HIDDEN MARKOV MODELS APPLIED IN AGRICULTURAL CROPS CLASSIFICATION

P. B. C. Leite[a*], R.Q. Feitosa[a], A.R. Formaggio[b], G. A. O. P. Costa[a], K.Pakzad[c], , I. D. A. Sanche[b],

[a]Catholic University of Rio de Janeiro (PUC-Rio) – paula.leite@gmail.com ;{raul,gilson}@ele.puc-rio.br;
[b]National Institute for Space Research (INPE) - formag@dsr.inpe.br, iedasanches@gmail.com
[c]Leibnitz University Hannover (IPI) - pakzad@ipi.uni-hannover.de;

KEY WORDS: Crop Identification, Hidden Markov Models, Multitemporal Analysis, Object-based Image Analysis

ABSTRACT:

This work proposes a Hidden Markov Model (HMM) based technique to classify agricultural crops, exploring information of temporal image sequences from TM and ETM+/Landsat sensors. It endeavours to combine two knowledge fields, the research on plant phenology and on multitemporal object-based classification techniques. HMMs are used to relate the varying spectral response along the crop cycle with plant phenology for different crop classes. The method recognizes different agricultural crops by analyzing their spectral profiles over a sequence of medium resolution satellite images. In our approach the temporal behaviour of each crop class is modelled by a specific HMM. A segment-based classification is performed using the average spectral values of each image segment across an image sequence, which is subsequently submitted to the HMMs of each crop class. The image segment is assigned to the crop class, whose corresponding HMM delivers the highest probability of emitting the observed sequence of spectral values. Experiments were conducted upon a set of 12 co-registered and radiometrically corrected LANDSAT images. The images cover an area of the State of São Paulo, Brazil with about 124.100ha, between 2002 and 2004. The following crop classes were considered: sugarcane, soybean, corn, pasture and riparian forest. Performance assessment was carried out upon a data set classified visually by two analysts and validated by extensive field work. While in our experiments a maximum likelihood monotemporal classifier delivered in average an overall accuracy close to 25%, the HMM method was able to achieve 86%. Considering the scarcity of training samples for some crop classes in our data set, it is fair to expect even higher performances, if more representative training sets can be made available.

1. INTRODUCTION

1.1 Motivation

Given the importance of agriculture worldwide, socially and economically, the availability of precise and efficient information about agricultural activities in an appropriate time interval is highly relevant for a number of strategic decisions. Rural producers, export and import agents, companies in the food industry, suppliers, investors and the government are some of the players interested in this kind of information.

With accurate information about the status of different crops it is possible to develop commercial plans, to regulate agricultural products internal stocks, to make decisions on subsidies, and to draw strategies for the negotiation of agricultural commodities in financial markets.

This work endeavours to combine two knowledge fields that have had a noticeable evolution in recent years, namely the research on multitemporal classification techniques using satellite imagery and on plant phenology. Here lies the main novelty of the present work. In fact there are few reports on using phenological models to support the image classification process (Aurdal et al., 2005). Hidden Markov Models were used to relate the varying spectral response along the crop cycle with plant phenology for different crop classes.

Thus the general objective of this work was to evaluate the potential of Hidden Markov Models for crop classification from remote sensing temporal image sequences. Instead of relying on single date images, the methodology investigated in this work identifies different agricultural crops by analyzing the crop specific temporal profiles of spectral features over a sequence of medium resolution satellite images.

Section 2 shows the problem characterization, followed by a description of Hidden Markov Model method in section 3. The proposed methodology is presented in section 4 and a performance analysis is presented in section 5 followed by final comments.

2. PROBLEM CHARACTERIZATION

2.1 Crops and their phenological cycles

The cycles and the planting and harvesting dates of the main crops found in a study area determine the quantity of foliar area, phytomass volume and soil coverage temporal variations. The knowledge of these peculiarities gives the basis for understanding the spectral behaviors presented by the studied crop types in a certain period of the year.

2.1.1 Sugarcane: In São Paulo, Brazil, the sugarcane (SC) (Saccharum spp.) cultivation follows basically two cycles: one of 12 months ("one-year" sugarcane) and another of 18 months ("one-year-and-half" sugarcane). The one-year-and-half sugarcane is planted between January and March and the one-year sugarcane, between October and November. It is important to highlight that each sugarcane crop can be harvested during five or six consecutive agricultural cycles. For this reason the cycle is named "semi-perennial", which is different from grain

crops' cycles, because of its duration, as well as of its phenological dynamics.

