sid.inpe.br/mtc-m19/2013/09.17.18.58-TDI

SPECTRAL-TEMPORAL AND BAYESIAN METHODS FOR AGRICULTURAL REMOTE SENSING DATA ANALYSIS

Marcio Pupin Mello

Doctorate Thesis Course Graduate in Remote Sensing, guided by Dr. Bernardo Friedrich Theodor Rudorff, approved in August 19, 2013.

URL of the original document:
<http://urlib.net/8JMKD3MGP7W/3ERM89S>

INPE
São José dos Campos
2013

PUBLISHED BY:

Instituto Nacional de Pesquisas Espaciais - INPE
Gabinete do Diretor (GB)
Serviço de Informação e Documentação (SID)
Caixa Postal 515 - CEP 12.245-970
São José dos Campos - SP - Brasil
Tel.:(012) 3208-6923/6921
Fax: (012) 3208-6919
E-mail: pubtc@sid.inpe.br

sid.inpe.br/mtc-m19/2013/09.17.18.58-TDI

SPECTRAL-TEMPORAL AND BAYESIAN METHODS FOR AGRICULTURAL REMOTE SENSING DATA ANALYSIS

Marcio Pupin Mello

Doctorate Thesis Course Graduate in Remote Sensing, guided by Dr. Bernardo Friedrich Theodor Rudorff, approved in August 19, 2013.

URL of the original document:
<http://urlib.net/8JMKD3MGP7W/3ERM89S>

INPE
São José dos Campos
2013

Cataloging in Publication Data

Mello, Marcio Pupin.

M489s Spectral-Temporal and Bayesian Methods for agricultural Remote Sensing Data Analysis / Marcio Pupin Mello. – São José dos Campos : INPE, 2013.

xxvi + 94 p. ; (sid.inpe.br/mtc-m19/2013/09.17.18.58-TDI)

Thesis (Doctorate in Remote Sensing) – Instituto Nacional de Pesquisas Espaciais, São José dos Campos, 2013.

Guiding : Dr. Bernardo Friedrich Theodor Rudorff.

1. spectral-temporal response surface 2. sugarcane pre-harvest burning 3. Bayesian Network 4. plausible reasoning 5. soybean mapping I.Título.

CDU 528.8:631

Aprovado (a) pela Banca Examinadora em cumprimento ao requisito exigido para obtenção do Título de *Doutor(a)* em

Sensoriamento Remoto

Dr. Antonio Roberto Formaggio

Formaggi

Presidente / INPE / SJCampos - SP

Dr. Bernardo Friedrich Theodor Rudorff

Orientador(a) / INPE / SJCampos - SP

Dr. Rafael Duarte Coelho dos Santos

Membro da Banca / INPE / SJCampos - SP

Dr. Getúlio Teixeira Batista

Convidado(a) / UNITAU / Taubaté - SP

Dr. Carlos Antônio Oliveira Vieira

Carlos Antonio O. Vieira

Convidado(a) / UFSC / Florianópolis - SC

Este trabalho foi aprovado por:

( ) maioria simples

(X) unanimidade

Aluno (a): ***Márcio Pupin de Mello***

São José dos Campos, 19 de Agosto de 2013

“At the moment of commitment, the Universe conspires to assist you.”

Johann Goethe

“Quando uma criatura humana desperta para um grande sonho e sobre ele lança toda a força de sua alma, todo o universo conspira a seu favor.”

Johann Goethe

(In Portuguese)

Dedico ao homem que me ensinou a lutar pelos meus sonhos,
que dedicou seu suor ao meu sustento,
que me mostrou o verdadeiro significado de coragem,
ao homem que me escolheu para ter a honra de tê-lo como pai.

ACKNOWLEDGEMENTS

In Portuguese:

Desde a pré-escola, ensino fundamental e médio, (também o cursinho pré-vestibular) universidade, mestrado e doutorado, sempre estudei em instituições públicas. Portanto, agradeço a todos os cidadãos brasileiros, que com o dinheiro de seus impostos permitiram que eu tivesse acesso a uma educação de qualidade, sem muitas vezes poder sequer proporcionar o mesmo aos próprios filhos.

Ao mestre Leandro Coelho Dalvi, que respondia com paciência às tolas perguntas de um adolescente curioso. Obrigado, meu amigo, por ter sido o catalizador no propulsor que tem me impulsionado nessa prazerosa viagem pelos belos mistérios da ciência.

Aos amigos Drs. Luiz Peternelli e Carlos Vieira. Certa vez, Sir Isaac Newton disse: “Se enxerguei longe, foi porque me apoiei nos ombros de gigantes”. Faço destas minhas palavras.

Ao meu orientador e amigo, Dr. Bernardo. Professor, sou grato por cada ensinamento, por me dar a liberdade necessária para descobrir quais caminhos trilhar, sempre à luz de sua serena e sabia orientação.

Ao Dr. Edzer Pebesma, do Instituto de Geoinformática da Universidade de Münster por ter me recebido de braços abertos durante meu doutorado sanduíche na Alemanha. Também agradeço aos amigos que fiz por lá, em especial ao Chris e ao Jim.

Aos Drs. Getúlio Batista, Antonio Formaggio, Carlos Vieira e Rafael Santos pela honra de ter uma banca de defesa de tese com tão alto padrão. Aproveito para agradecer pelas contribuições à versão final deste documento.

Aos pesquisadores do INPE (não poderia nomear todos aqui) que mesmo nas conversas pelos corredores ou durante um cafezinho na cantina sempre tinham algo a ensinar. Extendo meus agradecimentos também aos diversos colegas da ciência que tive a oportunidade de conhecer e trocar ideias pelo mundo a fora.

Aos amigos e amigas que fiz durante esses anos no INPE (também não posso nomeá-los aqui), desde funcionários, bolsistas, colegas de pós e amigos do futebol. Cada um teve sua parcela de contribuição nesta longa etapa.

A todos os meus familiares e amigos pelo apoio e carinho. Em especial aos meus pais José Antero e Angela; minhas irmãs Aline, Mayara e Lara; meu sobrinho e afilhado Nícolas; e à minha Bianka... Tudo por vocês – todo o esforço e recompensa, sempre: tudo por vocês!

Ao apoio financeiro do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq –142845/2011-6); da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES – 33010013005P0); da Fundação de Amparo à Pesquisa do Estado de São Paulo (Fapesp – 2008/56252-0); e do Serviço Alemão de Intercâmbio Acadêmico (*Deutscher Akademischer Austausch Dienst* – DAAD – 12/A/72898).

A todos e todas que direta ou indiretamente contribuíram.

ABSTRACT

Reliable agricultural statistics has become increasingly important to decision makers. Especially when timely obtained, agricultural information is highly relevant to the strategic planning of the country. Although remote sensing shows to be of great potential for agricultural mapping applications, with the benefit of further improving official agricultural statistics, its potential has not been fully explored. There are very few successful examples of operational remote sensing application for systematic mapping of agricultural crops, and they are strongly supported by visual image interpretation to allow accurate results. Indeed, despite the substantial advances in remote sensing data analysis, techniques to automate remote sensing data analysis focusing on agricultural mapping applications are highly valuable but have to maintain consistency and accuracy. In this context, there continues to be a demand for development and implementation of computer aided methods to automate the processes of analyzing remote sensing datasets for agriculture applications. Thus, the main objective of this thesis is to propose implementation of computer aided methodologies to automate, maintaining consistency and accuracy, processes of remote sensing data analyses focused on agricultural thematic mapping applications. This thesis was written as a collection of two papers related to a core theme, each addressing the following main points: (i) multitemporal, multispectral and multisensor image analysis that allow the description of spectral changes of agricultural targets over time; and (ii) artificial intelligence in modeling phenomena using remote sensing and ancillary data. Study cases of sugarcane harvest in São Paulo and soybean mapping in Mato Grosso were used to test the proposed methods named STARS and BayNeRD, respectively. The two methods developed and tested confirm that remotely sensed (and ancillary) data analysis can be automated with computer aided methods to model a range of cropland phenomena for agriculture applications, maintaining consistency and accuracy.

MÉTODOS SPECTRO-TEMPORAL E BAYESIANO PARA ANÁLISE DE DADOS EM SENSORIAMENTO REMOTO AGRÍCOLA

RESUMO

Informações agrícolas confiáveis tem se tornado cada vez mais importantes para os tomadores de decisões. Especialmente quando são obtidas em tempo hábil, essas informações são altamente relevantes para o planejamento estratégico do país. Apesar de o sensoriamento remoto mostrar-se promissor para aplicações em mapeamento agrícola, com potencial de melhorar as estatísticas agrícolas oficiais, esse potencial não tem sido amplamente explorado. Existem poucos exemplos bem sucedidos do uso operacional do sensoriamento remoto para mapeamento sistemático de culturas agrícolas e, para garantir resultados precisos, eles são fortemente baseados em interpretação visual de imagens. De fato, apesar dos substanciais avanços em análise de dados de sensoriamento remoto, novas técnicas para automatizar a análise de dados em sensoriamento remoto com aplicações agrícolas são desejáveis, especialmente no propósito de manter a consistência e a precisão dos resultados. Neste contexto, existe uma demanda crescente pelo desenvolvimento e implementação de métodos automatizados de análise de dados de sensoriamento remoto com aplicações em agricultura. Assim, o principal objetivo desta tese é propor o desenvolvimento e a implementação de métodos para automatizar a análise de dados de sensoriamento remoto em aplicações agrícolas, com foco na consistência e precisão dos resultados. Este documento foi escrito como uma coleção de dois artigos, cada um com foco nos seguintes pontos: (i) análise multitemporal, multiespectral e multisensor, permitindo a descrição das variações espectrais de alvos agrícolas ao longo do tempo; e (ii) inteligência artificial na modelagem de fenômenos usando dados de sensoriamento remoto e informações complementares de maneira integrada. Dois estudos de caso referentes ao mapeamento da colheita da cana em São Paulo e ao mapeamento da soja no Mato Grosso foram usados para testar as metodologias batizadas de STARS e BayNeRD, respectivamente. Os resultados dos testes confirmaram que ambos os métodos propostos foram capazes de automatizar processos de análises de dados de sensoriamento remoto com aplicações agrícolas, com consistência e precisão.

LIST OF FIGURES

LIST OF TABLES

LIST OF ABREVIATIONS

6S –	Second Simulation of the Satellite Signal in the Solar Spectrum
ANEEL –	*Agência Nacional de Energia Elétrica*
BayNeRD –	Bayesian Network for Raster Data
BH –	Burned Harvest
BN –	Bayesian Network
CEI –	Crop Enhanced Index
CS –	Collocation Surface
DAG –	Directed Acyclic Graph
ETM+ –	Enhanced Thematic Mapper Plus
EVI –	Enhanced Vegetation Index
FAO –	Food and Agriculture Organization of the United Nations
FUNAI –	*Fundação Nacional do Índio*
GeoDMA –	Geographic Data Mining Analyst
GH –	Green Harvest
GHG –	Greenhouse Gas
IB7 –	Instance-Based (classifier used with seven nearest neighbors)
IBGE –	*Instituto Brasileiro de Geografia e Estatística*
IR-MAD –	Iteratively Reweighted Multivariate Alteration Detection
J48 –	[refers to the J4.8 implementation of the algorithm used as decision tree classifier]
KL –	Kullback-Leibler
MAPA –	*Ministério do Meio Ambiente*
MCI –	Multi-Coefficient Image
MGS –	Modified Gram-Schmidt
MLP20 –	Multilayer Perceptron (classifier used with 20 neurons in the hidden layer)
MMA –	*Ministério do Meio Ambiente*
MODIS –	Moderate Resolution Imaging Spectroradiometer
NDVI –	Normalized Difference Vegetation Index
NIR –	Near Infrared
OBIA –	Object-Based Image Analysis
PAM –	*Produção Agrícola Municipal*
PCA –	Principal Component Analysis
PI –	Probability Image
PTS –	Polynomial Trend Surface
RBMC –	*Rede Brasileira de Monitoramento Contínuo*

ROC –	Receiver Operating Characteristic
SEPLAN-MT –	*Secretaria de Estado de Planejamento e Coordenação Geral do Mato Grosso*
SPOT –	*Système Pour l’Observacion de la Terre*
SRTM –	Shuttle Radar Topography Mission
STARS –	Spectral-Temporal Analysis by Response Surface
STS –	Spectral-Temporal Space
SWIR –	Short-Wave Infrared
TM –	Thematic Mapper
TPV –	Target Probability Value
UN –	Unharvested

CONTENTS

Page

1 Introduction

Reliable agricultural statistics has become increasingly important to decision makers. Especially when timely obtained, agricultural information is highly relevant to the strategic planning of the country (e.g., inventory control, pricing, etc.) (PINO, 1999). Until 1938, the official agricultural statistics of Brazil were the sole responsibility of the Ministry of Agriculture, Livestock and Food Supply (MAPA – *Ministério da Agricultura, Pecuária e Abastecimento*). Later, this responsibility was shared with the Brazilian Institute of Geography and Statistics (IBGE – *Instituto Brasileiro de Geografia e Estatística*). From 1938 until the 1970s, several methodologies were applied by MAPA and IBGE to estimate agricultural statistics under the responsibility of public agencies (IBGE, 2002). In January 1974, IBGE was decreed the official agency for agricultural statistics in Brazil. However, these statistics have been estimated using methods based on subjective techniques. According to IBGE (2002), the estimates are based on questionnaires distributed to producers or to regional representatives of the agricultural sector. Despite of the relevance of these estimates, two aspects shall be pointed out about the data produced by IBGE: (i) the estimates are carried out based on a subjective method, therefore, it is not possible to statistically treat the errors, and (ii) the municipality estimates (*Produção Agrícola Municipal* – PAM) are published with a time lag of about two years (BATISTA *et al.*, 1978; IBGE, 2012b).

A significant improvement in the quality of satellite imagery was observed in 1984 with the advent of the Thematic Mapper (TM) sensor aboard the Landsat-5 satellite, widespreading the use of satellite images to map agricultural areas (NELLIS *et al.*, 2009). Another important event that increased the use of satellite images for agricultural applications was the Moderate Resolution Imaging Spectroradiometer (MODIS) sensor aboard the Terra satellite launched in late 1999 (BECKER-RESHEF *et al.*, 2010; RUDORFF, B. F. T. *et al.*, 2007). Although the moderate spatial resolution of 250 m of the MODIS sensor restricts its use to areas of extensive farming (crops planted in large fields) (RUDORFF, C. M. *et al.*, 2007), it is compensated by a wide imaging swath allowing an almost dayly temporal resolution (PITARCH *et al.*, 2011). Indeed, the Earth observing imagery counts on a wide range of sensors with different characteristics, acquiring a huge amount of data with potential use for different applications. Thus, due to its synoptic and systematic characteristic (JENSEN, 2006), orbital remote sensing has been pointed out as a valuable tool for mapping and monitoring agricultural crops.

Although remote sensing has great potential for agricultural mapping applications, with the benefit of further improving official agricultural statistics in Brazil (MELLO e*t al.*, 2013a), its potential has not been widely explored for this purpose. There are very few successful examples of operational use of remote sensing for systematic mapping of agricultural crops (ATZBERGER, 2013). Among those few examples, we highlight the Canasat and the Soy Moratorium projects, developed by INPE in partnerships with the private and public institutions.

Since 2003, the Canasat Project mapped the sugarcane crop in the south-central region of Brazil (São Paulo, Paraná, Minas Gerais, Mato Grosso, Mato Grosso do Sul, Goiás, Rio de Janeiro and Espírito Santo States) (RUDORFF *et al.*, 2010). Moreover, since 2006, the Canasat Project mapped the type of sugarcane harvest (i.e., with or without the straw burning during preharvest) in São Paulo State (AGUIAR *et al.*, 2011). Furthermore, the Canasat Project also developed complementary research on topics such as understanding the dynamics of land use change in response to the expansion of sugarcane (ADAMI *et al.*, 2012b). The Soy Moratorium Project, in its turn, incorporated the use of satellite imagery in 2009 to identify annual crops in deforested land after July 24, 2006, followed by air survey to identify soybean plantations among the pre-selected annual crops (RUDORFF *et al.*, 2011, 2012).

An important point to be highlighted is that the two previously mentioned projects are strongly supported by visual image interpretation to allow accurate results (ADAMI *et al.*, 2012a). Hence, techniques to automate the remote sensing data analysis (LU; WENG, 2007) focusing on agricultural mapping applications are highly desirable (MELLO, 2009; VIEIRA, 2000) but have to maintain consistency and accuracy (LOVELAND *et al.*, 2002).

Lu and Weng (2007) made a literature review on the subject of image classification methods and the progress made in terms of improving the classification results. Among the main features listed by the authors, advances in terms of automating processes for remote sensing data analysis converge around: (i) the development and use of advanced algorithms for classification, especially those that incorporate the expert's knowledge; (ii) the use of multispectral, multitemporal and multisensor information; and (iii) the use of ancillary data (e.g., topography, soil, tabular data, etc.) to complement data collected by the sensors.

In recent decades, the literature offers some cases of new classification techniques, focusing on agricultural applications. Santos *et al.* (2012), for example, proposed a classification method that integrates the result of different classifiers and, according to the authors, achieved more than 80% of overall accuracy for classifications of coffee plantations in mountain areas of Minas Gerais State, Brazil. The technique of combining results from different classifiers had already been reviewed and detailed by Vieira (2000), who achieved an improvement of almost 20% in the value of the kappa index (COHEN, 1960) for a particular study of crop classification in England.

In addition, the combination of different techniques has also proved to be effective for classification. Vieira *et al.* (2012), for example, proposed the integrated use of data mining and object based image analysis (OBIA) to identify, with 94% overall accuracy, sugarcane areas in three municipalities of São Paulo State. In fact, there is a role for the use and development of new tools for OBIA. The Geographic Data Mining Analyst (GeoDMA), described by Körting *et al.* (2013), is a successful example of implementing an integrating set of tools.

However, despite the substantial advances in developing new classifiers [e.g., the Support Vector Machine (MOUNTRAKIS *et al.*, 2011)] and new approaches of automated methodologies for remote sensing data analyses [e.g., combination of classifiers (SANTOS *et al.*, 2012)], there is still a need for the development of robust methods (WILKINSON, 2005) focused on image classification for remote sensing applications in agriculture (ATZBERGER, 2013; VIEIRA, 2000).

In this context, there continues to be a demand for development and implementation of computer aided methods to automate the processes of analyzing remote sensing datasets for agriculture applications. Thus, this thesis proposes the implementation of methodologies to automate processes of remote sensing data analyses focused on agricultural thematic mapping applications.

1.1. Objective, Thesis Structure and Content

The main objective of this thesis is to automate, maintaining consistency and accuracy, processes of remote sensing data analysis with emphasis on thematic mapping of agricultural applications. Two main points were addressed:

a) Multitemporal, multispectral and multisensor image analysis that allow the description of spectral changes of agricultural targets over time;

b) Artificial intelligence in modelling phenomena using remote sensing and ancillary data.

The working hypothesis was that processes of remotely sensed data analysis focused on crop mapping can be automated with computer aided methods and produce highly accurate maps.

This thesis was written as a collection of two papers related to a core theme. Each paper describes the two aforementioned points. A brief description of the structure of each paper follows.

Chapter 2: This chapter aims at describing the development and implementation of a method to synthesize the full information content of a multispectral-multitemporal remote sensing dataset into a single synthetic image. It presents the full mathematical structure and conceptual definitions of the method named Spectral-Temporal Analysis by Response Surface (STARS). A case study was used to rigorously assess the STARS method, evaluating its potential to accurately characterize the sugarcane harvest practices in Brazil.

Chapter 3: In this chapter we proposed an innovative method to integrate remote sensing and ancillary data analysis in a logical perspective. It consists on the application of Bayesian theory using an artificial intelligence technique known as Bayesian Networks. The chapter aims at describing the development and implementation of this new method named Bayesian Networks for Raster

Data (BayNeRD). The method was used to model soybean plantations in Mato Grosso State, Brazil, based on vegetation indices, soil maps, roads network, topography and hydrography data stored in raster format.

2 STARS: a new method for multitemporal remote sensing[1]

Abstract:

There is great potential for the development of remote sensing methods that integrate and exploit both multispectral and multitemporal information. This paper presents a new image processing method: Spectral-Temporal Analysis by Response Surface (STARS), which synthesizes the full information content of a multitemporal-multispectral remote sensing image data set to represent the spectral variation over time of features on the Earth's surface. Depending on the application, STARS can be effectively implemented using a range of different models [e.g., polynomial trend surface (PTS) and collocation surface (CS)], exploiting data from different sensors, with varying spectral wavebands and acquiring data at irregular time intervals. A case study was used to test STARS, evaluating its potential to characterize sugarcane harvest practices in Brazil, specifically with and without preharvest straw burning. Although the CS model presented sharper and more defined spectral-temporal surfaces, abrupt changes related to the sugarcane harvest event were also well characterized with the PTS model when a suitable degree was set. Orthonormal coefficients were tested for both the PTS and CS models and performed more accurately than regular coefficients when used as input for three evaluated classifiers: instance-based, decision-tree, and neural network. Results show that STARS holds considerable potential for representing the spectral changes over time of features on the Earth's surface, thus becoming an effective image processing method, which is useful not only for classification purposes but also for other applications such as understanding land-cover change. The STARS algorithm can be found at www.dsr.inpe.br/~mello.

