related to the 2013/2014 crop year as descendant of *Land Use* variable.

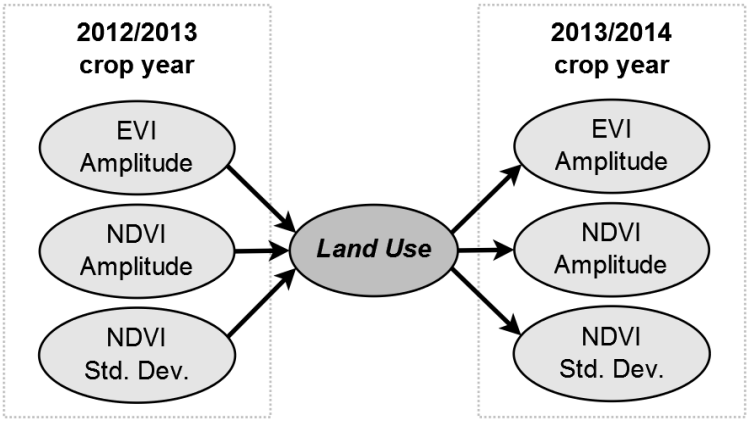

Figure 5. Bayesian Network model.

6. Results and discussion

Once the Bayesian Network graphical model is defined, context variables must be discretized and the interval limits should be appropriately chosen to describe as best as possible the context variable according to the *Land Use* variable classes. After the discretization, the probabilities associated to each variable are computed based on pixels frequency counting according to the dependence relations among the variables as well as their defined intervals. Prior probability is assigned to those variables without parents, whereas conditional probability is assigned to descendant ones. The reader is referred to Silva et al. (2014) for details about how to comput probability functions.

With all the probability functions computed, e-BayNeRD algorithm calculates the probability of target presence given the values observed in the context variables for each pixel in the study area. The number of classes in the target variable defines the number of e-BayNeRD's output layers, which represent the occurrence probability of each class. As *Land Use* variable has four classes (*other uses*, *pasture*, *sugarcane* and *annual agriculture*), four layers will be produced, as shown in Figure 6. Dark colored pixels represent areas with high occurrence probability of a class.

We can verify that the *other uses* (Figure 6-a) and *annual agriculture* (Figure 6-d) classes are well defined, i.e., these areas exhibit high occurrence probability values. Such classes present distinct temporal behavior and, consequently, their time series metrics values are different, as one can observe in Figure 4. Therefore, the algorithm can more easily differentiate *other uses* and *annual agriculture* classes.