For areas where this crop is recently planted, a green mass of one-year-and-half sugarcane starts to completely cover the soil from October, when there is more heat and pluviometric precipitation; however, new areas of one-year sugarcane, should have full green coverage in April and May and then the green phytomass tends to increase its foliar area until the next harvesting period.

Each year the period of harvesting starts in April and ends in November, this way, in a same date of satellite image it is possible to find: straw from harvested crop, recently planted sugarcane, as well as sugarcane in the growth phase and in the adult phase. It is also possible to find exposed soil, where the agricultural area is prepared for planting.

2.1.2 Short cycle crops (cereals): Soybean (SB) and corn (CO) are called "annual crops" or "short cycle crops", once they can complete their phenological cycle in 110 to 140 days. They are planted, in general, in the end of October or in the beginning of November and they germinate about 10 days after being planted, begging their vegetative growth and fully covering the soil surface around 60 days after the germination. In the sequence, these crops reach the peak of green phytomass and then they begin the grain filling process, when the quantity of green leaves starts to diminish while the quantity of yellow leaves increase. They then dry out and fall, exposing again the soil background until the harvesting period.

2.1.3 Pasture: Pasture (PS) presents different phenological and spectral dynamics from the other crops mentioned above. These dynamics depend on the types of soil management used by cattlemen, however, in general, pastures are more dry and scarce between April and September, when the rainy season starts along with their revigoration, which increases the foliar area index and sustain the green vegetative vigor from November to March.

2.1.4 Other classes: Besides these crops vegetation, riparian forest (RF) was also considered in this work. Other classes of land cover are present in the study area: urban areas, roads, forest and water bodies. They appear as few, large segments that practically do not change thorough all the image sequence, and for this reason, they were not included in this work.

3. HIDDEN MARKOV MODELS

A Hidden Markov Model (HMM) (Bunke & Caelli, 2001) represents a doubly embedded stochastic process. In an HMM, the observations (v_i) are regarded as symbols emitted by non observable states (S_i), following particular probabilistic functions, whereby the state sequence is a first order Markov Chain. An HMM is illustrated in Figure 1. N is the number of states in the model (the individual states are denoted as $\mathbf{S} = \{S_1,\ldots,S_N\}$, and the state at time t as q_t) and M is the number of distinct observation symbols per state (the individual symbols are denoted as $\mathbf{V} = \{v_1, \ldots, v_M\}$). A basic HMM consists of three sets of parameters:

a) the symbol emission probabilities b_{jk} – the probability that symbol v_k is emitted by state S_j, i.e.
$$b_{jk} = \mathrm{P}[v_k \text{ at } t | q_t = S_j], \quad 1 \le j \le N \text{ and } 1 \le k \le M$$

b) the state transition probabilities a_{ij} – the probability of being in state S_j in the subsequent time instant given that the current state is S_i, i.e.
$$a_{ij} = \mathrm{P}[q_{t+1} = S_j | q_t = S_i], \quad 1 \le i,j \le N$$

c) the prior probability distribution π_i that the system is in a given state S_i at the initial time instant (not shown in the figure), i.e.
$$\pi_i = \mathrm{P}[q_1 = S_i], \quad 1 \le i \le N.$$

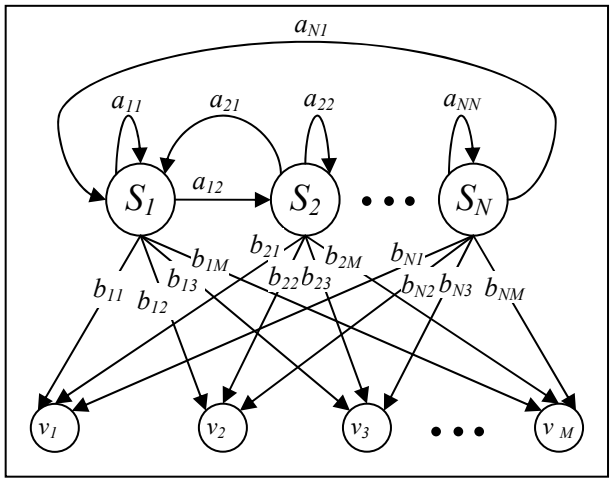

Figure 1. Example of a Hidden Markov Model ($S_i \rightarrow$ states, $v_k \rightarrow$ observation symbols, $a_{ij} \rightarrow$ state transition probability, $b_{ik} \rightarrow$ symbol emission probability).