[1] This chapter is an adapted version of the paper:
Mello, M. P.; Vieira, C. A. O.; Rudorff, B. F. T.; Aplin, P.; Santos, R. D. C.; Aguiar, D. A. STARS: a new method for multitemporal remote sensing. **IEEE Transactions on Geoscience and Remote Sensing**, v. 51, n. 4, p. 1897-1913, 2013.

2.1. Introduction to STARS

Spaceborne remote sensing is widely used to monitor land-cover change on the Earth's surface. However, due to the complexity of land-cover dynamics, it is difficult to establish patterns that can be standardized to represent and map such change (DeFRIES; BELWARD, 2000; LAMBIN; GEIST, 2006). By the 1990s, the scientific community had recognized the value of remote sensing as the chief source of spatial data for driving wide-area analysis (SELLERS *et al.*, 1995). The key characteristics of satellite sensor images are that they are being continuously recorded at specific spectral wavebands over the entire Earth and can facilitate observation of environmental change at local to global scales (APLIN, 2006). As pointed out by DeFries and Belward (2000), the continuity of spaceborne remote sensing observations is a key factor for the success of using these data for characterizing change on the Earth's surface.

A wide range of studies have been conducted over the last decade to improve spectral (DEMIR *et al.*, 2011b; LANDGREBE, 2005) and temporal (BOVOLO *et al.*, 2012; DEMIR *et al.*, 2011a; SMITS; BRUZZONE, 2004) analysis and comprehension of remotely sensed data related to changes on the Earth's surface. However, these studies are often limited spectrally or temporally, either by constraining examination of image spectral profile (LANDGREBE, 2005) (i.e., multispectral analysis) to only a single date image (LEE; ERSOY, 2007; SOUTH *et al.*, 2004) or by constraining examination of image temporal profile (i.e., multitemporal analysis) to only a single spectral layer (e.g., a vegetation index time series) (GALFORD *et al.*, 2008; LUNETTA *et al.*, 2006; SALMON *et al.*, 2011; WARDLOW *et al.*, 2007). Wilkinson (2005) suggests that satellite sensor image classification results have not significantly improved for a considerable period of time. Moreover, relatively few integrated multispectral-multitemporal approaches have been reported in the scientific literature (e.g., BRUZZONE; SMITS,2002; CARRÃO *et al.* 2008). Thus, there is great potential for the development of remote sensing methods that integrate and exploit both multispectral and multitemporal information (COPPIN *et al.*, 2004).

Novel multispectral-multitemporal methods are likely to be of particular benefit where they are sufficiently robust and adaptable to be used in a range of applications, such as land-cover inventorying [e.g., change detection (LAMBIN; LINDERMAN, 2006;

LAMBIN; STRAHLER, 1994)], environmental monitoring [e.g., deforestation (SILVA *et al.*, 2008)], or resource management [e.g., maximizing agricultural productivity (BARGIEL; HERRMANN, 2011)], and in a range of different circumstances. For instance, if we are interested in monitoring agricultural crops over the growing season, it may be desirable to take into account the gradual spectral change of each crop (VIEIRA, 2000). In contrast, if we are interested in detecting harvest, it may be desirable to consider the abrupt spectral change that occurs at the time the crop is harvested (MELLO, 2009). It may be also desirable to constrain data dimensionality to avoid both high computational costs and the Hughes phenomenon (HUGHES, 1968).

This paper presents an advanced image processing method to represent the spectral-temporal behavior of features on the Earth's surface: Spectral-Temporal Analysis by Response Surface (STARS). STARS uses the concept of response surfaces for spectral-temporal analyses of multitemporal-multispectral remote sensing data (VIEIRA, 2000). It allows the use of image data from different sensors with varying spectral wavebands and irregular time intervals. Moreover, different model options can be used to fit the response surfaces according to the application.

This work draws on earlier tests using response surfaces to map agriculture fields (e.g., Epiphanio *et al.* (2010)), although these tests were limited to classification analysis. This new work presents the full mathematical structure of STARS and its conceptual definitions and treats STARS as a generic image processing method that can be used not only for classification but also for other applications such as understanding land-cover change. Within this context, a case study was used to test the STARS method, evaluating its potential to characterize sugarcane harvest practices in Brazil. In the next section, the STARS methodology is described in full. Then, in Section 2.3, Brazilian sugarcane agriculture is introduced. This is followed, in Section 2.4, by an outline of the research materials and methods employed in the application of STARS for the sugarcane harvest case study. In Section 2.5, the results of STARS and subsequent classification of the STARS outputs are presented, discussed, and rigorously assessed in terms of accuracy. This leads to final concluding comments in Section 2.6.

2.2. STARS methodology

2.2.1. Rationale

The STARS method operates by representing the full information content of a multitemporal-multispectral remote sensing image data set as a single synthetic multicoefficient image (MCI).

A multispectral remote sensing image of a specific area contains *S* spectral wavebands, with *L* lines per *C* columns. At each ground resolution element, usually represented as a pixel, there is a spectral profile formed by the *S* spectral wavebands. When this pixel is imaged over time at *T* dates, a 3-D spectral-temporal space (STS) is formed. For modeling purposes, we shall assume that STS is formed by two independent variables, namely, time (**t**) and spectrum (**s**), and one dependent variable representing the observed values of the sensor (**r**), such as reflectance or band transformation (e.g., vegetation index[2]).

Thus, for each pixel, there are *n* points distributed within the STS, where *n* is given by the total sum of the number of the spectral wavebands for all *T* dates. These points can be obtained from several observations of sensors with different spectral wavebands that represent the spectral-temporal profile of the pixel. In short, the idea is to establish a model that describes the STS points based on the function

$$\mathbf{r} = f\,(\mathbf{t},\mathbf{s}). \qquad (2.1)$$

The model that represents the relationship between dependent (**r**) and independent (**t**,**s**) variables is denominated the spectral-temporal response surface model and will have *k* coefficients to be estimated for each pixel. Therefore, each coefficient will compose a specific synthetic band of the MCI. An overview of the STARS is presented in Fig. 2.1.

[2] The **s** variable can also represent arbitrary labels instead of spectral wavebands, but it may cause a lack of robustness (MELLO, 2009).

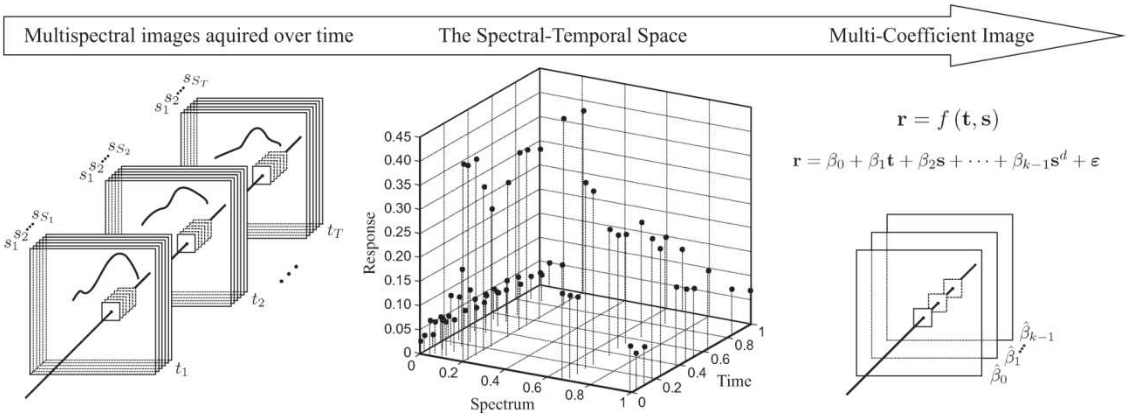


Figure 2.1 - Framework of the STARS method.

2.2.2. Spectral-temporal response surface model

The modeling of the spectral-temporal response surface assumes that the variables in the STS should be of the same magnitude, or else, they should be rescaled (VIEIRA, 2000).

As discussed by Watson (1992), there are several options for modeling the function shown in Eq. 2.1, and the choice used will depend on the purpose of the application. Two models tested by Vieira (2000) are particularly useful to represent Earth surface changes: (i) the polynomial trend surface (PTS) model that can generate relatively smooth surfaces representing gradual change such as crop growth and (ii) the collocation surface (CS) model that can generate relatively sharp surfaces representing abrupt change such as crop harvest.

2.2.2.1. Polynomial Trend Surface

PTS is a polynomial regression model that describes the distribution trend of the STS points (WATSON, 1992). Therefore, since PTS models tend to describe the general behavior of observed values on the spectral-temporal response surfaces, it is expected that their use minimizes problems associated with aberrant or noisy data such as cloud or cloud shadow in multitemporal images (VIEIRA, 2000). On the other hand, by describing general behavior, a PTS model may obscure important extreme values observed.

PTS is considered to be a special case of the general linear regression model (KUTNER *et al.*, 2005). In response surface interpolations, Watson (1992) describes the PTS as a bivariate linear combination expressed in terms of powers and cross products of the two independent variables (in this case, **t** and **s**). With the condition $k < n$, the k coefficients can be estimated using any method to solve overdetermined systems (e.g., least squares).

The system of simultaneous linear equations for the PTS model with degree d has the form

$$\mathbf{r} = \sum_{i=0}^{d}\sum_{j=0}^{i} \beta_{\left(\frac{i(i+1)}{2}+j\right)} \mathbf{t}^{i-j}\mathbf{s}^{j} + \boldsymbol{\varepsilon} \tag{2.2}$$

where the k coefficients are denoted by $\boldsymbol{\beta}$ (from β_0 to $\beta_{\text{k-1}}$) that will be estimated to their best unbiased point estimators, which are denoted by $\widehat{\boldsymbol{\beta}}$ (from $\hat{\beta}_0$ to $\hat{\beta}_{k-1}$), and $\boldsymbol{\varepsilon}$ represents the error vector assumed to be uncorrelated with mean equal to zero and variance equal to σ^2. For PTS models with two independent variables (e.g., **t** and **s**), k will depend on d following the relationship

$$k = \frac{(d+1)(d+2)}{2}. \tag{2.3}$$

For each pixel, the system in Eq. 2.2 can be written in matrix form as

$$\mathbf{r} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{2.4}$$

where **r** is a column vector with n observed values, **X** (commonly known as *design matrix*) is a full column rank matrix (i.e., linearly independent columns) that has n rows by k columns containing powers and cross product terms of independent variables, $\boldsymbol{\beta}$ is a column vector with k coefficients to be estimated, and $\boldsymbol{\varepsilon}$ is a column vector with n error elements.

As pointed out by Forsythe (1957), the solution of Eq. 2.4 to find $\widehat{\boldsymbol{\beta}}$ using

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{r} \tag{2.5}$$

might be inaccurate due to the fact that **X** becomes progressively more ill-conditioned as the degree of polynomials increases. However, Mather (1976) suggests that the modified Gram-Schmidt (MGS) orthonormalization (BJÖRCK, 1967) is an alternative to obtain a more accurate solution. This solution does not require matrix inversion but works upon the **X** matrix and is accurate even in badly conditioned problems, although the degree of accuracy is affected by the size of the residuals (MATHER, 1976). The use of the MGS orthonormalization has other advantages such as the minimization of computation roundoff errors and the independence of the terms of the equation (and then, the orthonormal coefficients are uncorrelated). This independence is important because it makes it possible to add or remove terms without the need to recalculate the existing ones. This fact enables some estimated orthonormal coefficients to be discarded in order to reduce dimensionality (see DRAPER; SMITH, 1966).

Thus, we can rewrite Eq. 2.4 using the orthonormal corresponding system given by

$$\mathbf{r} = \mathbf{Q}\boldsymbol{\alpha} + \boldsymbol{\epsilon} \tag{2.6}$$

where **Q** is an orthonormal matrix with *n* rows by *k* columns, calculated using MGS orthonormalization; $\boldsymbol{\alpha}$ is a column vector with *k* orthonormal coefficients (i.e., from α_0 to α_{k-1}); and ϵ is a column vector with *n* error elements associated with the orthonormalized system.

As pointed out by Mather (1976), the solution using MGS orthonormalization requires two steps: (i) the conversion of **X** to its orthonormalized corresponding **Q** and (ii) the determination of the orthonormal coefficients $\boldsymbol{\alpha}$. Considering that these orthonormal coefficients are abstract and cannot be directly interpreted, it might be interesting to convert them into the regular coefficients of the original PTS model ($\boldsymbol{\beta}$). This conversion can be done by the QR factorization (GOLUB; VAN LOAN, 1996), where the **Q** matrix is the same **Q** in the MGS process.

The MGS orthonormalization is carried out as follows: if we call $\mathbf{x}_{*1}$, $\mathbf{x}_{*2}$, …, $\mathbf{x}_{*k}$ the column vectors of **X** (the symbol "$*$" represents all row elements within a column), then the first orthonormal column vector of **Q** (i.e., $\mathbf{q}_{*1}$) is given by

$$\mathbf{q}_{*1} = \frac{\mathbf{x}_{*1}}{\|\mathbf{x}_{*1}\|} \tag{2.7}$$

where $\|\mathbf{x}_{*1}\|$ represents the norm of the column vector $\mathbf{x}_{*1}$. For the following column vectors of **Q**, the *i*th column vector ($\mathbf{q}_{*i}$) is calculated using an iterative process given by *i*-1 steps:

$$\begin{aligned}
\mathbf{q}_{*i}^{(1)} &= \mathbf{x}_{*i} - \left(\frac{\mathbf{q}_{*1}'\mathbf{x}_{*i}}{\mathbf{q}_{*1}'\mathbf{q}_{*1}}\right)\mathbf{q}_{*1} \\
\mathbf{q}_{*i}^{(2)} &= \mathbf{q}_{*i}^{(1)} - \left(\frac{\mathbf{q}_{*2}'\mathbf{q}_{*i}^{(1)}}{\mathbf{q}_{*2}'\mathbf{q}_{*2}}\right)\mathbf{q}_{*2} \\
&\vdots \\
\mathbf{q}_{*i}^{(i-1)} &= \mathbf{q}_{*i-1}^{(i-2)} - \left(\frac{\mathbf{q}_{*i-1}'\mathbf{q}_{*i}^{(i-2)}}{\mathbf{q}_{*i-1}'\mathbf{q}_{*i-1}}\right)\mathbf{q}_{*i-1}
\end{aligned} \tag{2.8}$$

and its subsequent normalization is given by

$$\mathbf{q}_{*i} = \frac{\mathbf{q}_{*i}^{(i-1)}}{\left\|\mathbf{q}_{*i}^{(i-1)}\right\|}. \tag{2.9}$$

The normalization to find $\mathbf{q}_{*i}$ is performed prior to the calculation of the next column vector $\mathbf{q}_{*i+1}$.

As discussed by Draper and Smith (1966) and Golub and van Loan (1996), the solution for the orthonormal system in Eq. 2.6 according to the orthonormality characteristic **Q'Q** = **I**, where **I** is the identity matrix, is given by the least squares as

$$\hat{\alpha} = \mathbf{Q}'\mathbf{r}. \tag{2.10}$$

If desirable, the regular coefficients of the PTS model ($\boldsymbol{\beta}$) can be calculated using the QR factorization. Thus, through the QR factorization, we can write

$$\mathbf{X} = \mathbf{QR} \tag{2.11}$$

where **R** is a square upper triangular matrix with *k* rows by *k* columns, with its elements calculated by

$$R_{ij} = \mathbf{x}'_{*j}\mathbf{q}_{*i}. \tag{2.12}$$

Thus, using Eq. 2.11 in Eq. 2.5, the least squares solution becomes

$$\mathbf{R}\hat{\boldsymbol{\beta}} = \mathbf{Q}'\mathbf{r}. \tag{2.13}$$

The solution of this upper triangular system will estimate the regular coefficients of the PTS model ($\boldsymbol{\beta}$) using, for instance, backward substitution. Another way to estimate $\boldsymbol{\beta}$ is using the relationship between Eqs. 2.10 and 2.13. Thus, the conversion of the orthonormal coefficients $\boldsymbol{\alpha}$ to $\boldsymbol{\beta}$ can be made by $\widehat{\boldsymbol{\beta}} = \boldsymbol{R}^{-1}\widehat{\boldsymbol{\alpha}}$.

2.2.2.2. Collocation Surface

CS is a model that uses the distances among the STS points to fit a surface that passes through all *n* points (WATSON, 1992). Spectral-temporal response surfaces fitted with CS models tend toward horizontal planes as the distance in the **t-s** plane of the STS increases (HARDY, 1971). Thus, this model is recommended for use both when remote sensing images are relatively free of noise and when the images were acquired from sensors with similar spectral characteristics.

Given the set of *n* points, the procedure solves the system of simultaneous linear equations estimating the *k* coefficients $\boldsymbol{\delta}$. The CS model in matrix form is given by

$$\mathbf{r} = \mathbf{D}\delta + \xi \tag{2.14}$$

where ξ is a column vector with *n* error elements associated with the collocation model, and **D** is the distance matrix, i.e.,

$$\mathbf{D} = \begin{pmatrix} e_1 & d(p_1,p_2) & \cdots & d(p_1,p_k) \\ d(p_2,p_1) & e_2 & \cdots & d(p_2,p_k) \\ \vdots & \vdots & \ddots & \vdots \\ d(p_k,p_1) & d(p_k,p_2) & \cdots & e_k \end{pmatrix} \tag{2.15}$$

with

$$d(p_i,p_j) = \sqrt{(t_i - t_j)^2 + (s_i - s_j)^2 + e_j^2} \tag{2.16}$$

where $d(p_i,p_j)$ is a function of the distance between the projections of the *i*th and *j*th points on the **t-s** plane of the STS. This distance is modified by the nonnegative constant *e*, which, in the case of relatively small-scale variations in distances, can be set to zero (HARDY, 1971).

It is worth mentioning that in the CS model, there is no dimensionality reduction since $k = n$. Therefore, Eq. 2.14 is a square system, and its stable solution can be found to be the same as for Eq. 2.4 using MGS orthonormalization, after replacing **X** and $\boldsymbol{\beta}$ for **D** and $\boldsymbol{\delta}$, respectively (GOLUB; VAN LOAN, 1996). Then, the conversion of the orthonormal coefficients into the regular coefficients of the CS model ($\boldsymbol{\delta}$) is performed the same way as in the PTS model.

2.2.3. Multi-Coefficient Image

Response surfaces may be comparated either directly (e.g., by difference) or indirectly (e.g., by estimated coefficients). The main advantage of using the estimated coefficients is that this approach tends to represent the form of the reponse surface, which makes the method more robust (VIEIRA, 2000).

The MCI has *L* lines per *C* columns and *k* synthetic bands formed by the *k* estimated coefficients for each pixel in a study area (with *L* lines per *C* columns). Thus, the estimated coefficients that compose the *k* bands of the MCI represent the information content of the multispectral-multitemporal remote sensing image data set for the study area. For example, in the PTS model, β_0 represents the surface offset with regard to the STS origin (KUTNER *et al.*, 2005). Indeed, each estimated coefficient, in some way, contains the description of the spectral variation over time.

2.3. Sugarcane agriculture context

The possible consequences of CO_2 accumulation in the atmosphere due to the use of fossil fuels, as well as the rise of oil prices, have triggered a considerable global interest in biofuels, which are considered relatively low-pollution energy sources (LEITE *et al.*, 2009). Hoogwijk *et al.* (2005) cite studies that project future growth of biofuels to supply between 5% and 50% of global energy demand. The potential for any biofuel to mitigate greenhouse gas (GHG) emissions is determined by the balance of emissions during all production steps, including agroindustrial ones, and biofuel consumption (MACEDO *et al.*, 2008). Thus, agricultural management techniques have a major role in calculating such balance (KIM *et al.*, 2009).

Of all biofuels, ethanol derived from sugarcane offers the highest GHG reduction rate when compared with gasoline (WALTER *et al.*, 2008). However, some management techniques such as preharvest burning (BH), which can make manual harvesting easier, decrease this biofuel mitigating potential, since CO, CH_4, and particulate material are emitted through this process (FIGUEIREDO; LA SCALA JR., 2011; GALDOS *et al.*, 2009; GOLDEMBERG *et al.*, 2008; KIRCHHOFF *et al.*, 1991). Moreover, the practice of BH has been identified as the cause of an increase in respiratory diseases, as measured by hospital admissions data, mainly among children and senior citizens (CANÇADO *et al.*, 2006; LARA *et al.*, 2005; URIARTE *et al.*, 2009). In addition, leaving straw on the fields (i.e., not burning) decreases soil and water loss and helps in the maintenance of soil carbon storage (GALDOS *et al.*, 2009).