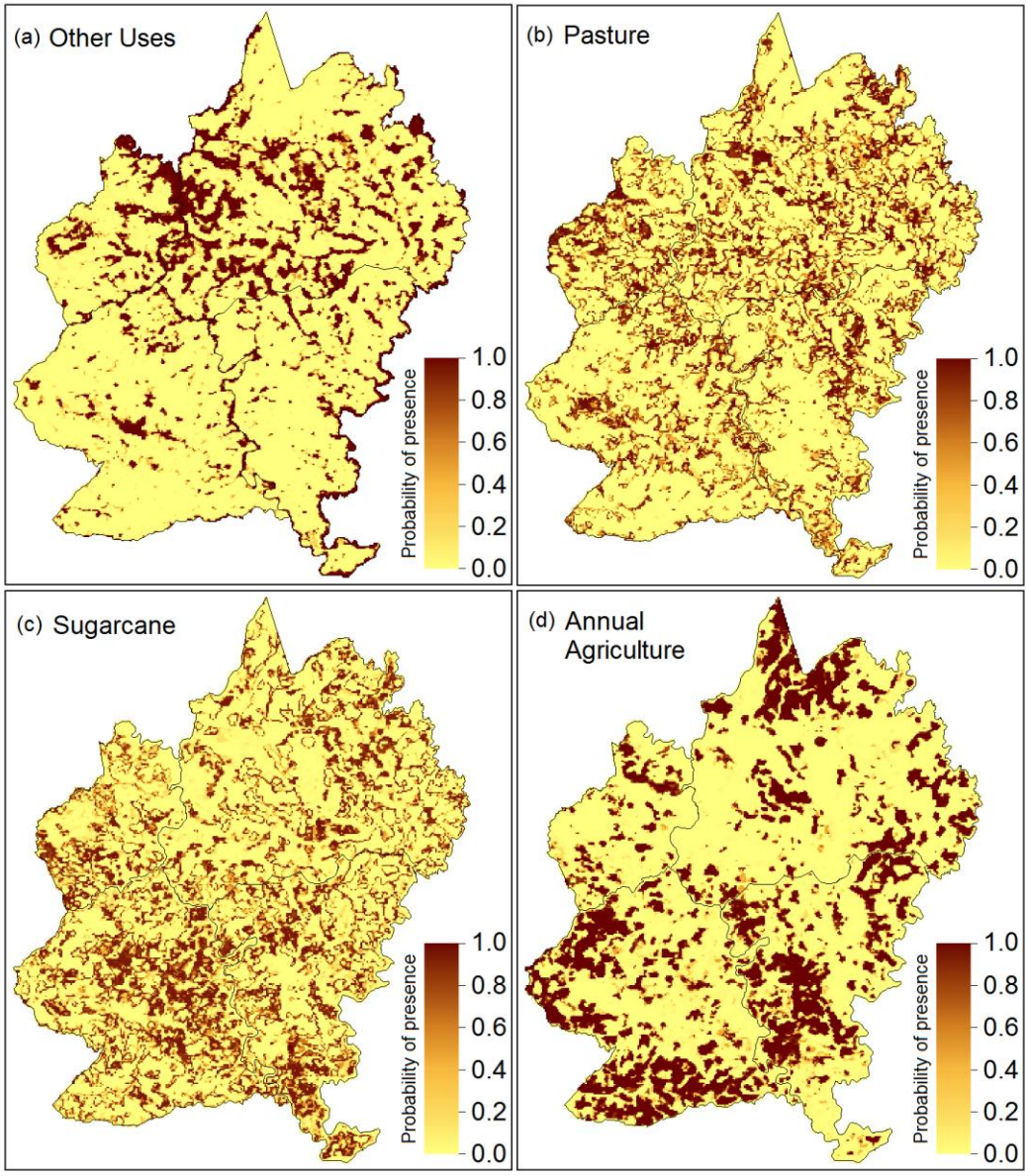

Figure 6. Resulting layers with values of occurrence probability of classes: (a) other uses; (b) pasture; (c) sugarcane; and (d) annual agriculture.

One the other hand, e-BayNeRD algorithm had more difficult to distinguish the *pasture* and *sugarcane* classes. The resulting layers for these classes (shown in Figures 6-b and 6-c) present many regions with intermediate probability values, mainly near the edge. It means that the algorithm had uncertainties about the occurrence of *pasture* and *sugarcane*. As both classes have similar temporal behavior, as reported by Rudorff et al. (2009), some metrics used to characterize them can present similar values (see the histograms in Figure 4) and then make difficult to discriminate them. Uncertainties in the edge regions were due the data spatial resolution of 250m employed in the Bayesian Network model.

From the e-BayNeRD's output layers, we can build several thematic maps with different scenarios by varying the layer probability values. We evaluated the model building a thematic map by stacking the resulting layers and analyzing the pixels in depth. Each pixel received the label of the class whose layer had the highest occurrence probability value. Figure 7 presents the resulting thematic map.