If a state S_i can reach another state S_j, $a_{ij}>0$ and if two states are not connected, $a_{ij} = 0$.

4. METHODOLOGY

The basic HMM shown in Figure 2 was chosen to model the temporal behaviour of sugarcane, soybean and corn. The arrows illustrate how the states are temporally related. According to plant phenology, states *PH, PP, GR* and *AD* correspond to stages Post-Harvesting, Prepared Soil, Growth phase and Adult phase respectively.

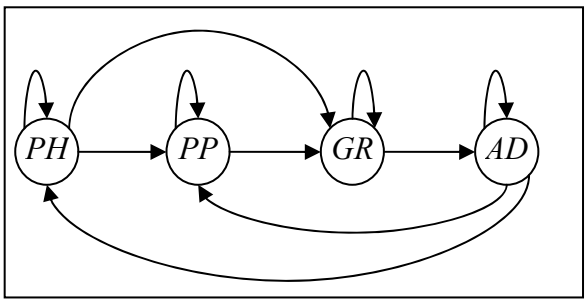

Figure 2. HMM used in this work for sugarcane, soybean and corn (*PH* = Post-harvesting, *PP* = Prepared soil, *GR* = Growth and *AD* = Adult phase).

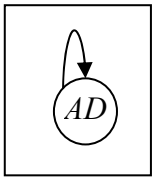

Figure 3. HMM used in this work for pasture and riparian forest (*AD* = Adult phase).

For pasture and riparian forest there is no significant change in the radiometric features throughout the phenological cycle, so a specific HMM is devised for these classes having a single state S_1, which in these cases correspond to Adult (Figure 3).

Even though pasture and riparian forest are actually not crop types, the term "crop" will be used hereafter to designate the set of all five classes to be recognized in our problem.

Considering the crops available in the study area, each one of these crops is associated with a different HMM, with different state transition probabilities and symbol emission probabilities. It is necessary to obtain such components, as well as the probability of occurrence of each state *Si* on the initial date of a sequence of observations being considered, in order to define each crop's model.

The problem being considered in this work deviates in a number of ways from the basic HMM description presented in the preceeding section. First, the symbol emission probabilities (b_{jk}) depend on seasonal effects that can not be fully compensated in the image pre-processing phase. Second, the prior probability distribution (π_i) is not constant along the year (see section 2). Third, the basic model depicted in section 3 assumes that the symbols are emitted at a constant time rate. In most real applications we don't have an usable image at a fixed time interval, mostly due to clouds at the moment when the satellite passes over the target geographical area. It is also worth mentioning that the basic model shown in Figure 2 may also change for a larger interval between two consecutive images in the data set. For instance, a transition from PP to AD may become possible in these cases.

In consequence, an HMM for our problem will have to consider distinct symbol emission probabilities, prior state probabilities, as well as state transition probability matrices for each pair of consecutive images in the available dataset.

Regarding the symbol emission probabilities, it is assumed throughout this paper that they have a Gaussian distribution. Hence the emission probability density of a symbol **x** (a vector consisting of the spectral bands and NDVI) will be given by:

$$p = \frac{1}{(2\pi)^{d/2}\left|\Sigma_{cs}\right|^{1/2}} \exp\left[-\frac{(x-\mu_{cs})^T \Sigma_{cs}^{-1}(x-\mu_{cs})}{2}\right] \quad (1)$$

where $\boldsymbol{\mu}_{cs}$, and $\boldsymbol{\Sigma}_{cs}$ denote respectively the mean vector, the covariance matrix for culture *c* and state *s,* and *d* is the dimension of **x**.

Once the HMM have been established and their parameters estimated, the classification of an image segment is done in the following way. The segment is represented at each date by a symbol vector comprising its average spectral values and NDVI observed at that date. From the symbol vectors representing the segment behaviour during a given succession of dates, the classifier computes for each model, the probability that the corresponding crop class emits the observed sequence of symbol vectors. The segment is assigned to the class whose model delivers the highest emission probability. A detailed description about how emission probabilities are computed in an HMM can be found in (Rabiner, 1989).

5. PERFORMANCE ANALYSIS

5.1 Data Set

This section shows the details of the database used in this work.