Given the detrimental environmental consequences of BHs, São Paulo State, which is responsible for 60% of ethanol production in Brazil, has pledged agreement to a "Green Ethanol" Protocol (for more details, see Lucon and Goldemberg (2010)). This protocol, agreed among the São Paulo State Secretary of Environment, the Sugarcane Industry Union, and the supplier associations and ethanol producers, aims to gradually reduce straw burning in sugarcane plantations, leading to a complete cessation of the practice by 2014 for mechanized areas (terrain slope $\leq 12\%$) and 2017 for nonmechanized areas (terrain slope $> 12\%$).

Since 2006, the Brazil's National Institute for Space Research (INPE) has monitored adherence to this protocol through remote sensing. To achieve this, visual interpretation has been used to analyze at least one image coverage per month over harvest seasons (AGUIAR *et al.*, 2011; RUDORFF *et al.*, 2010). Visual image interpretation is particularly effective for distinguishing BH fields and green harvested (GH) fields (i.e., nonburned) due to the strong contrast between dark burned fields, where bare soil is dominant after harvest (see Fig. 2.2), and bright GH fields, where a layer of dry leaves (straw) covers the ground after harvest (see Fig. 2.3). The dark or bright contrast indicates whether the sugarcane field was BH or GH, and this remains very evident for several days or even weeks after harvest (AGUIAR *et al.*, 2011). However, although visual interpretation can be effective for sugarcane harvest characterization, it is a laborious task and not practical for very large areas or very regular surveys. Alternative automated classification procedures hold considerable potential here, particularly since crop monitoring is required for such a large area throughout the April to December harvest season (RUDORFF *et al.*, 2010).

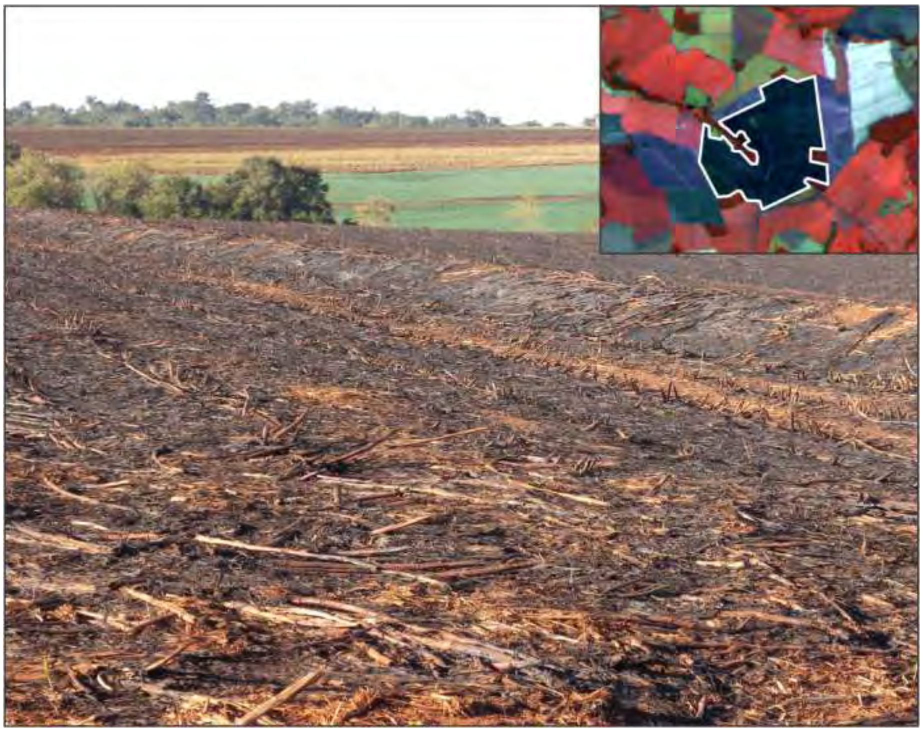


Figure 2.2 - BH sugarcane field, highlighting (inset) its appearance in a false color composite TM/Landsat-5 image.

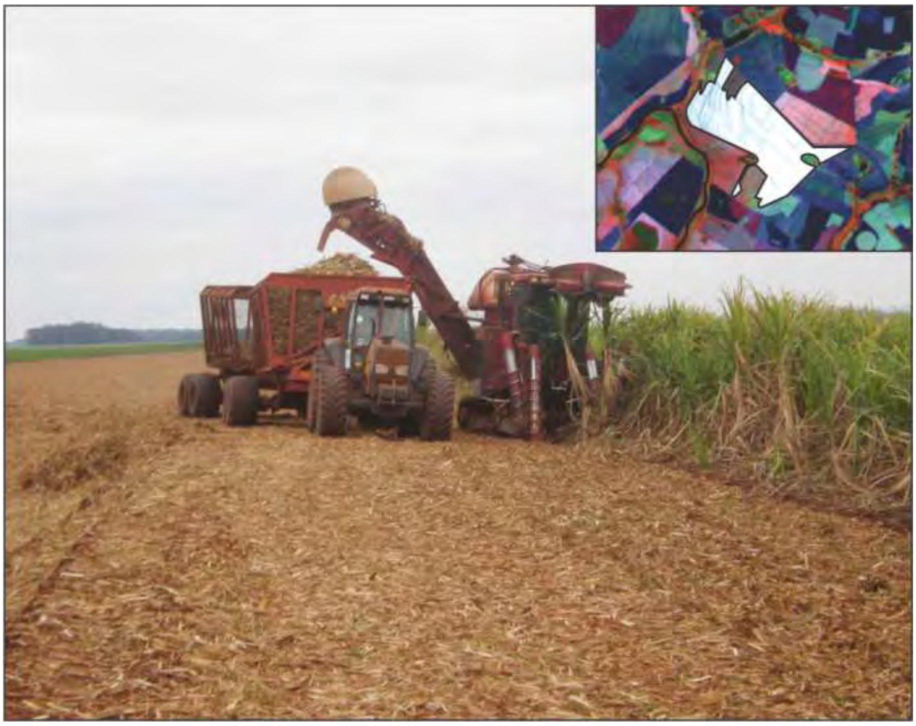


Figure 2.3 - Mechanized sugarcane harvest without BH (called GH), showing the straw remaining on the ground and highlighting (inset) its appearance in a false color composite TM/Landsat-5 image.

2.4. Research materials and methods

This section presents the materials and methods employed in the application of STARS to a sugarcane classification case study. Specifically, two types of sugarcane harvest practices are investigated: with BH and GH. Fig. 2.4 shows a flowchart summarizing the methods employed in the application of STARS to the case study. In short, 66 original wavebands (11 images, each with six spectral wavebands) of a multitemporal-multispectral Landsat data set were georeferenced and radiometrically corrected (including atmospheric correction). The resulting 66 georeferenced surface reflectance wavebands were then used as input for STARS, which ran under 20 different scenarios. Each scenario generated a different MCI, and the *k* synthetic bands of each MCI were used as input for three classifiers, producing 60 classified thematic maps. The georeferenced surface reflectance wavebands were also used together with a 5 m *Systeme Pour l’Observation dela Terre* (SPOT) image and field data to create a reference thematic map, which was used to select both the training and the testing samples for classification and accuracy assessment, respectively.

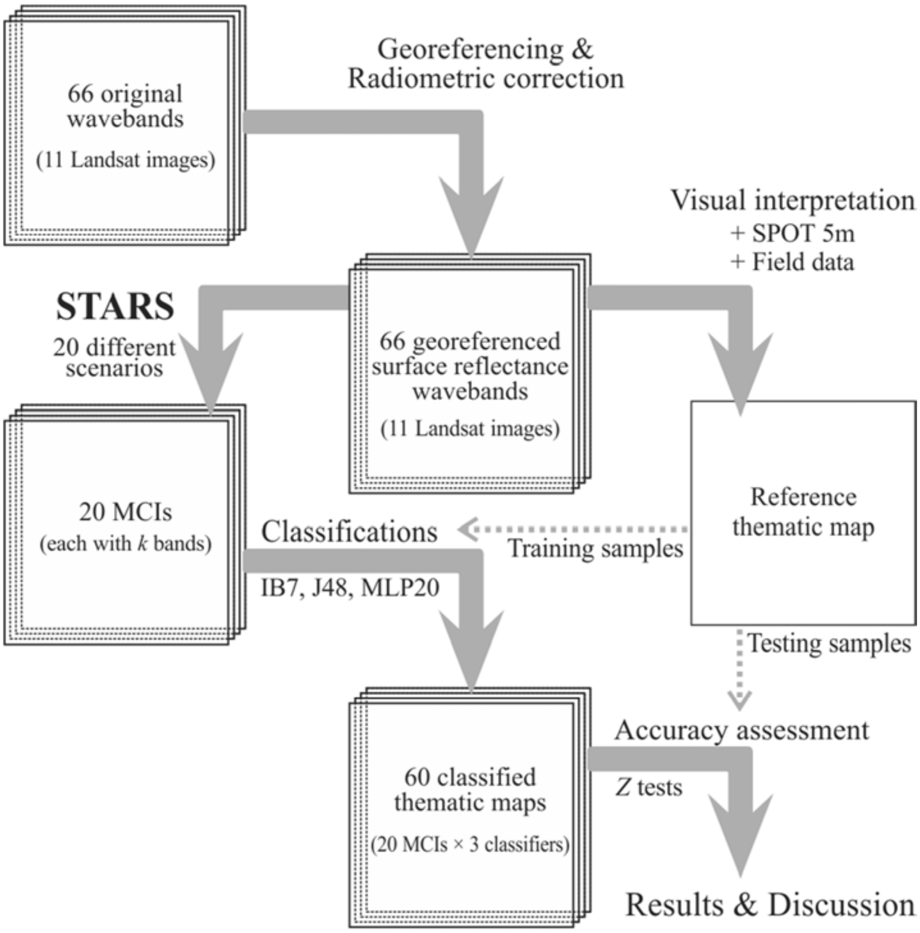


Figure 2.4 - Flowchart summarizing the application of STARS to the case study.

The research materials and methods employed in the application of STARS are detailed in the following subsections.

2.4.1. Study area

The study area, shown in Fig. 2.5, is a densely cultivated sugarcane region in São Paulo State, well represented by the two harvest types practiced: BH and GH. This region is located in the northern part of São Paulo State and comprises three municipalities, namely, Guará, Ipuã, and São Joaquim da Barra. In total, these municipalities cultivated about 60,000 ha of sugarcane in 2001 (IBGE, 2012b), with significant areas harvested both BH and GH. The year 2001 was used to test STARS due to the availability of a series of cloud-free Landsat sensor images, plus a strong field reference data set used to both train classifiers and assess classification results.

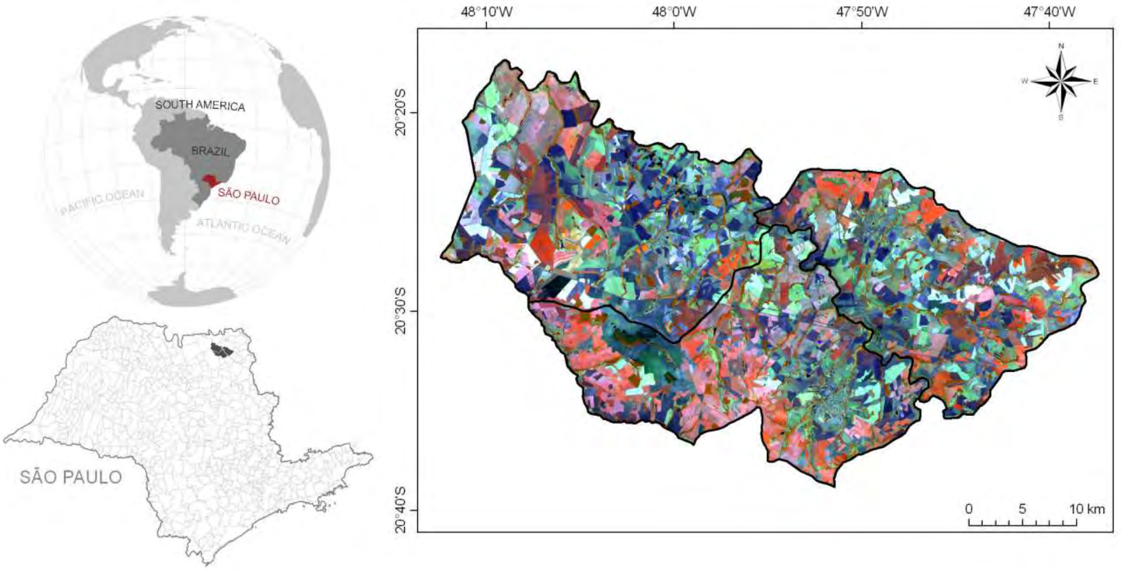


Figure 2.5 - Location of the study area, in São Paulo State, Brazil, highlighting the municipalities of Guará, Ipuã, and São Joaquim da Barra in a false color composite ETM+/Landsat-7 image acquired on July 29, 2001.

2.4.2. Multitemporal Landsat sensor images

Eleven cloud-free Landsat-5 Thematic Mapper (TM) and Landsat-7 Enhanced Thematic Mapper Plus (ETM+) images were acquired in 2001 ($T = 11$), covering dates from the beginning of the harvest season in April to the end of the harvest season in December (RUDORFF *et al.*, 2010) (see Table 2.1).

Table 2.1 - Summary of the 11 Landsat images used.

Image #	**Sensor/Satellite**	**Date**	**Day of year**
1	ETM+/Landsat-7	Apr. 08, 2001	98
2	TM/Landsat-5	Apr. 16, 2001	106
3	TM/Landsat-5	May 02, 2001	122
4	TM/Landsat-5	May 18, 2001	138
5	TM/Landsat-5	Jun. 03, 2001	154
6	TM/Landsat-5	Jul. 05, 2001	186
7	ETM+/Landsat-7	Jul. 29, 2001	210
8	ETM+/Landsat-7	Aug. 14, 2001	226
9	TM/Landsat-5	Sep. 07, 2001	250
10	TM/Landsat-5	Oct. 25, 2001	298
11	ETM+/Landsat-7	Dec. 04, 2001	338

Analysis was conducted using six of the Landsat sensors' spectral wavebands, corresponding to the blue (b1), green (b2), red (b3), near-infrared (NIR, b4), and shortwave infrared (SWIR, b5 and b7) parts of the electromagnetic spectrum. The spatial resolution of Landsat TM/ETM+ is 30 m.

2.4.3. Image georeferencing

The remote sensing images were georeferenced using 21 ground control points collected with a dual-frequency Global Positioning System (GPS) receiver with differential correction based on the two nearest stations in the Brazilian Network for Continuous Monitoring (RBMC: *Rede Brasileira de Monitoramento Contínuo*), i.e., UBER and MGUB, which are both located in the municipality of Uberlândia, Minas Gerais State, Brazil. The coordinates of these control points were projected according to the WGS84/UTM-23S map projection system, achieving positional errors of less than 50 cm. The images were georeferenced using first-degree polynomials and nearest-neighbor resampling, and the output images had a root-mean-square error of less than 0.5 pixels, as recommended by Dai and Khorram (1998).

2.4.4. Image radiometric correction

As recommended by Song *et al.* (2001), atmospheric correction should be taken into consideration in preprocessing for applications where a common radiometric scale is assumed among the multitemporal remote sensing data set. Thus, each Landsat image (see Table 2.1) was converted from digital number to radiance and then to top of atmosphere reflectance (apparent reflectance), as proposed by Markham and Barker (1986), using the parameters presented by Chander *et al.* (2009). When used for multitemporal analysis, Schroeder *et al.* (2006) recommended the use of radiometric normalization rather than atmospheric correction of each image. In this case, the July 29, 2001 image (image #7 in Table 2.1) was atmospherically corrected and used as a base image for subsequent radiometric normalization of the remaining ten images.

Image #7 was chosen for atmospheric correction since: (i) this is from a central period in the multitemporal data set (see Table 2.1), and (ii) this is a Landsat-7/ETM+ image, and the Landsat-7 satellite orbits in tandem with the EOS Terra satellite sensor for nearcoincident observations; hence, the aerosol optical depth product from Terra's MODIS could be used to estimate the visibility parameter in the atmospheric correction procedure (OLIVEIRA *et al.*, 2009). Image #7 was then converted from top of atmosphere reflectance to surface reflectance using the Second Simulation of the Satellite Signal in the Solar Spectrum (6S) algorithm (VERMOTE *et al.*, 1997). Eventually, using the processed image #7 surface reflectance as base, the other ten Landsat images listed in Table 2.1 were radiometrically normalized based on the iteratively reweighted multivariate alteration detection (IR-MAD) transformation (CANTY; NIELSEN, 2008; CANTY *et al.*, 2004) [with penalization parameter set to zero (NIELSEN; CANTY, 2005)]. After this normalization, it could be assumed that all Landsat sensor images were converted to surface reflectance and that all six spectral wavebands of the two Landsat sensors shared the same radiometric characteristics.

2.4.5. Spectral-temporal profile investigation

Since STARS aims to represent spectral response over time, it is important to have some knowledge about the spectral-temporal profile of the classes of interest prior to applying STARS. If the spectral profiles of the classes are different over time, it is expected that STARS will be able to represent each class of interest (enabling classifiers to distinguish these classes). Mello (2009) used six spectral wavebands of TM/Landsat-5 (b1 to b5 and b7) acquired at six different dates throughout the harvest season of 2007 to investigate the dynamic spectral-temporal nature of BH and GH sugarcane fields in São Paulo State (see Fig. 2.6).

Fig. 2.6 shows that the two harvest events (BH and GH) present distinct spectral responses over time. The spectral-temporal dynamic of the BH sugarcane field (see Fig. 2.6, left-hand side) is characterized by a minor reflectance decrease in the green waveband (b2) and a minor reflectance increase in the red waveband (b3). A significant reduction in the reflectance value is observed in the NIR waveband (b4) due to biomass removal during burning and harvest (see Fig. 2.2). The SWIR wavebands (b5 and b7) are strongly affected by soil type and moisture content (CARTER, 1991; GAUSMAN *et al.*, 1969; MELLO, 2009).

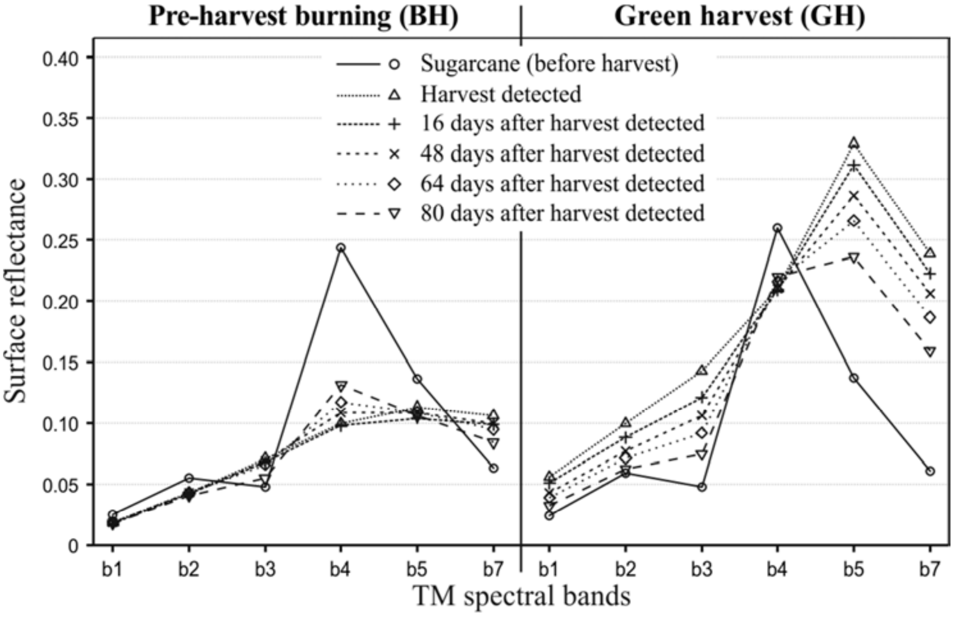


Figure 2.6 – Spectral-temporal dynamic of BH and GH sugarcane fields. Source: Adapted from Mello (2009).

The spectral-temporal dynamic of the GH sugarcane field (see Fig. 2.6, right-hand side) is characterized by relatively high reflectance values in all visible spectral wavebands (b1, b2, and b3) due to the bright reflectance of dry matter (straw). In the NIR waveband (b4), reflectance is generally quite low due to biomass loss (GAUSMAN *et al.*, 1969). In the SWIR wavebands (b5 and b7), reflectance is relatively high due to the low water content of the straw that remains on the ground (CARTER, 1991) (see Fig. 2.3). As the sugarcane crop gradually regrows after harvest, either BH or GH, the fields' spectral profiles over time tend to become similar to those observed before harvest.