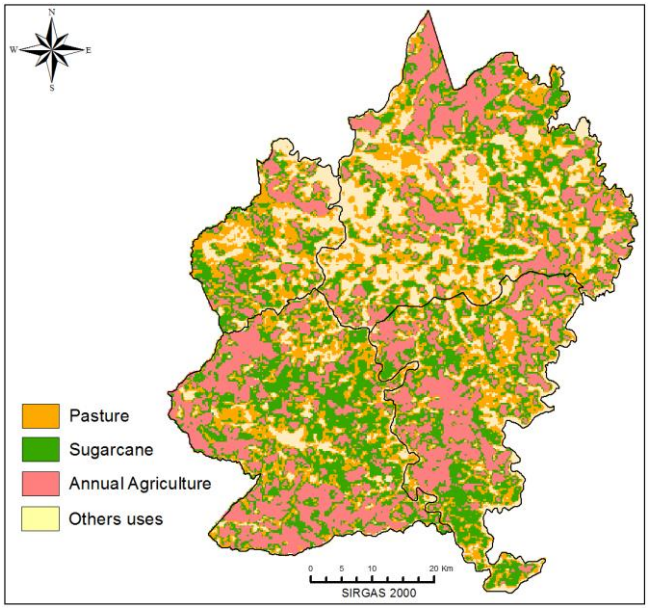

Figure 7. Land use thematic map of the study area.

The thematic map was evaluated using the confusion matrix and the Monte Carlo simulation. In each step, 100 random samples from each class were selected and the overall accuracy and Kappa index were computed. Hence, it was possible to create an interval for the Kappa index. In 95% of the experiment, Kappa index was within interval [0.51, 0.63]. Figure 8 shows the results for 5000 iterations along with the Kappa values interval.

Table 1 shows an example of a confusion matrix that generated a Kappa index = 0.63. One can notice that the lowest classification accuracy values occurred for the *pasture* and *sugarcane* classes. This confirms the uncertainty about the occurrence probability for both classes (Figure 6-b and 6-c).

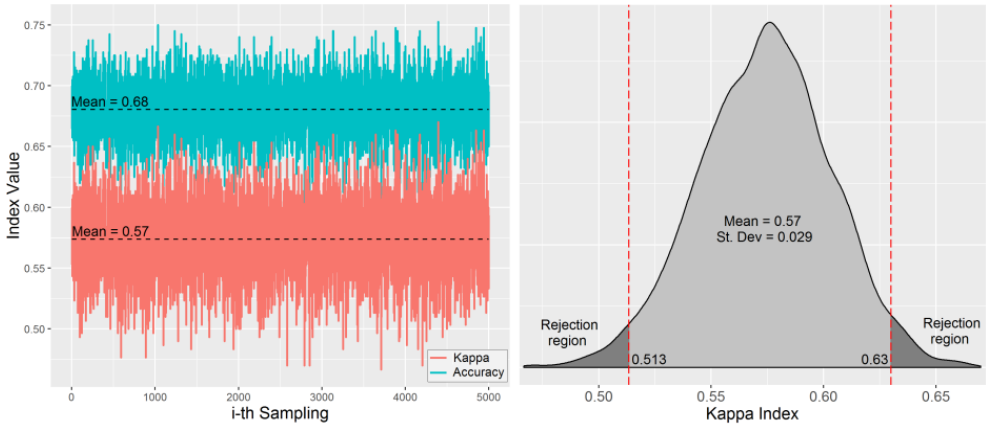

Figure 8: Overall accuracy and Kappa index in the Monte Carlo simulation (left); and interval of Kappa values (right).

Table 1. Confusion matrix.

		Reference				
		Other uses	**Pasture**	**Sugarcane**	**Annual Agriculture**	**Commission Error**
Classification	**Other uses**	78	18	1	0	19.6%
	Pasture	21	58	17	3	41%
	Sugarcane	1	24	67	10	34.3%
	Annual Agric.	0	0	15	87	14.7%
	Omission Error	22%	42%	33%	13%	

7. Conclusion

This paper presented a method for land use mapping using Bayesian Network model based on multitemporal data features. The method was tested to map *pasture*, *sugarcane*, *annual agriculture* and *other uses* classes in the southern of Goiás state, Brazil. The features to characterize theses classes were computed based on NDVI and EVI time series.