5.1.1 Study area: The study area corresponds to three cities in the State of São Paulo, Brazil: Ipuã, Guará e São Joaquim da Barra (inside a rectangle defined by the following coordinates: 20º16'30"S to 20º40'00"S x 47º37'36"W to 48º13'50"W), covering an area of 124.100ha (Figure 4). Agriculture is the main activity in this area. The main crops found are: sugarcane, soybeans and corn. This region has a plane to slightly undulated relief, a tropical climate with dry winter, with annual mean temperature of 22,9ºC and annual mean precipitation of 1480mm.

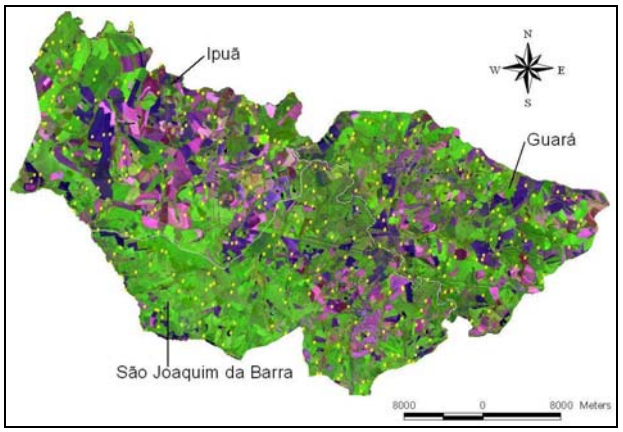

Figure 4. Study area in state of São Paulo, Brazil

5.1.2 Image Sequence: The dataset contains a total of 12 images from the Landsat satellite, orbit/point WRS 220/74, from 2002 to 2004 (Table 1), from TM/Landsat-5, as well as from ETM+/Landsat-7 sensors (Sanches, 2004). Bands 1 to 5 and 7 were used in this work. Only the thermal band was not used.

	2002	2003	2004
January		(ETM+) 08/01/03	(TM) 19/01/04
February		(ETM+) 09/02/03 - 25/02/03	
April		(ETM+) 14/04/03 - 30/04/03	
May		(ETM+) 16/05/03	
July		(TM) 27/07/03	
August		(TM) 12/08/03	
September	(ETM+) 02/09/02		
October	(ETM+) 20/10/02	(TM) 15/10/03	

Table 1. Images available

5.1.3 Image pre-processing: The Landsat images were in geotiff format and for the geometric corrections, 13 control points gathered by GPS were used. The nearest neighbour resampling method was applied, considering that it well preserves the original image's radiometry (Mather, 1993; Richards, 1995).

A correction was applied to the multitemporal images to diminish atmospheric effects, once that the atmosphere, by its spread-spectrum, absorption and refraction phenomena, affects the radiance measured by the orbital sensors. The Dark-object

subtraction technique, developed by Chavez (1988), was applied.

As the same object may present distinct digital values in different acquisition dates' images, due to difference in the solar angles and to spread-spectrum effect, multitemporal groups of images must be radiometrically normalized. In this work, this process was done according to the methodology proposed by Gürtler et al. (2003).

Classification algorithms are based on the spectral appearance of the objects being classified in images from different dates, so the grayscale values were converted to reflectance values, which have a physical meaning, in order to correctly represent the different objects and their conditions at the images' acquisition moments. This conversion was based on the methodology proposed by Luiz et al. (2003).

5.1.4 Image Segmentation and Attributes: After gathering all the images available, they were stacked up and segmented. A watershed based technique was applied, which is presented in details in (Mota et al., 2007). The average spectral values of eachwere are measured across an image sequence and, subsequently, a seventh attribute was generated, the NDVI (Normalized Difference Vegetation Index).

5.1.5 Reference Data: The agricultural vegetations considered in this work were the main ones found in the study area: sugarcane, soybeans, corn, pasture and riparian forest.

For soybeans, corn and sugarcane, the phenological-spectral cycle was divided in four phases: Prepared soil (when the surface appears as exposed soil in the satellite images), Growth phase (when the crop fully covers the soil), Adult phase (when crops are in maximum green vegetative vigor and may be beginning their senescence period) and Post-Harvesting (when areas, where there were crops before, are covered with dry straw remains after harvesting).

As mentioned before, the cycles of pasture and riparian forest were represented by a single state.

There were 316 reference segments selected in the study area, and each one of them was visually classified by two experts, considering the acquisition dates and according to the classes indicated above. This classification was validated by field works conducted in March and August of 2003 respectively.

5.1.6 Training Procedure: The training procedure consists in estimating the symbol emission probability, as well as the state transition probability and the prior probability distribution.