In the spectral-temporal analysis of sugarcane harvested fields, it is important to consider the timing of the harvest event. Indeed, although the spectral dynamic of a field GH in May can be similar to that of a field GH in October, the spectral-temporal response surfaces of these fields will be different as a result of the different dates of the harvest event, since the spectral profile of a harvested field (BH or GH) changes over time (see Fig. 2.6). Thus, for labeling purposes, each BH or GH sugarcane field appearing on any of the 11 Landsat images (see Table 2.1) is labeled according to the image number in which the harvest event was observed. The thematic classes are summarized in Table 2.2. If the sugarcane field was not harvested, it is labeled as unharvested (UH).

Table 2.2 - Summary of the 23 thematic classes used in classifications.

Description	**Label**
Pre-harvest burning identified on the image #1	BH01
Pre-harvest burning identified on the image #2	BH02
Pre-harvest burning identified on the image #3	BH03
Pre-harvest burning identified on the image #4	BH04
Pre-harvest burning identified on the image #5	BH05
Pre-harvest burning identified on the image #6	BH06
Pre-harvest burning identified on the image #7	BH07
Pre-harvest burning identified on the image #8	BH08
Pre-harvest burning identified on the image #9	BH09
Pre-harvest burning identified on the image #10	BH10
Pre-harvest burning identified on the image #11	BH11
Green harvest identified on the image #1	GH01
Green harvest identified on the image #2	GH02
Green harvest identified on the image #3	GH03
Green harvest identified on the image #4	GH04
Green harvest identified on the image #5	GH05
Green harvest identified on the image #6	GH06
Green harvest identified on the image #7	GH07
Green harvest identified on the image #8	GH08
Green harvest identified on the image #9	GH09
Green harvest identified on the image #10	GH10
Green harvest identified on the image #11	GH11
Unharvested sugarcane	UH

2.4.6. Reference map creation

The reference map was populated using field data and visual interpretation of the Landsat sensor images (see Table 2.1) in two steps: (i) a thematic map with the classes sugarcane and nonsugarcane was generated using the method described by Rudorff *et al.* (2010) and evaluated by Adami *et al.* (2012), and (ii) the thematic map with these classes was then used to evaluate the harvested sugarcane (BH and GH), as described by Aguiar *et al.* (2011). This interpretation generated a thematic map with 22 classes of sugarcane harvested with BH and GH, depending on the image number in which the harvest event was visually observed. Moreover, the sugarcane fields that were not harvested throughout the harvest season were classified as UH. Eventually, the reference map has 23 thematic classes, as described in Table 2.2.

To improve the detail of the reference map to a spatial resolution of 5 m, a panchromatic SPOT-5 high-resolution geometry image acquired on October 7, 2002 was used to delineate the sugarcane fields precisely. Finally, an erosion filter[3] (HARALICK *et al.*, 1987) was applied to the reference map to discard border pixels, preventing both misclassification due to spectrally mixed pixels and underestimation of the classification accuracy due to positional uncertainty (FOODY, 2002).

2.4.7. Multi-Coefficient Image testing

It is expected that the MCI is able to represent the spectral-temporal information content of the 11 Landsat images (see Table 2.1) and, according to their spectral-temporal profile, indicates not only the harvest date but also whether the harvest practiced was BH or GH.

Based on the 11 multitemporal-multispectral Landsat sensor images (see Table 2.1) used in this case study, the variables that define the coordinates of the STS points were: the image date acquisition ($\mathbf{t} = \boldsymbol{\tau}$), given in day of year; the central wavelength of each Landsat spectral waveband[4] ($\mathbf{s} = \boldsymbol{\lambda}$), given in micrometers; and the observed values given by the surface reflectance ($\mathbf{r} = \boldsymbol{\rho}$), varying from 0 to 1. In order to standardize all variables to the same magnitude, the variables $\boldsymbol{\tau}$ and $\boldsymbol{\lambda}$ were rescaled to a closed interval between 0 and 1, as suggested by Vieira (2000), before running STARS.

The MCI is the result of STARS and depends on the model used to describe the spectral-temporal response surface for every pixel in the study area. The PTS model can be simple, by setting the degree to one ($d = 1$), and the complexity is increased as d increases. As discussed by Kutner *et al.* (2005), the degree must be correctly chosen to

[3] At a spatial resolution of 5 m, an erosion filter with a window of 13 × 13 eliminates a border of two Landsat pixels.
[4] After the radiometric normalization procedure, the six corresponding spectral wavebands of all Landsat images were considered to having the ETM+ central wavelengths, since an ETM+ image was used as base in the radiometric normalization procedure. Thus, λ has six levels. The central wavelength values considered were presented by Chander *et al.* (2009).

produce a suitable response surface, i.e., neither too small to inadequately describe the surface nor too big to produce large anomalies in the surface. In order to evaluate different degrees for the PTS model and also differences between the orthonormal and the regular coefficients for both the PTS and CS models, we ran STARS with 20 different scenarios and generated 20 MCIs, as summarized in Table 2.3.

Table 2.3 - Summary of the 20 MCIs tested.

MCI #	**Model**	*d*	*k*	**Type of coeff.**
MCI 01	PTS	1	3	Orthonormal
MCI 02	PTS	2	6	
MCI 03	PTS	3	10	
MCI 04	PTS	4	15	
MCI 05	PTS	5	21	
MCI 06	PTS	6	28	
MCI 07	PTS	7	36	
MCI 08	PTS	8	45	
MCI 09	PTS	9	55	
MCI 10	CS	-	66	
MCI 11	PTS	1	3	Regular
MCI 12	PTS	2	6	
MCI 13	PTS	3	10	
MCI 14	PTS	4	15	
MCI 15	PTS	5	21	
MCI 16	PTS	6	28	
MCI 17	PTS	7	36	
MCI 18	PTS	8	45	
MCI 19	PTS	9	55	
MCI 20	PTS	-	66	

d: degree of polynomials; *k*: number of coefficients, given by Eq. 2.3.

Finally, the *k* synthetic bands of each MCI, instead of the 66 Landsat multitemporal-multispectral wavebands, were used as input attributes for classification. Three classifiers were tested, as described below. As recommended by Vieira (2000), each synthetic band of each MCI was individually rescaled to a closed interval between 0 and 1 before classification to avoid significant differences in magnitude, which can affect the performance of some classifiers (TSO; MATHER, 2009). Since we compared 20 MCIs as the input for three different classifiers, we generated 60 classification products.

2.4.8. Classification

Three classification techniques were selected for comparison, enabling rigorous evaluation of the STARS method. First, instance-based classification was performed, which does not rely on a model to classify the data. Second, decision-tree classification was performed, which creates a simple interpretable model from the data, although it may be relatively inefficient. Third, neural-network classification was performed, which creates an accurate model, although it may not necessarily be easily interpretable (TAN *et al.*, 2006).

Instance-based classification is based on the instances themselves, instead of a model derived from labeled instances. This type of classifier uses the labeled data themselves to classify unlabeled data. Unlike most classification algorithms, there is no need for a preliminary step to create a model from the labeled data — unlabeled instances are compared with all labeled ones, and a majority vote determines the label for that instance, assigning a label to an instance based on the majority count of labels on nearby instances. (The algorithm is also known as k-nearest neighbors.) Usually, a limiting number (a small positive integer) is used; hence, only this number of labeled instances is considered when deciding a label. We set this value to 7 (hence, the classifier will be referred to as IB7). This algorithm is computationally more demanding than the other two classifiers, particularly if the labeled data set is very large, but it has the advantage of being able to deal with practically any kind of data distribution (AHA *et al.*, 1991).

Decision trees are classification algorithms that attempt to classify single instances of the data by comparing the values of their attributes with decision rules. These rules are stored in the classifier model, created in a preliminary step, which uses the labeled instances to create a set of hierarchical rules for posterior classification. The main advantage of this algorithm is that a decision-tree is easily interpretable by humans as long as it is kept simple (i.e., without too many rules). The main disadvantage of the algorithm is that in its canonical form, the rules correspond to orthogonal cuts or separations in the attribute space — if the classes on the data are orthogonally separable, the algorithm will yield accurate classification results and relatively simple trees, but otherwise, the rule set may become too large for interpretation, even with accurate classification results. In this experiment, we used the J4.8 implementation (WITTEN *et al.*, 2011) of the C4.5 algorithm (QUINLAN, 1993) for decision-tree classification. This classifier will be referred to as J48.

The third classification method is based on a neural network trained with the back-propagation algorithm (LOONEY, 1997). Neural networks are generally considered effective classifiers and, through the combination of linear classifiers, are able to classify nonlinear data distributions and even disjoint data distributions accurately, as long as there are enough neurons in the hidden layers to create these combinations. Neural networks must be trained in a preliminary step to classification. Training with this algorithm requires a set of labeled data and several passes through the algorithm, which may be time consuming depending on the neural-network architecture and on the number of labeled samples. In the classification step, the trained network is used to derive the classes for unlabeled samples. The main advantage of this algorithm is its ability to classify data with any type of distribution accurately; its main disadvantage is that the model (i.e., the trained neural network) is not easily interpretable, and the determination of the network architecture, particularly the number of neurons in the hidden layer, is somehow empirical (FAUSETT, 1993). We used a multilayer perceptron model, i.e., a feedforward neural network, and set the number of neurons in the hidden layer at 20. Thus, we will refer to this classifier as MLP20.

In each classification, two thirds of the pixels in each class (see Table 2.2) were randomly selected from the reference map for the training of the classifier. The remaining third of the reference map pixels was set aside to be used for accuracy assessment.

2.4.9. Accuracy assessment

The accuracy assessment was conducted by comparing classified and reference data. The sample size was computed according to the multinomial statistical distribution, as recommended by Congalton and Green (2009). The sample size should be neither too small such that it could not detect a difference that is actually important nor too large such that tiny differences in accuracy are declared statistically significant (FOODY, 2009) at a specific significance level (α). Thus, considering the 23 thematic classes (see

Table 2.2), $\alpha = 5\%$ and the worst case (where one single class covers about 50% of the entire study area), the minimum sample size per class was computed as 41 pixels. Therefore, we used 50 pixels per class, which were randomly collected, for each classification based on the third of reference data not used for training the classifier.

Statistical tests based on the standardized Gaussian distribution (*Z* distribution) were performed, as detailed by Congalton and Green (2009) – pag. 107, for testing the significance of and between the classifications (represented by their confusion matrices) using their estimates of kappa index ($\hat{\kappa}$) and kappa's variance ($\sqrt{\hat{\sigma}(\hat{\kappa})}$) (COHEN, 1960; HUDSON, 1987). The test of significance of a classification was performed under the hypotheses H_o: $\kappa = 0$ and H_a: $\kappa \neq 0$, whereas the significance between the differences of two estimated kappa indices (pairwise test) was performed under the hypotheses H_o: $(\kappa_1 - \kappa_2) = 0$ and H_a: $(\kappa_1 - \kappa_2) \neq 0$.

For these two *Z*-tests, H_o is rejected if the calculated statistic $z_{calc} \geq z_{\alpha/2}$, where $\alpha/2$ is the confidence level of the two-tailed *Z*-test with degrees of freedom assumed to be infinity. Another interpretation can be made using the *p*-value related to z_{calc}: H_o is rejected for *p*-values smaller than α.

In this paper, pairwise *Z*-tests were conducted to evaluate differences in accuracy values: (i) between classifications of MCIs based on orthonormal and regular coefficients, (ii) according to MCI complexity used as input for classifiers, and (iii) among classifiers. We assumed $\alpha = 5\%$ (i.e., 0.05) for all *Z*-tests conducted.

2.5. Results and discussion of STARS

This section presents and discusses the results of applying STARS for sugarcane harvest classification. Examples of fitted spectral-temporal response surfaces will be presented for both PTS and CS models, followed by illustration of one of the 20 MCIs tested (see Table 2.3). Next, we will present the classifications of the STARS outputs (MCIs) and the accuracy assessment.

2.5.1. Spectral-temporal response surface

The main rationale behind STARS is to use a spectral-temporal response surface model to describe the spectral behavior of pixels over time. To facilitate the comprehension of how a response surface model can be used in this context, example spectral-temporal response surfaces are illustrated for a pixel from a sugarcane field GH (GH04, see Table 2.2), using both the PTS model (see Fig. 2.7) and the CS model (see Fig. 2.8).

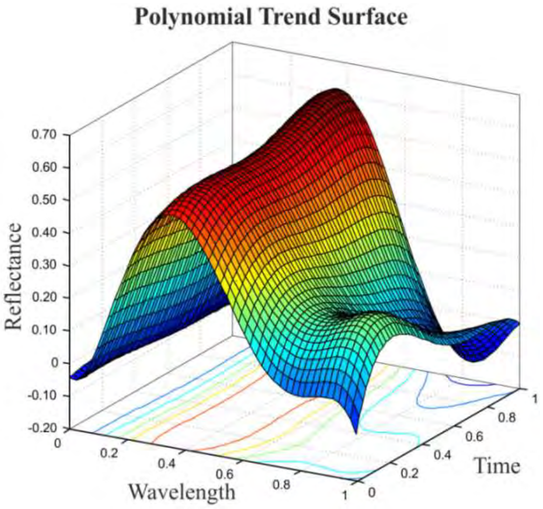


Figure 2.7 – Adjusted spectral-temporal response surface, for a pixel from a sugarcane field GH (GH04, see Table 2.2), using the PTS model with *d* = 5 and considering the regular coefficients.

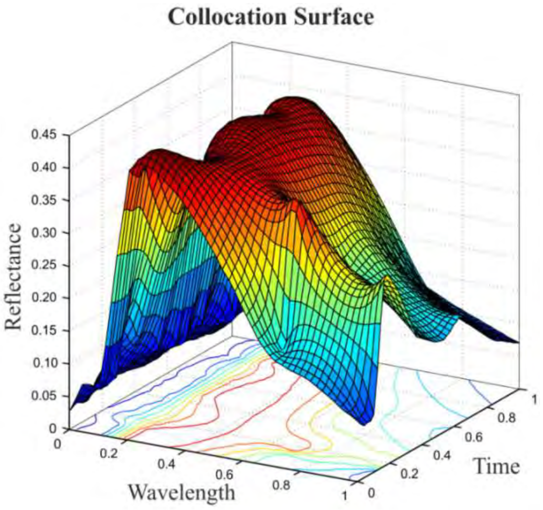


Figure 2.8 – Fitted spectral-temporal response surface, for a pixel from a sugarcane field GH (GH04, see Table 2.2), using the CS model and considering the regular coefficients.

The response surface drawn in Fig. 2.7 relates to the PTS model with $d = 5$, considering the regular coefficients (i.e., it corresponds to the MCI 15 in Table 2.3). This degree was chosen because it produces a suitable response surface [i.e., $d = 5$ is neither too small to inadequately describe the surface nor too big to produce large anomalies in the surface (KUTNER *et al.*, 2005)]. With $d = 5$, the PTS model has 21 coefficients ($k = 21$) and has the form

$$\boldsymbol{\rho} = \beta_0 + \beta_1 \boldsymbol{\tau} + \beta_2 \boldsymbol{\lambda} + \beta_3 \boldsymbol{\tau}^2 + \beta_4 \boldsymbol{\tau}\boldsymbol{\lambda} + \cdots + \beta_{20} \boldsymbol{\lambda}^5 + \boldsymbol{\varepsilon}. \tag{2.17}$$

The spectral-temporal profile of the chosen pixel (see Fig. 2.7) represents typical spectral behavior and change over time in response to the harvest event that can be observed at $\tau = 0.167$ (image #4 of the Landsat images listed in Table 2.1). The straw left on the ground after the mechanical sugarcane harvest (see Fig. 2.3) causes a substantial increase in the reflectance of the SWIR wavebands (b5 and b7, see Fig. 2.6) (MELLO, 2009). After the harvest event, a gradual regrowth of the sugarcane field can be observed from $\tau = 0.167$ to $\tau = 1$. Apparently, there are two inconsistencies on the estimated reflectance surface shown in Fig. 2.7: (i) the maximum estimated reflectance is higher than expected, and (ii) there are negative reflectance values. These inconsistencies are likely associated with both the high degree of the polynomial and the distribution of the STS points along the wavelength axis since most of them are concentrated in the visible and NIR wavebands generating anomalies in the fitted surface. However, these apparent inconsistencies do not pose a problem for subsequent analysis since we are actually interested in the estimated coefficients and not in the estimated reflectance values.

The CS model presents the number of coefficients (k) equal to n (i.e., $k = n = 66$ in this work). In Fig. 2.8, it can be noticed that the response surface described by the CS model is considerably sharper when compared with the surface described by the PTS model (see Fig. 2.7). It seems that the CS model can better describe the abrupt spectral change related to the harvest event than the PTS model; although even subtle changes in the estimated coefficients of the PTS model should be able to detect the harvest event, as will be described further.

2.5.2. Multi-Coefficient Image

The two spectral-temporal response surfaces presented (see Figs. 2.7 and 2.8) correspond to only one pixel, chosen as an example. By fitting a response surface for every pixel in the study area, each estimated coefficient will compose a synthetic band of the MCI. In this paper, we tested 20 different MCIs by varying the model (PTS or CS), the degree for PTS model, and considering either the orthonormal or the regular coefficients (see Table 2.3). It is worth mentioning that, although actual specialized hardware systems can satisfy the time-critical constraints introduced by the large amounts of computations usually regarding remote sensing data processing (LEE *et al.*, 2011), it was noticed that the computational time demanded by STARS tended to increase with the complexity of the models.

Fig. 2.9 presents the estimated orthonormal α_0 synthetic band of the MCI 14, described in Table 2.3. Bright areas correspond to areas where the average spectral responses over time tend to be high, considering all the six spectral wavebands used. They can be associated with bare soil areas, which have average reflectance higher than vegetation (except in the NIR waveband) (PONZONI; SHIMABUKURO, 2007), or they can also be associated with areas where straw, after mechanical harvest, remained on the ground for a long period of time (AGUIAR *et al.*, 2011; MELLO, 2009) (see Fig. 2.3). Thus, since an MCI presents not only the spectral information of the pixels in the study area but also their spectral behavior over time, a classification using an MCI as input can be considered a multitemporal-multispectral classification. Indeed, the MCI can be used for various purposes in the context of multitemporal analysis. For instance, band ratios can be used to create spectral-temporal indices or even land-cover change indices. Alternatively, synthetic bands can be used for visual interpretation or as input for land-use and land-cover change models, etc.

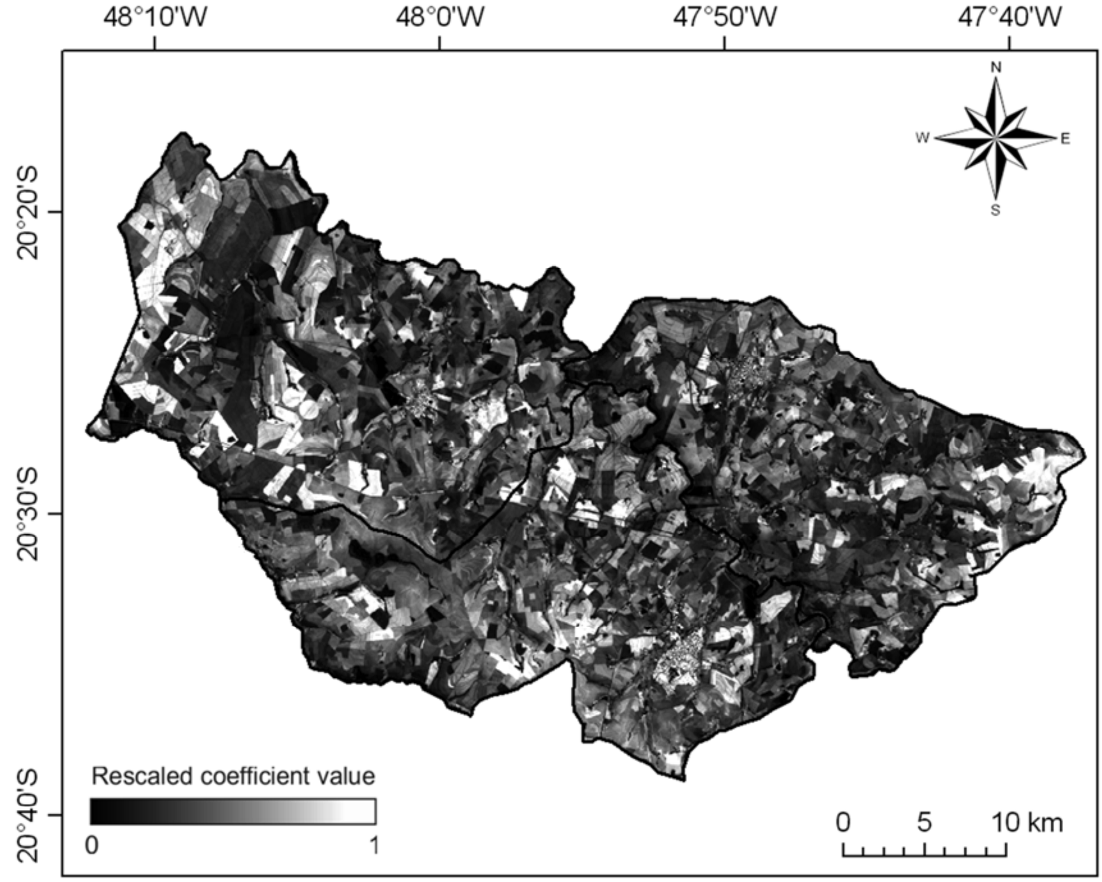


Figure 2.9 – Orthonormal $\hat{\alpha}_0$ synthetic band of the MCI computed for the study area.