The model's output layers indicate the occurrence probability of each class. The model had more uncertainties about pasture and sugarcane occurrence due their similar temporal behavior. The *annual agriculture* and *other uses* classes were easily differentiated. Decision uncertainties were also observed in the edge regions, which can be explained by the coarse spatial resolution of the raster data used in this study.

Although the classification accuracy values were not good, the proposed method showed potential for land use mapping and it can be improved by using raster data with higher spatial resolution and other time series features to better characterize the temporal patterns.

References

ABADE, N. A. et al. Comparative analysis of MODIS time-series classification using support vector machines and methods based upon distance and similarity measures in the brazilian cerrado-caatinga boundary. **Remote Sensing**, v. 7, n. 9, p. 12160–12191, 2015.

AGUILERA, P. A. et al. Bayesian networks in environmental modelling. **Environmental Modelling & Software**, v. 26, n. 12, p. 1376–1388, dez. 2011.

ARVOR, D. et al. Classification of MODIS EVI time series for crop mapping in the state of Mato Grosso, Brazil. **International Journal of Remote Sensing**, v. 32, n. 22, p. 7847–7871, 2011.

HUETE, A. et al. Overview of the radiometric and biophysical performance of the MODIS vegetation indices. **Remote Sensing of Environment**, v. 83, n. 1–2, p. 195–213, 2002.

HÜTTICH, C. et al. On the suitability of MODIS time series metrics to map vegetation types in dry savanna ecosystems: A case study in the Kalahari of NE Namibia. **Remote Sensing**, v. 1, n. 4, p. 620–643, 2009.

KÖRTING, T. S.; GARCIA FONSECA, L. M.; CÂMARA, G. GeoDMA—Geographic Data Mining Analyst. **Computers & Geosciences**, v. 57, p. 133–145, ago. 2013.

NEAPOLITAN, R. E. **Learning Bayesian Networks**. 2. ed. New Jersey: Person Prentice Hall, 2004.

R CORE TEAM. **R: A Language and Environment for Statistical Computing**Vienna, Austria, 2016. Disponível em: <https://www.r-project.org/>

RUDORFF, B. F. T. et al. **Temporal series of EVI/MODIS to identify land converted to sugarcane**. 2009 IEEE International Geoscience and Remote Sensing Symposium. **Anais**...IEEE, jul. 2009Disponível em: <http://ieeexplore.ieee.org/document/5417326/>

SHIKIDA, P. F. A. Expansão canavieira no Centro-Oeste: limites e potencialidades. **Revista de Política Agrícola**, v. XXII, n. 2, p. 122–137, 2013.

SILVA, A. A.; MIZIARA, F. Avanço do setor sucroalcooleiro e expansão da fronteira agrícola em Goiás. **Pesquisa Agropecuária Tropical**, v. 41, n. 3, p. 399–407, 6 jul. 2011.

SILVA, A. C. O.; FONSECA, L. M. G.; KÖRTING, T. S. **Bayesian network model to predict areas for sugarcane expansion in Brazilian Cerrado**. XVII Brazilian Symposium on GeoInformatics. **Anais**...Campos do Jordão: [s.d.]

SILVA, A. C. O.; MELLO, M. P.; FONSECA, L. M. G. **Enhancements to the Bayesian Network for Raster Data (BayNeRD)**. (C. A. Davis Jr, K. R. Ferreira, Eds.)Proceedings of XV Brazilian Symposium on Geoinformatics. **Anais**...Campos do Jordão: Proceedings of the XV Symposium on GeoInformatics - GEOINFO, 2014

XAVIER, A. C. et al. Multi-temporal analysis of MODIS data to classify sugarcane crop. **International Journal of Remote Sensing**, v. 27, n. 4, p. 755–768, 2006.