The value returned by equation (1) was used in place of the symbol emission probability as these values are proportional. Hence, the problem of estimating symbol emission probabilities turned into the estimation of the sample mean and covariance matrix for each crop type and state.

After isolating the samples of one crop, in a given date, the proportion between the occurrence of one phenological stage and all the others was calculated. The prior probability distribution was fully defined after having calculated such proportions for all phenological stages and all the crops, in each date.

At last, the state transition probability was calculated considering pairs of consecutive dates. To calculate the transition probability of state *i* to state *j*, the proportion between the occurrence of this particular transition and the transitions of state *i* to all the others was calculated. This was done for each crop and each transition possible in all the pairs of consecutive dates.

The need to provide model parameter estimates for each date (see section 4) brought about a considerable demand for training samples, which in some cases could not be met by the available data set. This was especially critical for the estimation of the covariance matrices (equation 1). To cope with this problem some strategies were applied, namely:
Prior-knowledge: To estimate prior state probabilities, the number of possible states with no sample in the training set was set to 1; this guaranteed a non zero probability for all possible states. The information about what are the possible states for each crop type and for each date was treated as prior-knowledge. A similar strategy was applied to estimate state transition probabilities.
Leave-one-out: all sequences in the data set excluding the one being classified was used to estimate the model parameters; this procedure was repeated for each tested sequence in the data set.
Dimensionality reduction: principal component analysis was applied to reduce dimensionality, and consequently the demand for training samples.
Linear Regression: in cases where, despite the aforementioned strategies, available training samples were still insufficient, linear regression was applied to provide estimates based on samples from a different date.

5.2 Experiment Results

This experiment aims at identifying crop types, as well as the phenological stages during the dates in the test sequences.

The sequence used to test the classifier was not used for training. A "leave-one-out", as well as the strategies briefly described in section 5.1.6 to deal with scarce training sets, were applied.

Only complete sequences were used here, meaning that they had all the phenological stages represented. Additionally, there was only one crop type per sequence.

Finally, a maximum likelihood single-date classifier was applied for comparison.

Table 2 and 3 show the accuracies and the confusion matrix for crop class classification respectively. Table 4 and 5 refer to stage identification, considering again only sequences correctly identified by the HMM classification model.

Class Accuracy (crops)	
Crops	**Rates (%)**
Soybeans (SB)	96
Corn (CO)	47
Sugarcane (SC)	90
Pasture (PS)	76
Riparian forest (RF)	75
Overall accuracy:	**86**
Average class accuracy:	**77**

Table 2. Crop classification accuracy.

Confusion matrix (crops)					
	SB	CO	SC	PS	RF
SB	96	0	4	0	0
CO	5	14	5	4	2
SC	8	1	179	11	0
PS	1	0	1	19	4
RF	1	0	4	3	24

Table 3. Crop classification confusion matrix.

Class Accuracy (states)	
States	**Rates (%)**
Post-harvesting (PH)	78
Prepared soil (PP)	84
Growth phase (GR)	38
Adult phase (AD)	94
Overall accuracy:	**84**
Average class accuracy:	**74**

Table 4. State classification accuracy.

Confusion matrix (states)				
	PH	PP	GR	AD
PH	167	31	5	11
PP	16	431	32	32
GR	4	79	137	139
AD	7	31	67	1678

Table 5. State classification confusion matrix.

The tables show high overall and average class accuracy for both crop and phenological stage classification.

Table 2 shows that corn crops had the lowest value for class accuracy. This can be explained by the scarce data available for training. When leaving one of the sequences out for testing, for some dates, the only sample of this culture was taken out, making it hard to estimate the model parameters. It is important to highlight that this is a problem with the data available and not with the method itself.

When looking at the confusion matrix shown in Table 3, one may be mislead to think that the aforementioned problem of missing samples in some dates for corn crops should also affect pasture and riparian forest as they have approximately the same number of samples. Recall that these two crops are represented by single-state models, meaning that there are fewer parameters to be estimated and thus, fewer samples needed.

The phenological stages were also well identified, in exception of the Growth phase (Table 4). This can be explained by the temporal evolution of the crops throughout the phenological cycle. During the prepared soil, adult and post-harvesting phases, there is no significant changes in the crop's spectral response. However, the spectral response of the growth phase is continuously changing from prepared soil to post-harvesting. So its spectral response could be close to the response of these other two stages, or something in between, which can lead to misclassification.