2.5.3. Classifications

The 20 MCIs summarized in Table 2.3 were classified using three classifiers: instance-based, considering seven nearest neighbors (IB7); a decision-tree, using the J4.8 implementation of the C4.5 algorithm (J48); and a multilayer perceptron model, which is a feedforward neural network, with 20 neurons in the hidden layer (MLP20). The classification accuracy values were measured using the estimations of both kappa ($\hat{\kappa}$) and kappa's variance ($\hat{\sigma}(\hat{\kappa})$) values, computed through the confusion matrix for each classification (see an example in Table 2.7). The $\hat{\kappa}$ and $\hat{\sigma}(\hat{\kappa})$ for all 60 classifications are listed in Table 2.4.

By observing Table 2.4, one can see that $\hat{\kappa}$ tends to increase as the complexity of the MCI increases from MCI 01 to MCI 10 (orthonormal coefficients) and from MCI 11 to MCI 20 (regular coefficients). The number of (input) attributes is defined by k in the classifications, and as presented in Table 2.3, k increases with the MCI complexity.

Indeed, the accuracy of the classifications tends to increase when MCI complexity increases due to the fact that more attributes are considered (TSO; MATHER, 2009). It can also be noticed in Table 2.4 that $\hat{\sigma}(\hat{\kappa})$ decreases with increasing MCI complexity.

Table 2.4 - Summary of the estimated kappa and kappa's variance.

MCI #	**IB7**		**J48**		**MLP20**	
	$\hat{\kappa}$	$\hat{\sigma}(\hat{\kappa})$ x 10^3	$\hat{\kappa}$	$\hat{\sigma}(\hat{\kappa})$ x 10^3	$\hat{\kappa}$	$\hat{\sigma}(\hat{\kappa})$ x 10^3
MCI 01	.5273	.234366	.5245	.234445	.3936	.225083
MCI 02	.9227	.065034	.8264	.131529	.6236	.215125
MCI 03	.9727	.024146	.9064	.077487	.8000	.146586
MCI 04	.9864	.012235	.9409	.050672	.8773	.098368
MCI 05	.9918	.007380	.9445	.047734	.9273	.061505
MCI 06	.9918	.007380	.9573	.037251	.9527	.041017
MCI 07	.9918	.007380	.9573	.037252	.9509	.042527
MCI 08	.9973	.002473	.9618	.033439	.9573	.037246
MCI 09	.9918	.007380	.9673	.028819	.9673	.028817
MCI 10	.9945	.004933	.9718	.024928	.9591	.035730
MCI 11	.5291	.234211	.5045	.235161	.3636	.217061
MCI 12	.8236	.133161	.7482	.173685	.6182	.217496
MCI 13	.9291	.060081	.8755	.099716	.8055	.143632
MCI 14	.9527	.041029	.9009	.081524	.8327	.127162
MCI 15	.9600	.034971	.9091	.075451	.9282	.060784
MCI 16	.9673	.028819	.9064	.077476	.9400	.051401
MCI 17	.9736	.023361	.9091	.067853	.9345	.055755
MCI 18	.9809	.017038	.9227	.065048	.9482	.044762
MCI 19	.9764	.021000	.9091	.075426	.9582	.036491
MCI 20	.9900	.009004	.9500	.043273	.9691	.027266

Fig. 2.10 presents the estimated kappa values ($\hat{\kappa}$) in a graph, which shows that the accuracy of the three classifiers significantly increases from MCI 01 to MCI 04 (orthonormal coefficients) and from MCI 11 to MCI 14 (regular coefficients). Some further classification improvement is observed for the MPL20 classifier when using MCI 05 and particularly MCI 15. Classification with MCIs of greater complexity does not significantly increase classification accuracy. However, the Hughes phenomenon (HUGHES, 1968) was not observed, even for MCI 10 and MCI 20, which can be attributed to the satisfactory training sample size (ABEND; HARLEY JR., 1969).

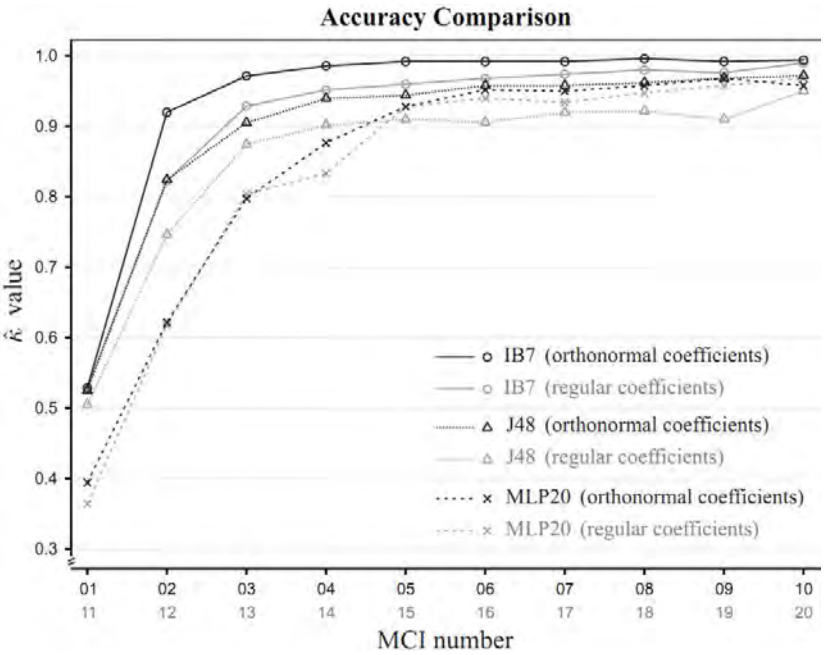


Figure 2.10 – Estimated kappa values ($\hat{\kappa}$) for the three classifiers (IB7, J48, and MLP20) applied to the MCIs with (black) orthonormal and (gray) regular coefficients.

In order to analyze the performance of STARS in terms of representing the full information content of the whole multitemporal-multispectral Landsat image data set, we performed each of the three classifiers on all 66 georeferenced surface reflectance wavebands (11 images, each with six spectral wavebands; see Table 2.1). The estimated kappa values of the three classifications were IB7 = 0.998, J48 = 0.975, and MLP20 = 0.961. In terms of the *Z*-test, the IB7 classification using 66 wavebands presented similar results for the classification using MCI from the MCI 06 on, for orthonormal coefficients (i.e., MCI 06, MCI 07, ..., MCI 10). The decision-tree classifier (J48) presented similar results for MCI 08, MCI 09, and MCI 10. Eventually, the neural network (MLP20) presented similar results from the MCI 06 on, for the orthonormal coefficients, and from MCI 17 on, for the regular ones (i.e., from MCI 17 to MCI 20). These results suggest that, depending on the choice of the model used to fit the spectral-temporal response surfaces (e.g., PTS), STARS does not lead to a loss of information but represents and also synthesizes the spectral changes over time inside the synthetic bands of the MCI.

To examine STARS in a wider context, its performance was also confronted with a classification considering the normalized difference vegetation index (NDVI) (ROUSE

JR *et al.*, 1973). Eleven NDVI images were computed (i.e., one for each Landsat image, see Table 2.1) and used as input for the IB7 classifier. The kappa value, estimated at 0.909, attested that this classification was statistically less accurate than the classifications using MCI from the MCI 03 on, for orthonormal coefficients. This result means that the ten synthetic bands of MCI 03 (see Table 2.3) performed considerably more effectively than the 11 NDVI images when used as input for the IB7 classifier. Indeed, the main disadvantage of using vegetation indices compared with the use of STARS is that vegetation indices usually account for only two or three wavebands, whereas STARS exploits all available spectral wavebands of the multitemporal data set.

Furthermore, the IB7 classifier was also performed using as input attributes the results of a principal component analysis (PCA) (JOLLIFFE, 2002) conducted upon the 66 georeferenced surface reflectance wavebands. The main characteristic of PCA is to capture and represent the great majority of variability in multilayer data (i.e., the useful information) within the first few principal components (PCs), and as such, it has been widely used for dimensionality reduction in remote sensing (FARRELL; MERSEREAU, 2005). We performed 66 classifications: the first classification using only the first PC, the second classification using the first two PCs, and so on, until the 66th classification that used all 66 PCs. The kappa value considerably increased from the first classification until the classification using ten PCs and then stabilized at around 0.995. According to *Z*-tests, this value represented similar performance to the IB7 classifications using MCI 05 on, for orthonormal coefficients. Despite the high accuracy values presented in the classifications using PCA, the performance of PCA depends on the data distribution becoming worse when the data set is not normally distributed (JIMENEZ; LANDGREBE, 1999; LEE *et al.*, 2010; YATA; AOSHIMA, 2009), whereas STARS is able to deal with practically any kind of data distribution. Moreover, some models used in STARS (e.g., PTS) minimize problems associated with aberrant or noisy data (VIEIRA, 2000), whereas PCA is significantly sensitive to outliers (YANG *et al.*, 2008). The results suggest, therefore, that the choice of a suitable model (e.g., PTS with $d = 5$ and orthonormal coefficients) for STARS can reduce dimensionality representing the data variability in the synthetic bands of the MCI. Complementarily, the use of feature selection techniques can reduce k dramatically without decreasing kappa values (VIEIRA, 2000).

Other studies have explored alternative techniques to automate the process of mapping sugarcane harvest in Brazil. For instance, Mello (2009) compared the performance of the maximum likelihood classifier using a range of different input attributes, including fitted coefficients from spectral-temporal response surfaces, fraction images derived from linear mixture models (SHIMABUKURO; SMITH, 1991), and the full set of spectral wavebands from a multispectral-multitemporal Landsat data set. He concluded that the classification upon the fitted coefficients from spectral-temporal response surfaces showed better results in terms of accuracy than the classifications using the other input attributes. Mello *et al.* (2010b) also used fraction images derived from linear mixing models as input for the maximum likelihood classifier, and although a relatively high overall accuracy index was found (≈ 90%), the authors pointed out the gap for alternative methods to handle with spectral-temporal dynamics in face of the difficulty in defining suitable endmembers for the solution of mixing models. Similarly, Aguiar *et al.* (2009) used fraction images derived from linear mixing models in a decision-tree procedure to identify sugarcane harvest for the entire São Paulo State. Once again, although the work found a harvested area with 97.7% in accordance with the São Paulo State Environmental Secretary data for the 2006/2007 crop year, the authors found difficulties in defining suitable endmembers.

Alternatively, El Hajj *et al.* (2009) proposed an approach using high spatial resolution multitemporal images to detect sugarcane harvest on Reunion Island, which is an overseas department of France in the Indian Ocean. However, the success of this method is highly dependent on the integration of crop model outputs with expert knowledge, such as the understanding of sugarcane physiology or cultural practices that can vary considerably over different regions or different timescales (e.g., crop growing and harvesting seasons) (XAVIER *et al.*, 2006). This makes its uptake impractical where expert knowledge is limited, as is often the case. Indeed, despite the great overall potential of remote sensing for agricultural applications (VIEIRA, 2000) including sugarcane agriculture (ABDEL-RAHMAN; AHMED, 2008), there remains a considerable need for more robust and widely applicable models (LIN *et al.*, 2009).

In fact, the high kappa values presented in Table 2.4 and Fig. 2.10 indicated that STARS was effective in describing the spectral-temporal change associated with either

BH or GH. Aguiar *et al.* (2011) mapped sugarcane harvest practices during five crop years in São Paulo State, Brazil, to evaluate the "Green Ethanol" Protocol based on visual interpretation of Landsat-type images. In the last evaluated year (2010), about 4.7 million ha of sugarcane were harvested, which means that visual interpretation of the harvest type is a very substantial undertaking, and an alternative automated procedure would assist the task significantly. STARS seems to offer such an alternative.

In order to evaluate statistically the classification accuracy values, *Z*-tests for each single classification were performed, and these indicated that all classifications were significant and, therefore, different from a random classification, at $\alpha = 5\%$. Furthermore, pairwise *Z*-tests were performed to compare accuracy values of all classifications. Results are presented in the following three subsections.

2.5.3.1. Orthonormal versus regular coefficients

The results of the pairwise *Z*-tests comparing the accuracy of classifications based on orthonormal and regular coefficients are presented in Table 2.5.

Table 2.5 – *p*-values of the pairwise *Z*-tests comparing the classifications based on orthonormal and regular coefficients.

Orthonormal x Regular	**IB7**	**J48**	**MLP20**
MCI 01 x MCI 11	.933	.356	.154
MCI 02 x MCI 12	≈ 0	≈ 0	.793
MCI 03 x MCI 13	≈ 0	.020	.749
MCI 04 x MCI 14	≈ 0	.001	.003
MCI 05 x MCI 15	≈ 0	.001	.934
MCI 06 x MCI 16	≈ 0	≈ 0	.186
MCI 07 x MCI 17	.001	≈ 0	.099
MCI 08 x MCI 18	≈ 0	≈ 0	.315
MCI 09 x MCI 19	.004	≈ 0	.261
MCI 10 x MCI 20	.233	.008	.208

Gray cell indicates significance at $\alpha = 5\%$.

According to Tables 2.4 and 2.5, for the three classifiers tested, all classifications based on orthonormal coefficients presented accuracy values equal to or greater than classifications based on regular coefficients. In fact, orthonormal coefficients are expected to be better for distinguishing thematic classes since they tend to present little or no correlation between themselves, whereas the regular coefficients are generally more correlated with each other (MATHER, 1976). The pairwise *Z*-test presented in Table 2.5 revealed that for the IB7 and J48 classifiers, the orthonormal coefficients from the PTS model actually performed better than the regular coefficients for all MCIs, except the simplest ones. Indeed, the simplest MCIs for orthonormal (MCI 01) and regular (MCI 11) coefficients were not statistically different in terms of accuracy for all three classifiers. For the MLP20 classifier, all comparisons indicated that they were not significantly different, except for MCI 04 and MCI 14 ($p = 0.003$). The most complex MCI (MCI 10 and MCI 20) from the CS model were significantly different in terms of accuracy only for the J48 classifier ($p = 0.008$) with the best results using the orthonormal coefficients.

Considering that the model created by the neural-network classifier (MLP20) uses hyperplanes for the data separation (classification) and these hyperplanes can be in any possible direction (FAUSETT, 1993; LOONEY, 1997), its performance is not significantly improved using the orthonormal coefficients. For the decision-tree classifier (J48), which makes orthogonal separations (QUINLAN, 1993), the classification performance is improved when using orthonormal coefficients instead of regular ones. The performance observed for the instance-based classifier (IB7) can be attributed to the smaller variance of the classes when orthonormal coefficients are used.

Thus, based on the results, we suggest the use of the orthonormal coefficients instead of the regular ones in order to improve classification performance and avoid both unstable computational solution (MATHER, 1976) and multicollinearity (KUTNER *et al.*, 2005).

2.5.3.2. Multi-Coefficient Image complexity

Pairwise *Z*-tests were performed on each classification to evaluate differences in κ according to the complexity of the MCI used. The tests were conducted considering orthonormal or regular coefficients.

Orthonormal: For the classifications based on the ten MCIs with orthonormal coefficients (from MCI 01 to MCI 10), the IB7 presented no significant differences from the MCI 05 on. Moreover, the MCI 04 also presented good accuracy but was slightly less accurate than MCI 08 (p = 0.004) and MCI 10 (p = 0.048). No significant differences were found from MCI 06 on, for the J48 classifier. J48 also presented no differences among MCI from MCI 04 to MCI 07 and from MCI 05 to MCI 08. Eventually, the MLP20 classifier presented the same accuracy from MCI 06 on.

Regular: For the ten MCIs with regular coefficients (from MCI 11 to MCI 20), the IB7 presented the best results for MCI 18 and MCI 20 (p = 0.075). There were no significant differences detected among MCIs from MCI 14 to MCI 16; from MCI 15 to MCI 17; among MCI 16, MCI 17, and MCI 19; and also from MCI 17 to MCI 19. On the other hand, the J48 classifier presented equivalent accuracy from MCI 14 on for the PTS model. However, the best performance of the J48 was presented for the CS model (MCI 20) with $\hat{\kappa}$ = 0.9500. Moreover, there was no significant difference between MCI 13 and MCI 14, even though its Z-test p-value was close to 5% (p = 0.059). Eventually, the MLP20 presented the best accuracy for MCI 19 and MCI 20 (p = 0.172).

In general, as shown in Fig. 2.10, accuracy values tended not to rise after a certain number of synthetic bands (k) were used as input for classifiers. In terms of classifications based on MCIs with orthonormal coefficients, the IB7 classifier, for example, did not improve accuracy from MCI 05 on (i.e., it presented the same accuracy for classifications based on MCIs with 21, 28, 36, 45, 55, and 66 synthetic bands). The same was observed for the J48 and MLP20 classifiers, which did not improve accuracy from MCI 06 on. These results show that STARS can also be used to reduce the dimensionality of multitemporal-multispectral data sets without significant loss of information. Moreover, for operational purposes, the results indicate that sugarcane harvest monitoring can be carried out using STARS with a PTS model and a suitable degree (e.g., MCI 05) that demands less computational processing than the more complex models.

2.5.3.3. Comparing classifiers

Pairwise *Z*-tests were also performed comparing the accuracy values of the three classifiers for each MCI. The results for MCIs with orthonormal coefficients are presented in Table 2.6.

Table 2.6 – *p*-values of the pairwise *Z*-tests comparing the accuracy values of the three classifiers for each MCI with orthonormal coefficients.

MCI #	IB7 x J48	IB7 x MLP20	J48 x MLP20
MCI 01	.900	≈ 0	≈ 0
MCI 02	≈ 0	≈ 0	≈ 0
MCI 03	≈ 0	≈ 0	≈ 0
MCI 04	≈ 0	≈ 0	.003
MCI 05	≈ 0	≈ 0	.098
MCI 06	≈ 0	≈ 0	.607
MCI 07	≈ 0	≈ 0	.476
MCI 08	≈ 0	≈ 0	.589
MCI 09	≈ 0	≈ 0	≈ 1
MCI 10	≈ 0	≈ 0	.102

Gray cell indicates significance at $\alpha = 5\%$.

In terms of orthonormal coefficients, according to the pairwise *Z*-tests (see Table 2.6), IB7 and J48 presented the same accuracy only for the simplest MCI (MCI 01) with $p = 0.900$, but they were more accurate than MLP20. For all other MCIs with orthonormal coefficients, IB7 presented more accurate indices than J48 and MLP20 classifiers. The best performance observed for the instance-based classifier (IB7) can be attributed to its ability to deal with practically any kind of data distribution and to the fact that it is based on the instances themselves, instead of a model derived from labeled instances. However, other parameters such as the k-nearest neighbor number should be tested to confirm the superiority of the instance-based classifier.

The J48 classifier presented greater accuracy than MLP20 when classifications were performed with MCIs from MCI 01 to MCI 04. For the more complex MCIs with orthonormal coefficients, these two classifiers presented equivalent accuracy values.

For all MCIs with regular coefficients, IB7 also presented stronger results than the other classifiers, except for the simplest one (MCI 11) where J48 presented accuracy equivalent to IB7 ($p = 0.257$). J48 also presented accuracy values greater than MLP20 for MCIs from MCI 11 to MCI 14, whereas MLP20 was more accurate than J48 for MCIs 16, 18, 19, and 20. For MCIs 15 and 17, these two classifiers presented equivalent accuracy values.

According to the accuracy assessment analyses comparing classifications of MCIs based on orthonormal and regular coefficients, from the simplest to the most complex model, and evaluating three different classifiers, we suggest using MCI 05 with IB7 classifier for operational sugarcane harvest monitoring in Brazil with STARS. Table 2.7 shows the confusion matrix regarding the accuracy assessment of the classification by the IB7 classifier running upon MCI 05. Both inclusion and omission errors were distributed over BH, GH, and UH classes, indicating that the classifier did not tend to misclassify any particular harvest type. The overall accuracy index of 99.22% confirms the high accuracy of this classification. Moreover, the ability to distinguish the harvest type (BH or GH) is even better if we consider the fact that in approximately 44% of (inclusion/omission error) cases, the classifier chose the right class in terms of harvest type (BH or GH) and only mistook the harvest date.

Table 2.7 – Confusion matrix for the classification by the IB7 classifier running upon MCI 05.

		Reference																						
		BH01	BH02	BH03	BH04	BH05	BH06	BH07	BH08	BH09	BH10	BH11	GH01	GH02	GH03	GH04	GH05	GH06	GH07	GH08	GH09	GH10	GH11	UH
Classified	BH01	50																						
	BH02		50																					
	BH03			50											1									
	BH04				50	1																		
	BH05					47																		
	BH06					1	50																	1
	BH07					1		50																
	BH08								49															
	BH09									50														
	BH10										50													
	BH11											50												1
	GH01												50											
	GH02													50										
	GH03														49									
	GH04															49								
	GH05																50							
	GH06																	50						
	GH07																		50					
	GH08																			49				
	GH09								1											1	50			
	GH10																					50		
	GH11																						50	
	UH															1								48

Descriptions of the 23 thematic classes are presented in Table 2.2. Empty cells are equal to zero.

2.6. Conclusion to STARS

The STARS method enables representation of the entire information content of a multitemporal-multispectral remote sensing data set in a single MCI. Using a case study of sugarcane harvest, it is shown that STARS holds considerable potential for representing spectral change over time of features on the Earth's surface. Indeed, the example presented in this paper demonstrates that this method could be introduced to automate regional agricultural monitoring activities such as sugarcane harvest classification.

Two models were tested in modeling spectral–temporal response surfaces. The PTS model presented smooth spectral-temporal surfaces that can be effective in describing gradual change on the Earth's surface such as crop phenology. In contrast, the CS model presented sharper and more defined spectral-temporal surfaces, useful for characterizing abrupt change. However, our results showed that abrupt changes related with the sugarcane harvest event were well characterized also with the PTS model when a suitable degree was set. Two types of coefficients were tested, and of these, orthonormal coefficients performed more accurately than the regular ones when using MCI for classification purposes.

Advantages of STARS include that the method can provide a description of features' spectral change over time; that image data from different sensors with varying spectral wavebands and irregular time intervals can be used; that the method is robust, enabling different model options according to the application; and for some models (e.g., PTS), that it is economical, as the number of coefficients is smaller than the sum of the spectral wavebands. Moreover, the synthetic bands of the MCI can be used as input features for a range of operations, including image classification, visual interpretation, and creating spectral-temporal indices.

The STARS algorithm has been implemented in R software (R CORE TEAM, 2013) and can be found at www.dsr.inpe.br/~mello.

Acknowledgment

The authors would like to thank Dr. M. J. Canty, from the Julich Research Center, for helping with IR-MAD radiometric normalization; Dr. I. D. Sanches, from the Institute of Geosciences of the State University of Campinas (Unicamp), and Dr. A. R. Formaggio, from INPE, for providing the ground data set; M.Sc. W. F. Silva, from the Brazilian National Water Agency (ANA), for his support in the GPS field work for georeferencing; SPOT Image from France and Brazil for providing, through Dr. J. V. Soares' contact, the SPOT 5 image; Dr. Y. E. Shimabukuro and Dr. C. D. Rennó, from INPE, for their general advice; the team of the Laboratory of Remote Sensing in Agriculture and Forestry (LAF) of INPE; and the anonymous referees for their valuable and constructive comments on the original manuscript.

3 BayNeRD: plausible reasoning from observations[5]

Abstract:

This paper describes the basis functioning and implementation of a computer-aided Bayesian Network (BN) method that is able to incorporate experts' knowledge for the benefit of remote sensing applications and other raster data analyses: Bayesian Network for Raster Data (BayNeRD). Using a case study of soybean mapping in Mato Grosso State, Brazil, BayNeRD was tested to evaluate its capability to support the understanding of a complex phenomenon through plausible reasoning based on data observation. Observations made upon Crop Enhanced Index (CEI) values for the current and previous crop years, soil type, terrain slope and distance to the nearest road and water body were used to calculate the probability of soybean presence for the entire Mato Grosso State, showing strong adherence to the official data. CEI values were the most influencial variables in the calculated probability of soybean presence, stating the potential of remote sensing as a source of data. Moreover, the overall accuracy of over 91% confirmed the high accuracy of the thematic map derived from the calculated probability values. BayNeRD allows the expert to model the relationship among several observed variables, outputs variable importance information, handles incomplete and disparate forms of data, and offers a basis for plausible reasoning from observations. The BayNeRD algorithm has been implemented in R software and can be found on internet.

[5] This chapter is an adapted version of the paper:
Mello, M. P.; Risso, J.; Atzberger, C.; Aplin, P.; Pebesma, E.; Vieira, C. A O.; Rudorff, B. F. T. Bayesian Network for Raster Data (BayNeRD): plausible reasoning from observations. **Remote Sensing**, *(submitted).*

3.1. Introduction to BayNeRD

Understanding complex phenomena in the field of Earth observation sciences represents a considerable challenge for scientific analysis (DONNER *et al*., 2009; MELESSE *et al*., 2007). Regarding investigation of large scale phenomena, great progress has been achieved through recent advances in spaceborne remote sensing data acquisition (LI *et al*., 2008), together with the availability of high performance computing for remotely sensed data analysis (LEE *et al*., 2011). To Lu and Weng (2007) the most important factors driving the success of an inference based on remotely sensed data are: (i) the availability of high-quality observations (e.g., accurate imagery corrected for atmospheric effects and ancillary data such as topography, soil, road and census data); (ii) the design of a suitable analytical procedure; and (iii) the analyst's skills and knowledge. However, some phenomena are often too complex to be investigated by conventional methods (RICHARDS, 2005), demanding new computer aided methods to help characterise phenomena through plausible reasoning inferences based on consistent data observations (i.e., evidence).

Interactions of probabilities have been identified as the most promising way for a computer to effect plausible reasoning (JAYNES, 2003). The Bayes' theorem updates the knowledge (*prior probability*) of a specific event in the light of new/additional evidence (*conditional probabilities*), allowing one to have a plausible reasoning based on a degree of belief (*posteriori probability*) (McGRAYNE, 2011). Thus, observations made upon variables that are related to a particular phenomenon can be used to develop plausible reasoning about the phenomenon, its causes and consequences (JAYNES, 2003). When the number of variables increases or even when the complexity of the interactions among the variables involved in a phenomenon rises, the Bayesian Network (BN) is a representation suited to models and handles such tasks (JENSEN; NIELSEN, 2007; PEARL, 1988).

Neapolitan (2003) defines BNs as graphical structures for representing the probabilistic relationship among a set of variables via a Directed Acyclic Graph (DAG), and for calculating probabilistic inference with those variables. BNs can also be defined as representational structures that are meant to organize one's knowledge about a particular phenomenon into a coherent whole (DARWICHE, 2009). The advantages of

BNs are that they: (i) can deal with a large number of variables and can also handle incomplete data sets (i.e., missing data); (ii) can deal with both numeric and categorical data simultaneously; (iii) are able to incorporate experts' knowledge via a participatory modelling procedure of causal relationships; and (iv) are easy to understand and visualize through DAGs (HECKERMAN, 1997; UUSITALO, 2007). Notwithstanding these advantages, BNs have rarely been used in the field of Earth sciences and remote sensing, and their potential is, as yet, largely unexploited (AGUILERA *et al*., 2011).

Although researchers have made substantial advances in developing the theory and application of BNs (NEAPOLITAN, 2003), the actual use of these networks often remains a difficult and time-consuming task (AGUILERA *et al*., 2011). In the Earth sciences, where investigations commonly involve numerous layers of data (e.g., maps and images), analysis can be difficult because of the need to know both the relationships among the variables (i.e., conditional (in)dependences) and their probability functions. In addition, tasks can be time-consuming because they are typically performed manually. Until now, only a limited number of computer aided methods have been implemented. Therefore, there is potential for the use of probability theory as a basis for computer aided plausible reasoning, and BNs as a tool for representing and computing probabilistic beliefs in the field of Earth sciences (UUSITALO, 2007). Moreover, there is demand for the development and implementation of computer aided methods that offer a basis for Earth science researchers to understand and model phenomena through plausible reasoning inferences based on data observations (AGUILERA *et al*., 2011).

The aim of this paper is to describe, implement and test a computer aided BN method that is able to incorporate experts' knowledge for the benefit of remote sensing applications and other raster data analyses. The freely available algorithm is named Bayesian Networks for Raster Data (BayNeRD). Following development of the approach, BayNeRD was tested on a case study for soybean identification and mapping in Mato Grosso State, Brazil. The test enabled evaluation of the capability of BayNeRD to support the understanding of a complex phenomenon through plausible reasoning based on data observation.

3.2. Bayesian Networks

A BN for a set of *n* variables consists of: (i) a network structure, graphically represented by a DAG with nodes and arcs, that encodes a set of conditional (in)dependence assertions about the variables; and (ii) a set of probability functions associated with each variable (NEAPOLITAN, 2003). We use upper-case letters (e.g., V_1, V_n) to denote both a variable and its corresponding node, and the same but lower-case letters (e.g., v_1, v_n) to denote the state or value (defining a particular instantiation) of the variable. Then, the joint probability distribution for any particular instantiation of all *n* variables in a BN is given by

$$P(V_1 = v_1, \dots, V_n = v_n) = \prod_{i=1}^{n} P(V_i = v_i \mid \Phi_i = \phi_i) \tag{3.1}$$

where v_i represents the instantiation of variable V_i and ϕ_i represents the instantiation of its parents Φ_i, with *i* varying from 1 to *n*. Parent variables are those whose instantiations directly influence other, descendent variables. The arcs (represented by arrows in the DAG) encode the conditional dependencies (i.e., which variables are parent/descendant of other variables) (NEAPOLITAN, 2003; PEARL, 1988).The joint probability of any instantiation of all the variables in a BN can be computed as the product of only *n* probabilities. Thus, we can determine any probability of the form

$$P(V_1|V_2, \dots, V_n) \tag{3.2}$$

where V_i are sets of variables with known values (v_i – i.e., instantiated variables). This ability to compute posterior probabilities given some evidence is called inference. In the case of using Eq. (3.2) for inferences about certain phenomena using BayNeRD, we named the variable that represents the phenomenon as the *target variable* and the variables that can be used to describe an outline of the phenomenon as *context variables* (i.e., those variables that are somehow related to the phenomenon).

To illustrate the concept, suppose we are interested in inferring soybean occurrence based on observations of other variables. It is well known that soybean plantations have certain peculiarities (GARRET *et al.*, 2013), such as: (i) it is preferably not sown in areas with steep terrain slope because mechanization may be hindered, and (ii) it is preferably sown in soils that are apt for agricultural cultivation. Then, *soybean occurrence* (*S*) is the target variable with a Binomial statistical distribution and can be represented by a thematic map with the classes soybean and non-soybean, that mean soybean presence (S = 1) and soybean absence (S = 0), respectively. On the other hand, *terrain slope* (*T*) and *soil aptitude* (*A*) could be, in our example, two context variables. Since we are interested to infer about *S*, in this example, Eq. (3.2) becomes

$$P(S|T,A). \tag{3.3}$$

Indeed, *T* and *A* directly influence *S* and so are said to be parents of *S*. Moreover, since soil formation processes are strongly influenced by terrain slope (PARK *et al.*, 2001), *T* also influences *A* and, therefore, *T* is also a parent of *A*. These (in)dependence relationships among the variables are represented by a DAG as shown in Fig. 3.1.

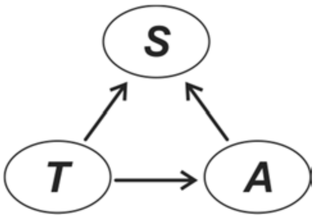


Figure 3.1 – Directed Acyclic Graph (DAG) representing a hypothetical BN graphical model where the target variable *soybean occurrence* (*S*) is influenced by two context variables: *terrain slope* (*T*) and *soil aptitude* (*A*). Since soil formation processes are strongly influenced by terrain slope, *T* is also parent of *A*. Variables are represented by nodes and dependences are represented by arcs between pairwise nodes.

The representation of conditional (in)dependencies is the essential function of BNs. For each node in a BN structure, there is a conditional-probability function that relates this node to its immediate parents. If a node has no parents (e.g., *T*) then a prior-probability function is specified (JENSEN; NIELSEN, 2007). Eventually, once all probability functions are specified, it is possible to compute the probability of soybean presence (S = 1) in a certain area based on the observed values for both *T* and *A* in the same area.

In practical terms, the definition of these probability functions is often the most complicated part of BN modelling. However, the empirical Bayesian approach suggests that the functions can be defined based on observations, i.e., from the data (COOPER; HERSKOVITS, 1992). Mello *et al.* (2010a) proposed use of pixel counting in discretized variables to compute probability functions in a BN when employing raster data (described further).

Aware of the great demand for implemented computer algorithms to help handle and understand phenomena in the field of Earth observation science, we implemented BayNeRD in R software (R CORE TEAM, 2013). The algorithm provides researchers a means of modelling any phenomenon of interest, whereby plausible reasoning inferences are made based on observations stored in raster data format.

3.3. Framework of the implemented BayNeRD algorithm in R software

R software was used to implement BayNeRD because it is a high-level language and environment for data analysis and graphics. It is growing in popularity and uptake, and is freely available for the research community (CRAWLEY, 2007). Furthermore, among all packages already implemented in R software, there are several developed for both handling spatial data (BIVAND *et al.*, 2008) and computing Bayesian analysis (ALBERT, 2009), especially catnet (BALOV; SALZMAN, 2011) which was designed for categorical BN.

The BayNeRD algorithm handles data in the GeoTIFF format, which has been widely used to represent raster data with geographical coordinates. For use in BayNeRD all raster data (i.e. one GeoTIFF representing each variable) must represent the same geographic area. Each GeoTIFF corresponds to a variable (node) used in the BN model. These variables and their (in)dependence relations are used to compute the probability functions.

3.3.1. Target variable

The variable which directly represent the phenomenon is called the target variable. A GeoTIFF with data representing the target variable as *reference data for training* must be provided. It is later used in the definition of the probability functions. The GeoTIFF representing the target variable usually has four labels representing the following thematic classes: (i) target presence observed; (ii) target absence observed; (iii) missing data, i.e., no observations were made; and (iv) pixels outside the study area. The latter is simply used to mask out any pixels that are outside the study area from any of the raster data layers to be used in BayNeRD. Although *reference data for training* may contain more than these four labels, it must have at least two: (i) and (ii). Thus, the target variable, represented in the general model as Y, is expected to have a Binomial statistical distribution that can be instantiated ($Y = y$) with y assuming either 1 for the target presence or 0 for the target absence.

3.3.2. Context variables

The context variables are those that exhibit any kind of relationship with the target variable (such as *terrain slope* and *soil aptitude*, as previously discussed). Moreover context variables may exhibit relationships among themselves, such as the *terrain slope* influencing the *soil aptitude* due to the influence of slope in soil formation processes (PARK *et al.*, 2001). The context variables may contain any sort of observations such as numerical values (e.g., *terrain slope* given in percentage) or categorical data (e.g., thematic classes representing *soil aptitude* for agriculture cultivation). Moreover, the context variables may also contain missing data.

One of the main difficulties of using BNs for real problems is the definition of the probability functions of the model (COOPER; HERSKOVITS, 1992). Therefore BayNeRD was developed to interact with the user to define, through discretization processes, the probability functions of the model based on both observed data and users' knowledge about the phenomenon of interest. Discretization is the process of representing (approximating) the observed values of a variable using discrete quantities (e.g., intervals, such as in the process of drawing a histogram).

After the target variable has been entered as reference data for training and the context variables have been read, the user will be able to design the BN graphical model.

3.3.3. Designing the Bayesian Network graphical model

To design the BN graphical model the user is asked about the (in)dependence relations among all variables read (i.e., both target and context variables). Since the dependencies are represented by arcs in a DAG, BayNeRD asks whether an arc exists between pairwise variables. For example, if the *terrain slope* (*T*) influences *soil aptitude* (*A*), and both *T* and *A* influence *soybean occurrence* (*S*), there will be an arc from *T* to *A*, an arc from *T* to *S* and another arc from *A* to *S* (see Fig. 3.1).Once the graphical representation of the BN model is defined stating the variables and their (in)dependence relations, BayNeRD is able to compute the probability functions, which is done based on pixel counting in discretized variables (MELLO *et al.*, 2010a).

3.3.4. Discretization and probability functions

The discretization divides the range of the observed values for a variable into intervals and codes the values in the variable according to which interval they belong. In BayNeRD the discretization is based on choosing the number of intervals defined for each context variable and can be computed following three implemented criteria: (i) equidistant intervals, where each interval has the same width; (ii) quantiles, where each interval tends to have the same number of elements (i.e., pixels); and (iii) manually defined intervals, where the user defines the upper and lower limits of each interval.

The discretization will have an impact on the computed probability functions. These probabilities are computed through pixel counting according to both the (in)dependence relations defined in the BN graphical model and the intervals defined in the discretization processes. Indeed, both the definition of the BN graphical model and the discretization processes enable users to add their knowledge about the phenomenon into the model. The more a data set is accurate and a user is skilled in defining both BN graphical model and interval limits during discretization processes, the more the data-based probability functions computed are representative of the real probability functions (MELLO *et al.*, 2010a).

Let us suppose that the *terrain slope* (*T*), which does not have parents in the designed BN model represented in Fig. 3.1, was discretized using four equidistant intervals between 0 and 100%. By dividing the number of pixels with values lower than 25% by the total number of pixels observed for the study area one can compute the probability for the first interval of the discretized *T*. The probabilities for the remaining intervals of the discretized *T* are computed by pixel counting as described above and the probabilities for all intervals must sum to 1. Indeed, such as for *T*, for all variables that do not have parents in a designed BN these probabilities define the prior-probability function. In the case of variables that have parents, such as the *soil aptitude* (A), which is a descendent of T (Fig. 3.1), BayNeRD uses the intervals defined for T and the ones defined for A to compute the conditional probability function for A in the BN model, also based on pixel counting (MELLO *et al.*, 2010a).

The user should be sufficiently expert to define suitable discrete intervals for each context variable so that all scenarios (i.e. combination of parents' and variable's intervals) have representative data to compute probability functions, where a minimum user-defined quantity of pixels is considered as a representative number. The process of computing the probability functions of the model is called training, when BayNeRD defines the probability functions based on the observed values from the data (i.e., by counting pixels). Using values of probability for plausible reasoning, BNs are able to infer based on evidence (observed data). Indeed, once BayNeRD is trained, it is able to answer the question: "what is the probability of target presence (e.g., *soybean*), given the observed values for the context variables (e.g., *terrain slope* and *soil aptitude*)?". When the probability that answers this question is calculated for every pixel in the entire study area, a Probability Image (PI) is formed.

3.3.5. Computing the Probability Image

The PI consists of a raster data (i.e., a matrix matching the same coordinates of the entered *reference data for training*) where each pixel contains the probability of presence of the target given the values observed (instantiations) for the input variables, i.e.,

$$P(Y = 1|V_1 = v_1, \dots, V_n = v_n) \tag{3.4}$$

If any context variable presents missing data for any specific pixel in the study area, it is considered as "unobserved" in the model but Eq. (3.4) is computed anyway. It is also possible to find P(Y = 1) for pixels where no observation was made for any context variable. In this case, the computed probability will be the marginal probability for Y when Y = 1.

BayNeRD also allows the user to quantify the influence of each context variable on the probabilities computed for the target variable. This is done through the Kullback-Leibler (KL) divergence, which is a non-symmetric measure of the difference between two probability distributions (KULLBACK; LEIBLER, 1951). Thus, it is possible to measure how much V_1, V_2, … and V_n individually influence the probability computed for Y by computing KL divergences between conditional and marginal probabilities in the BN model.

The main result of BayNeRD is the PI and it can be used in several applications. For example, the PI can be used to generate a thematic map with classes target and non-target (e.g., soybean and non-soybean) just by slicing the PI using a limiting probability value named the Target Probability Value (TPV). Thus, by setting TPV at 50%, for instance, all pixels with values equal to or greater than 0.5 in the PI will be labelled as target and the remaining pixels (with values smaller than 0.5) will be labelled as non-target. But what if the best TPV was 70% instead of 50%? Or even 80%?

3.3.6. Selecting the Target Probability Value

Apart from a user-defined value, six criteria are implemented in BayNeRD to select the TPV which best meets a chosen criterion, making use of available reference information (i.e., a *reference data for testing*). These implemented criteria are: (i) nearest 100% sensitivity and 100% specificity point (ZWEIG; CAMPBELL, 1993) (for a description of these two indices see Altman and Bland (1994)); (ii) minimum difference between sensitivity and specificity; (iii) highest overall accuracy index; (iv) highest kappa index (COHEN, 1960; HUDSON, 1987); (v) most similar area (number of pixels) matching the *reference data for testing*; and (vi) minimum difference between omission and commission errors (CONGALTON; GREEN, 2009).