Table 5 confirms this interpretation, as the confusion matrix shows that the growth stage was often misclassified as adult phase and prepared soil.

For this particular data set, which had scarce data for some crops, the number of attributes used in the classification is highly influent on the results. For example, when using all 7 attributes available, the accuracy for corn (crop with the least number of samples) was much (from 13% to 47%) worse than when applying PCA to reduce the dimension to 3 attributes.

A maximum likelihood single-date classification method was performed for comparison purposes. The experiment was only concerned about the crop type classification. The samples on each date were randomly separated, half for training and half for testing. We discarded from this analysis the dates in which there were insufficient training samples. To make it a more realistic experiment, the information of all the crops in the prepared soil stage was discarded as, in this stage, the spectral response is basically deternined by the bare soil, once the crop is still not aparent. PCA was not applied here. The average class accuracy of all valid dates was 25%.This is certainly not a surprising result, once one single crop class may have at the same date quite distinct spectral responses depending on their phenological stage. Nevertheless, the poor performance observerd in this single-date classification emphasizes the convenience of using a multi-date approach, as the HMM method proposed in this work.

6. FINAL COMENTS

This work evaluated the potential of Hidden Markov Models for crop classification. The experimental evaluation based on sequence of 12 Landsat images for 5 crop types indicated that a remarkable superiority of the HMM-based method, over a monotemporal maximum likelihood classification approach.

An analysis of the experimental results revealed that the performance of HMM-based classifier was severely impacted by the scarcity of training samples of some crop types. Hence, even better results could have been achieved if a more representative training set were available.

The HMM approach also performed well to recognise the phonological stages. Exception was the growth phase, which were frequently confused with prepared-soil and adult phase. This observation suggests that symbol vectors used to characterize the growth-phase should take into account not only the absolute spectral values but also their variation along the time.

For this work, only sequences with one crop type were considered. It would be interesting to test, in future works, the behaviour of the method with sequences containing samples of more than one crop type.

7. ACKNOWLEDGEMENTS

The authors acknowledge CNPq (National Counsel of Technological and Scientific Development) and DLR (German Aerospace Center) for supporting this research.

8. REFERENCES

Aurdal, L.; Huseby, R. B.; Eikvil, L.; Solberg, R. Vikhamar, D. Solberg, A., 2005. Use of hidden Markov models and phenology for multitemporal satellite image classification: applications to mountain vegetation classification; *2005 International Workshop on the Analysis of Multi-Temporal Remote Sensing Images*, pp.220-224.

Chavez Jr., P.S., 1988. An improved dark-object subtraction technique for atmospheric scattering correction of multispectral data. Remote Sensing of Environment, Vol.24, n.9, pp. 459-479.

Gürtler, S., 2003. "Estimativa da área agrícola a partir de sensoriamento remoto e banco de pixels amostrais". Dissertação de Mestrado. 179 p. sid.inpe.br/jeferson/2003/06.02.07.29 (accessed june 2003.) (INPE-9774_TDI/858).

Bunkle, H.; Caelli, T., 2001. Hidden Markov Models – applications in computer vision, *World Scientific*

Lillesand, T. M.; Kiefer, R.W., 1994. *Remote Sensing and Image Interpretation*. 3. ed. New York: John Wiley & Sons. 750 p.

Luiz, A.J.B.; Gürtler, S.; Gleriani, J.M.; Epiphanio, J.C.N.; Campos, R.C., 2003. Reflectância a partir do número digital de imagens ETM. *Simpósio Brasileiro de Sensoriamento Remoto*, 11., Belo Horizonte. Anais. São José dos Campos, Brasil: INPE, pp. 2071-2078.

Mather, P.M. *Computer processing of remotely-sensed images: an introduction*, 1993. 3.ed. Chichester: John Wiley & Sons, 352 p.

Mota, G. L. A., Feitosa, R. Q., Coutinho, H. L. C., Liedtke, C-E., Müller, S., Pakzad, K., Meirelles, M. S. P., 2007. Multitemporal fuzzy classification model based on class transition possibilities. *ISPRS Journal of Photogrammetry and Remote Sensing*. Vol.1, pp.1-2.

Rabiner, L. R., 1989. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE*, Vol. 77, n.2, pp. 257-286.

Richards, J.A., 1995. *Remote sensing digital image analysis: an introduction*. 3. ed. Berlin:Springer-Verlag,. 340 p.