3.4. Case study of soybean mapping in Brazil: Material and research methods

The case study involves soybean identification and mapping in Mato Grosso, which is a major Brazilian soybean producer (about 30% of the total domestic production) and an important global hub for tropical agricultural production (CONAB, 2013). Mato Grosso State is located in the Southwest of Legal Brazilian Amazon encompassing an area around 900,000 km^2 (BRASIL, 2002). Fig. 3.2 shows the location of Mato Grosso State, highlighting thirty 30 x 30 km plots (and the Landsat path/row covering them) of reference data produced by Epiphanio *et al.* (2010) for the crop year 2005/2006 (i.e., from August 2005 to July 2006) using visual interpretation of Landsat-5/TM images and field data. Additional data such as indigenous lands, conservation units, mapped forests and floodplains were used to mask out areas of no interest for mapping soybean (as will be described further).

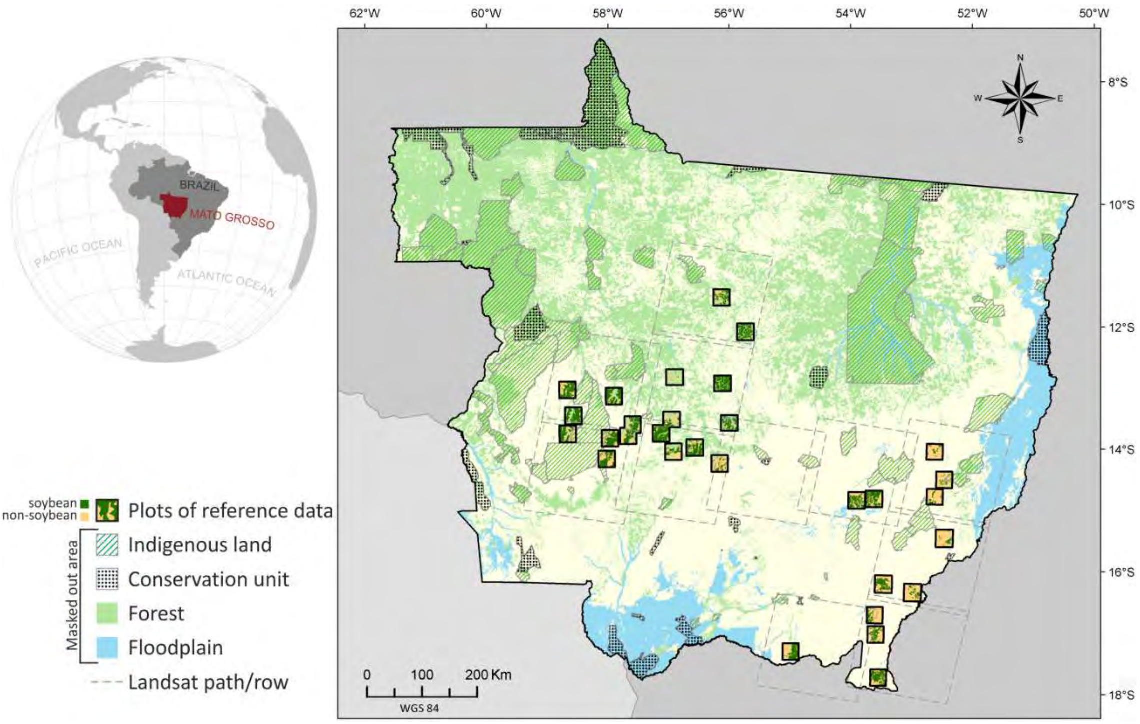


Figure 3.2 – Study area corresponding to the Mato Grosso State, Brazil. The analysis was only performed in areas that were not masked out.

Although Brazil is the second largest producer of soybean worldwide (FAO, 2012), the country does not have a systematic nationwide mapping system for this oilseed. Tabulated agricultural statistics at municipality level are only released with a delay of about two years after harvest. The absence of timely and spatial data restricts investigations related to crop monitoring and forecast. It also hinders the monitoring of the possible spread of this crop into new, sometimes environmentally-sensitive, areas. As such, there is demand for the use of satellite sensor images as an accurate, efficient, timely and cost-effective way to monitor agricultural crops (ATZBERGER, 2013). Several studies have demonstrated the value of Landsat-like images to monitor agricultural crops in Brazil using visual interpretation (RIZZI; RUDORFF, 2005; RUDORFF *et al.*, 2010) or even automatically (MELLO *et al.*, 2013b; VIEIRA *et al.*, 2012). However, these methods have certain constraints, notably the limited number of cloud-free images that are routinely acquired during the crop cycle (ASNER, 2001; SANO *et al.*, 2007). Alternatively, multitemporal approaches using Moderate Resolution Imaging Spectroradiometer (MODIS) time series images have been successfully used to monitor soybean plantations in tropical regions such as Mato Grosso, since the 1-2 day temporal resolution of MODIS minimizes the constraints related to cloud coverage on satellite sensor images (ARVOR *et al.*, 2011; MACEDO *et al.*, 2012; MORTON *et al.*, 2006).

Besides remotely sensed spectral and temporal information, several other context variables are closely related with soybean occurrence in a given field (e.g., soil type and infrastructure facilities) (GARRETT *et al.*, 2013). In the present study, this information is combined within a BN structure to optimize soybean identification and mapping. Fig. 3.3 shows a flowchart summarising the research material and methods employed in the soybean mapping case study application of BayNeRD.

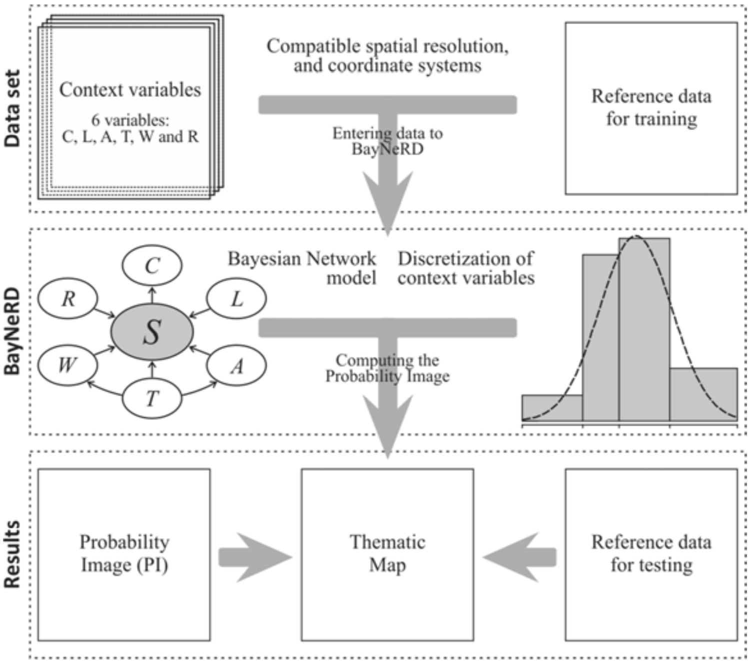

Data set
Context variables
6 variables:
C, L, A, T, W and R
Compatible spatial resolution, and coordinate systems
Entering data to BayNeRD
Reference data for training
BayNeRD
C
R
L
S
W
A
T
Bayesian Network model
Discretization of context variables
Computing the Probability Image
Results
Probability Image (PI)
Thematic Map
Reference data for testing

Figure 3.3 – Summary of the procedures used in the case study of applying BayNeRD to identify soybean plantations in Mato Grosso State, Brazil. Table 3.1 provides a description of the variables used.

In summary, six context variables and a reference thematic map were used as inputs in BayNeRD, where a BN model was defined based on experts' knowledge. Probability functions were computed based on pixel counting of discretized variables, allowing BayNeRD to compute the PI, which was eventually used to produce a thematic map of soybean occurrence over the study area. This thematic map was then assessed using reference data. The following subsections describe the research materials and methods in detail.

3.4.1. Variables

All variables used in this case study, each represented by a raster GeoTIFF, were resampled to match the grid of the MODIS vegetation indices product (MOD13Q1), with a nominal spatial resolution of 250 x 250 m (JUSTICE *et al.*, 2002).

Next, two classes of variables were entered:

a) Target variable – *soybean occurrence* (*S*) corresponding to the studied phenomenon, represented by a thematic map with four classes for the crop year 2005/2006: (i) target presence observed (i.e., soybean); (ii) target absence observed (i.e., non-soybean); (iii) missing data (i.e., no observations); and (iv) pixels outside the study area. This thematic map, produced by Epiphanio *et al.* (2010), was used as a reference in this study. In the BayNeRD modelling, *S* assumes a Binomial distribution with $S = s$, where $s = 1$ for soybean presence and $s = 0$ for soybean absence. Two thirds of the pixels in each of the thematic class soybean and non-soybean were randomly selected from the reference map to compose the *reference data for training*. The remaining third of the reference map pixels was set aside to be used for accuracy assessment (*reference data for testing*).

b) Context variables – the selected and available variables to compose the model are listed in Table 3.1. From expert knowledge it is known that each context variable influences soybean occurrence (*S*).

Table 3.1 – Summary of the six context variables used in the soybean mapping case study.

Variable	**Description**
C	CEI* value in the **C**urrent crop year (2005/2006)
L	CEI* value in the **L**ast crop year (2004/2005)
A	Soil **A**ptitude
T	**T**errain slope (given in %)
W	Distance to the nearest **W**ater body (given in Km)
R	Distance to the nearest **R**oad (given in Km)

*Crop Enhancement Index (RIZZI *et al.*, 2009).

As a remote sensing input, the Crop Enhancement Index [CEI (RIZZI *et al.*, 2009)] was used. CEI was designed to capture the high seasonality of annual crops, particularly soybean. It uses the Enhanced Vegetation Index [EVI (HUETE *et al.*, 2002)] values derived from MODIS images observed at two specific periods of the soybean crop calendar in the study area. CEI values may vary between [-1,+1] and are calculated, for each pixel, as

$$CEI = 100\frac{MaxEVI - MinEVI}{MaxEVI + MinEVI + 200} \tag{3.5}$$

where *MinEVI* is the minimum observed EVI value between June and August or prior to the beginning of the crop growing season, when EVI values are close to the minimum for annual crops; and *MaxEVI* is the maximum EVI value observed at the full soybean development period, occurring between December (earliest sowing) and March (latest sowing) when EVI values are at their highest for soybean (ARVOR *et al.*, 2011; RIZZI *et al.*, 2009).

In BayNeRD we used *CEI values in the current crop year* (*C* variable) for 2005/2006. It is expected that soybean presence leads to high values of CEI (RIZZI *et al.*, 2009). Therefore, since *soybean occurrence* influences the CEI value for the current crop year, *S* should be a parent of *C* in the BN model. In addition we also used *CEI values in the last crop year* (i.e., 2004/2005 – *L* variable). We used *L* because soybean plantations in Mato Grosso present spatially persistent characteristics over time, i.e., if soybean is sown on a given plot in a given year it is likely that soybean will be sown on the same plot in the following crop year (RISSO, 2013). Thus *L* should be a parent of *S* in the BN model.

Soybean occurrence is also influenced by soil type (RISSO, 2013), represented here by the variable *soil aptitude* (*A*). To set the *soil aptitude* for soybean production, we used a thematic soil map (1:250,000 scale) provided by the Secretariat of Planning and Coordination of Mato Grosso State (SEPLAN-MT, 2012). This map was produced within the scope of an ecological-economic zoning project, according to the Brazilian System of Soil Classification (PALMIERE *et al.*, 2002; SANTOS *et al.*, 2006). Originally, the soil map contained 26 classes (types of soil), which were pooled into two aptitude classes, low and high, defined by skilled soil experts according to soil properties such as soil composition, water holding capacity and fertility. The low aptitude class encompasses the following soils: rock outcrops, gleysols, lithic soils, quartz sands, planosols, plinthosols, podzolic soils, solonetzic soils, alluvial soils, cambisols, concretionary soils, organic soils and brunizem soils. On the other hand, the high aptitude class encompasses ultisols and oxisols (SANTOS *et al.*, 2006). Hence, since *A* influences *S*, *A* is a parent of *S* in the BN model.

The fourth context variable used was the *terrain slope* (*T*). To compute *T* we used altitude data derived from the Shuttle Radar Topography Mission [SRTM (RABUS *et al.*, 2003)]. *T* is critical in defining which fields are suitable for soybean production since it defines suitable areas for large scale mechanized agriculture such as soybean cultivation (SEERUTTUN; CROSSLEY, 1997; SHAXSON, 1999). Furthermore, land's erosive potential increases as slope increases, particularly if soil tilling practices are employed. Therefore *T* is a parent of *S* in the BN model. It is also known that *T* has a noticeable influence on soil formation (PARK *et al.*, 2001); thus *T* is also a parent of *A* in the BN model.

Another variable that influences soybean occurrence is the *distance to the nearest water body* (*W*), computed using the hydrographic network provided by the Brazilian Electricity Sector (ANEEL, 2012). This information includes the major river courses in Brazil, at a 1:1,000,000 scale. *W* was incorporated in this model for several reasons: (i) the rainfall pattern in Mato Grosso makes irrigation unnecessary, leading farmers to sow soybean preferably not close to river edges; (ii) Brazilian law safeguards preservation of natural vegetation in a buffer area around water bodies – up to 500 m, depending on the width of the water body, based on Brazilian Forest code in force at this evaluation time

(SILVA *et al.*, 2012); and (iii) short distances to water bodies are generally associated with higher terrain slopes, hampering the use of these areas for soybean production. Thus, we expect that the probability of soybean presence increases as the distance to water body increases. Therefore, *W* is both a parent of *S* in the BN model, since *soybean occurrence* is direct influenced by *W*, and a descendent of *T*, since *terrain slope* directly influences the path of flowing water channels.

The *distance to the nearest road* (*R*) was computed using the road map, provided by the Brazilian Institute of Geography and Statistics (IBGE, 2012a). This information includes the paved and unpaved road network for the entire country at a 1:5,000,000 scale. A close relationship between *soybean occurrence* and distance to roads is expected because of the logistical issues involved in accessing agricultural areas and transporting crops. That is, soybean production is expected to occur relatively close to major roads (FEARNSIDE, 2002). Therefore, *R* is a parent of *S* in the BN model.

Finally, areas that have no realistic role for commercial soybean production or are safeguarded by environmental protection laws in Mato Grosso were masked out. These include: (i) natural forest, identified from the Amazon Deforestation Monitoring Project (PRODES), carried out by INPE (2013) using the methodology described by Shimabukuro *et al.* (1998); (ii) floodplains, identified from SEPLAN-MT (2012); (iii) indigenous lands, identified from Brazil’s National Indian Foundation (FUNAI, 2013); and (iv) protected areas (also called Conservation Units), which are those without authorization for agricultural exploration, identified from the Brazilian Ministry of the Environment (MMA, 2013). These layers were overlaid to create a composite mask and all masked areas were omitted from analysis (see Fig. 3.2). Since some masked areas are suitable for soybean production in terms of physical properties, this step is important to minimize compromising the definition of the probability functions when counting pixels.

3.4.2. Bayesian Network model

Given the (in)dependence relationships among the context variables and between each context variable and the target variable (*S*) we designed a BN graphical model using a DAG (JENSEN; NIELSEN, 2007). Fig. 3.4 shows the designed model, where each node represents a variable and arcs between pairwise variables represent their dependence relationships.

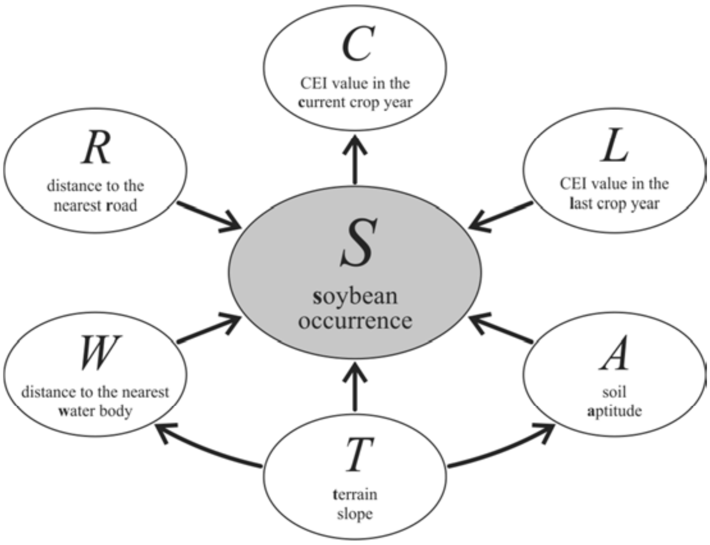

C
CEI value in the current crop year
R
distance to the nearest road
L
CEI value in the last crop year
S
soybean occurrence
W
distance to the nearest water body
A
soil aptitude
T
terrain slope

Figure 3.4 – Directed Acyclic Graph (DAG) encoding assertions of conditional (in)dependence among the variables and representing the designed Bayesian Network graphical model for the case study of *soybean occurrence* in Mato Grosso.

3.4.3. Discretization and probability functions

The first step after the definition of the BN graphical model is the discretization of continuous variables. The number of intervals must be appropriately chosen, i.e., neither too few to incorrectly describe the variable in the context of the phenomenon of interest nor too many to compromise the definition of the probability function associated to the variable and its descendants.

Regarding *T*, it is well known that soybean is preferably not sown on steep terrain slopes because mechanized cultivation processes may be hindered. Instead, soybean is usually sown in flat plateau areas with terrain slope < 6% (RISSO, 2013). A slope of 12% is considered the upper limit for mechanized cultivation (SHAXSON, 1999). Based on this knowledge, *T* was discretized into three intervals: one for slopes smaller than 6%, another for slopes equal to or larger than 6% but smaller than 12%, and the last for slopes equal to or larger than 12%. Since *T* has no parents, a prior probability function is defined. By pixel counting, BayNeRD computed the prior probability function for *T*, considering the defined intervals, i.e., $P(-\infty \leq T = t < 0.06)$, P(0.06 ≤ T = t < 0.12) and P(0.12 ≤ T = t < +∞). T is a parent of S, so the probabilities of soybean occurrence given each defined interval for T were also computed, i.e., P(S = s | -∞ ≤ T = t < 0.06), P(S = s | 0.06 ≤ T = t < 0.12) and P(S = s | 0.12 ≤ T = t < +∞). Fig. 3.5 shows a histogram of the discretized T variable and computed probabilities.

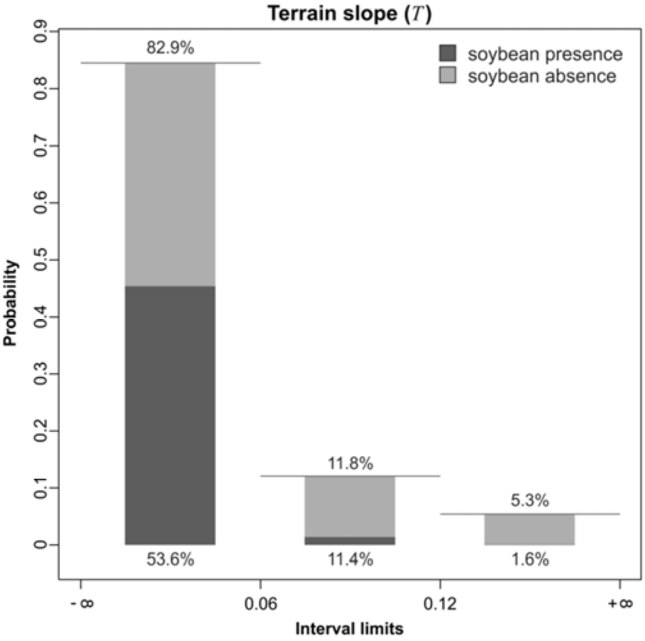
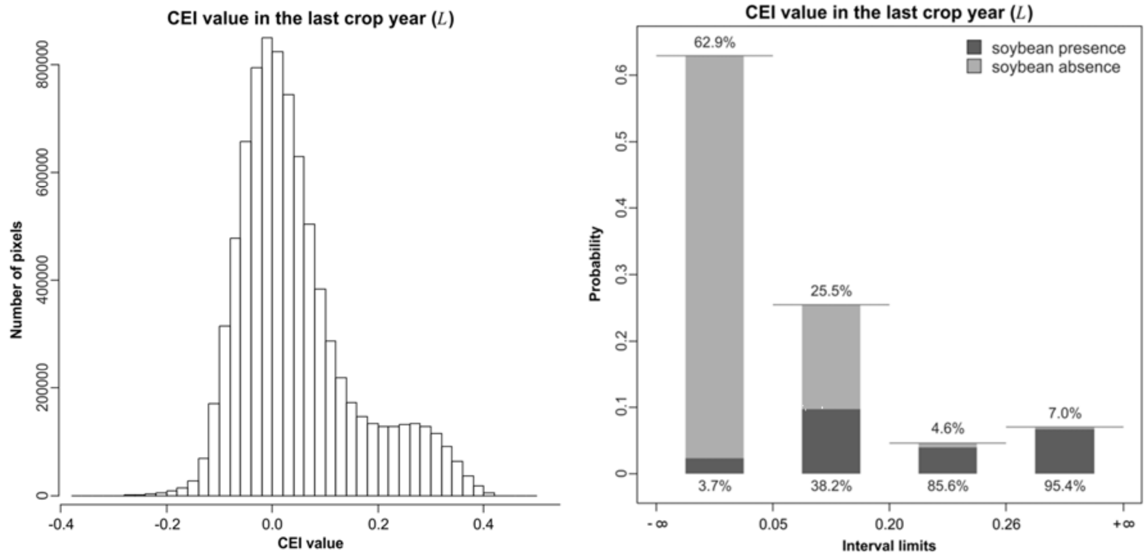


Figure 3.5 – Discretization of context variable *terrain slope* (T) into three intervals. The percentage at the top of each bar represents the probability of finding a pixel within the defined interval limits, e.g. P(-∞ ≤ T = t < 0.06) = 82.9%; and the percentage at the bottom of each bar represents the conditional probability of soybean presence given the defined interval limits for T, e.g. P(S = 1 | -∞ ≤ T = t < 0.06) = 53.6%.

Fig. 3.5 shows that almost 83% of the analysed area consists of flat areas, i.e., terrain slope smaller than 6%. Additionally, it shows that finding soybean plantations in these flat areas (probability of 53.6%) is more likely than in areas where slope is ≥ 12% (probability of 1.6%).

CEI (C and L) observations are also critical variables for this case study as they are closely related to *soybean occurrence* (RIZZI *et al.*, 2009).Fig. 3.6a shows a histogram of L values in the analysed area with bimodal appearance. CEI values less than 0.2 are usually associated with targets with low (e.g., forest) or medium seasonality (e.g., Cerrado or pasture) (GALFORD *et al.*, 2008; RISSO *et al.*, 2012). On the other hand, CEI values greater than 0.2 are strongly associated with high seasonality targets such as annual crops like soybean (RUDORFF *et al.*, 2011, 2012). Based on this knowledge, we empirically defined four intervals for *L*, as presented in Fig. 3.6b.



(a) (b)

Figure 3.6 – (a) Histogram of context variable *CEI value in the last crop year* (*L*); (b) discretization of *L* into four intervals. The percentage at the top of each bar represents the probability of finding a pixel within the defined interval limits, e.g., P(0.26 ≤ *L* = *l* < +∞) = 7.0%; and the percentage at the bottom of each bar represents the conditional probability of soybean presence given the defined interval limits for *L*, e.g.,P(*S* = 1 | 0.26 ≤ *L* = *l* < +∞) = 95.4%.

Indeed, Fig. 3.6b demonstrates the strong relationship between *S* and *L*. Although only 11.6% (4.6 + 7.0) of Mato Grosso State presented CEI values equal to or greater than 0.2 in the 2004/2005 crop year, the probability of finding soybean plantations in these areas in the 2005/2006 crop year is considerably greater than in the remaining part of the State.

Fig. 3.7 shows a histogram of *C* values with the same bimodal appearance as discussed for *L*, and a boxplot, where a strong relationship between soybean presence and *C* greater than 0.2 is also evident. Indeed, the relationship between *C* and *S* is similar to that between *L* and *S* because most soybean plantations of crop year 2005/2006 were sown over the same areas of crop year 2004/2005 due to the spatially persistent characteristic of soybean crop over time in Mato Grosso (RISSO, 2013). Thus, for *C* we used the same interval limits defined for *L*.

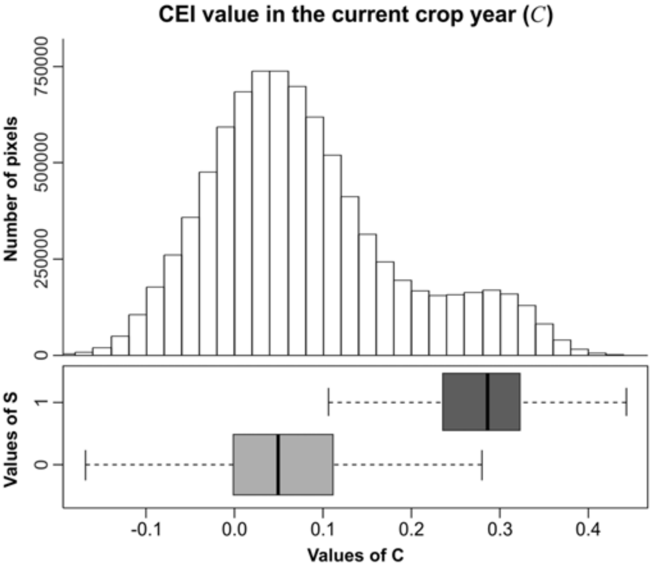


Figure 3.7 – Histogram of *CEI values observed in the current crop year* (C) and boxplot showing the strong relationship between soybean presence (S = 1) and C greater than 0.2.

As with T, L and C, we manually defined the upper and lower limits for the remaining context variables, as stated in Table 3.2. The main advantage of manual definition of interval limits is that it optimizes experts' knowledge during the discretization process.

Table 3.2 – Summary of the intervals limits defined for each of the six context variables, described in Table 3.1.

Interval #	***C***	***L***	***A***	***T***	***W***	***R***
1	[-∞ ; 0.05)	[-∞ ; 0.05)	low	[-∞ ; 0.06)	[-∞ ; 0.5)	[-∞ ; 3.0)
2	[0.05 ; 0.20)	[0.05 ; 0.20)	high	[0.06 ; 0.12)	[0.5 ; 1.0)	[3.0 ; 8.0)
3	[0.20 ; 0.26)	[0.20 ; 0.26)		[0.12 ; +∞)	[1.0 ; 2.0)	[8.0 ; +∞)
4	[0.26 ; +∞)	[0.26 ; +∞)			[2.0 ; +∞)	
# of intervals	4	4	2	3	4	3

Intervals are closed on the left and opened on the right, as denoted by [and), respectively.

3.4.4. Probability Image

Based on the designed BN model and the probability functions defined, BayNeRD computes, for each pixel in the study area, the probability of soybean presence given observations made on the context variables, i.e.,

$$P(S = 1|C = c, L = l, A = a, T = t, W = w, R = r) \tag{3.6}$$

where lower-case letters denote a state or value (defining a particular instantiation) of the respective discretized variable. The resulting PI was assessed visually and based on official data (i.e., from IBGE). The PI was also used to generate thematic maps that were statistically assessed, based on the *reference data for testing*, to determine the effectiveness of BayNeRD for characterising soybean cultivation.

3.5. Results and discussion of BayNeRD

3.5.1. Probability Image

The resulting PI (Fig. 3.8) is an image in which every pixel value represents the calculated probability as defined in Eq. 3.6.

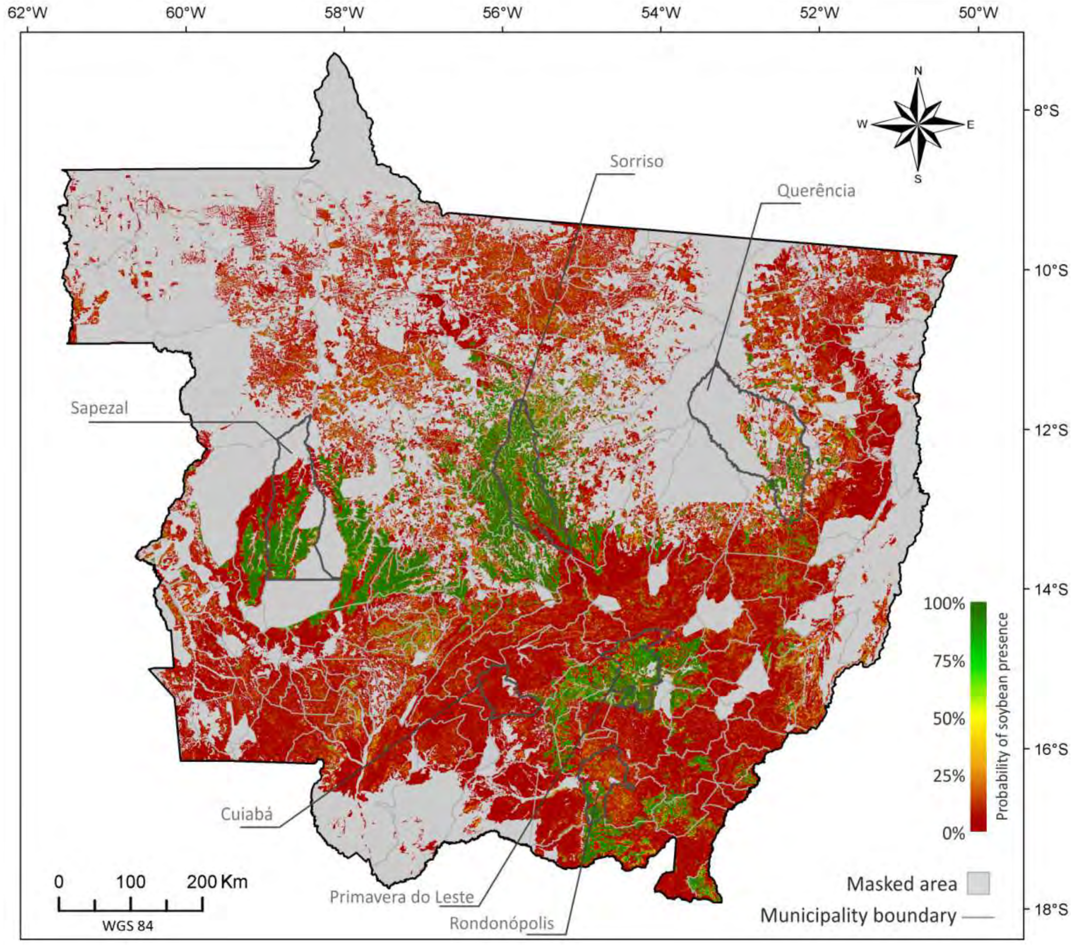


Figure 3.8 – Probability Image (PI) of soybean presence for the entire Mato Grosso State, Brazil. Main soybean producer centres and the capital, Cuiabá, are highlighted. The colour indicates the calculated probability of soybean presence in 2005/2006 given the observations made for the context variables, as expressed by Eq. 3.6.

The PI shows the spatial distribution of (the probability of) soybean crops throughout Mato Grosso territory in crop year 2005/2006. Green coloured pixels represent areas with higher probability of soybean presence based on observation of the context variables. Some of the main soybean production centres according to IBGE (2012b) are highlighted on the PI and allow us to verify the spatial coherence between PI and official soybean statistics. Previous studies that assessed soybean spatial distribution in Mato Grosso were also consistent with the regions of higher PI values identified here (ARVOR *et al.*, 2011; RIZZI *et al.*, 2009).

The higher probabilities shown in Fig. 3.8 highlight traditional centres of soybean production in the Cerrado biome of Mato Grosso, i.e Primavera do Leste, Rondonópolis, Sapezal and the central region (Sorriso Southward). More recent soybean frontiers are in transition regions between Cerrado and the Amazon biome. In Sorriso municipality Northward (along the BR 163 highway) and Querência region, which are considered to be the newer agricultural frontiers in Mato Grosso (JEPSON, 2009), pasturelands have been converted to soybean production in an accelerated way (RISSO, 2013).

Fig. 3.9 shows, for a small subset of the study area, the set of variables within different conditions leading to variations in the calculated soybean presence in crop year 2005/2006 (PI). The region labelled 1 is on a plateau and exhibits ideal conditions for soybean cultivation based on the designed BN model. CEI values (*C* and *L*) are high, predominantly in the upper discretized interval (≥ 0.26); *A* is high; *T* is flat (< 6%); *W* is ≥ 2 km; and a road crosses this plateau so *R* is < 3 km. Since every context variable exhibits favourable conditions for soybean presence, the combination of these conditions results in high probability of soybean presence. The region labelled 2 is on the edge of the plateau and represents an area where soybean plantations are usually close to pasture lands. In this case three context variables are favourable for soybean presence based on the criteria discussed above (*A*, *T* and *R*), but CEI values (*C* and *L*) are unfavourable (≤ 0.20). Moreover, there are two water bodies in this region further reduce the probability of soybean plantations. As a result, the probability of soybean presence in region 2 tended to range between 25 and 50%. The region labelled 3 corresponds to an area of Cerrado, and exhibits more or less the opposite condition to that of region 1. In this case, all context variables present unfavourable conditions for soybean presence, leading to probability values close to zero in the PI.

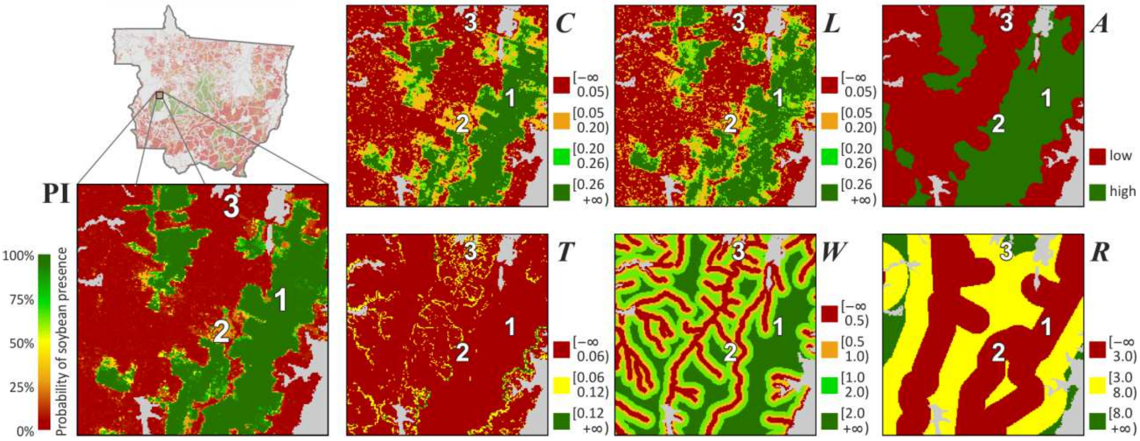


Figure 3.9 – Probability Image (PI) of soybean presence and six context variables (described in Table 3.1) zoomed in on the central part of the Sapezal municipality. The legend for the context variables followed the intervals stated in Table 3.2. Regions labelled 1, 2 and 3 show respectively, ideal, intermediate and flawed conditions for soybean cultivation.

Various other combinations of context variables can be found in the study area. The BN network is adept at dealing with such occurrences. According to KL divergence (KULLBACK; LEIBLER, 1951), *C* and *L* were the most important variables used to infer about soybean occurrence ($KL_C = 0.28$ and $KL_L = 0.16$). It means that, as pointed out by Risso *et al.* (2012), a proper vegetation index taken at key dates over the crop calendar can be used to identify specific crops such as soybean (RIZZI *et al.*, 2006). In fact, due to its ability and practicability to detect soybean areas, CEI is also used to monitor soybean plantations in the Brazilian Amazon Biome in the context of the Soy Moratorium (RUDORFF *et al.*, 2011, 2012). For the remaining context variables *A*, *T*, *W* and *R*, the KL divergences were 0.009, 0.002, 0.003 and 0.0001, respectively. This result means that soil type influenced more the calculated probability of soybean presence then terrain slope, water distance and especially the distance to a road.

The relatively small influence of *R* on the calculated probability of soybean presence could be explained by the fact that soybean fields are usually very large, particularly in Mato Grosso. Hence, even very high transportation costs do not hinder soybean cultivation (GARRETT *et al.*, 2013). Additionally, most soybean areas in Mato Grosso are consolidated (i.e., traditional areas planted with soybean), especially those surrounding Sapezal, Sorrizo, Rondonópolis and Primavera do Leste, where transportation logistics have been developed to fit the available road facilities. However, we expect *R* to be more influential close to agricultural frontiers such as in the region of Querência (JEPSON, 2009). Indeed, the close relationship between cash crops' occurrence and proximity to roads has been widely explored, often using models to predict future scenarios of agriculture expansion (JASINSKI *et al.*, 2005) and deforestation (SOARES-FILHO *et al.*, 2006). Although modelling such knowledge is possible in principle using BayNeRD, it was beyond the scope of the present study.

The influence of *T* on the calculated probability of soybean presence was minimized by the fact that most parts (83%) of the study were relatively flat ($T < 6\%$ – Fig. 3.5). Nevertheless, results showed that soybean is not likely to be sown in steep areas, corroborating that steep areas are unsuitable for large scale mechanized agriculture (SEERUTTUN; CROSSLEY, 1997; SHAXSON, 1999). Historically, landholders sow soybean on flat areas, such as *Chapada dos Parecis* and those surrounding the BR-163 highway in Mato Grosso central (e.g., Sorriso region), where the large soybean hubs are located (FEARNSIDE, 2002).

In general, where only one context variable is unfavourable and/or is not strongly related to *soybean occurrence* (such as *W*, which presented $KL_W = 0.003$), any decrease in the calculated probability of soybean presence is likely to be very small. However if the context variable has a strong relationship with *soybean occurrence* (for example *C*, which presented $KL_C = 0.28$), any unfavourable condition of this variable is likely to decrease soybean probability values substantially. Additionally, the mixing within a pixel size of 250 x 250 m (defined as our nominal spatial resolution), especially over the boundaries of defined discretized intervals, could be noted in Fig. 3.9, which presented yellow coloured pixels surrounding green pixels in the PI.

3.5.2. Creating thematic maps from the Probability Image

The PI, as shown in Fig. 3.8, is the main output of BayNeRD and may be used in a range of different applications. For example, if one is looking for soybean areas for environmental supervision of soybean plantations in recent deforested areas, as defined in the Soy Moratorium context in Brazil (RUDORFF *et al.*, 2011, 2012), then areas where the probability of presence of soybean is high could be prioritized and the PI could be used to guide the logistics of field inspection by regulatory agencies (MELLO *et al.*, 2010a). The PI can also be used as input for classifiers (e.g., as *prior probability* for the maximum likelihood classifier) or to mask out low probability areas before running a classification.

Additionally, the PI can also be used to produce a thematic map (e.g., for acreage estimates) by applying a threshold probability value where all pixels with values above the threshold are allocated to the target thematic class (e.g., soybean). This value, herein called TPV, can be defined as any real value between 0 and 100%. Apart from a manually defined TPV, six criteria were implemented in BayNeRD to select a TPV according to some criterion, as defined in section "*3.3.6. Selecting the Target Probability Value*", using reference information (e.g. *reference data for testing*). The TPV that produces the most suitable thematic map, following the chosen criterion, is then called the best TPV.

The goal is to find the TPV that generates the most suitable thematic map showing two classes: target (soybean) and non-target (non-soybean). Several metrics are discussed in the literature to access map accuracy (FOODY, 2002; LIU *et al.*, 2007). The most widely used one is the kappa index (COHEN, 1960; SMITS *et al.*, 1999). However, in the case of binary classifications, Foody (2010) pointed out the advantages of two complimentary indices: sensitivity and specificity (ALTMAN; BLAND, 1994). These indices indicate the ability to find true positives (e.g. soybean areas which are correctly labelled soybean) and true negatives (e.g. non-soybean areas which are correctly labelled non-soybean), respectively.

By varying the TPV from 0% to 100% different thematic maps were produced. Obviously, TPV = 0% produced a thematic map where all pixels within the study area were labelled as soybean. When all pixels were labelled soybean, all true soybean areas were then labelled as soybean and consequently sensitivity was equal to 100%. On the other hand, all true non-soybean areas were also labelled as soybean, and consequently specificity was 0%. With TPV increasing from 0 to 100%, sensitivity decreases while specificity increases. A useful graph to represent accuracy assessment in terms of these two indices is known as a Receiver Operating Characteristic (ROC) curve (HANLEY; MCNEIL, 1982). In a ROC curve the sensitivity is plotted on the Y-axis while the X-axis represents 1-specificity. Thus, the upper left corner represents the ideal point of 100% sensitivity and 100% specificity. According to Zweig and Campbell (1993), the closer the point is to the upper left corner in a ROC curve, the higher the overall accuracy of the thematic map. Therefore, the used *nearest 100% sensitivity and 100% specificity point* criterion aimed at selecting the TPV that produces a thematic map where its corresponding point in a ROC curve is closest to the upper left corner, based on the *reference data for testing*. Fig. 3.10 shows a ROC curve produced by varying TPV from 0 to 100%.

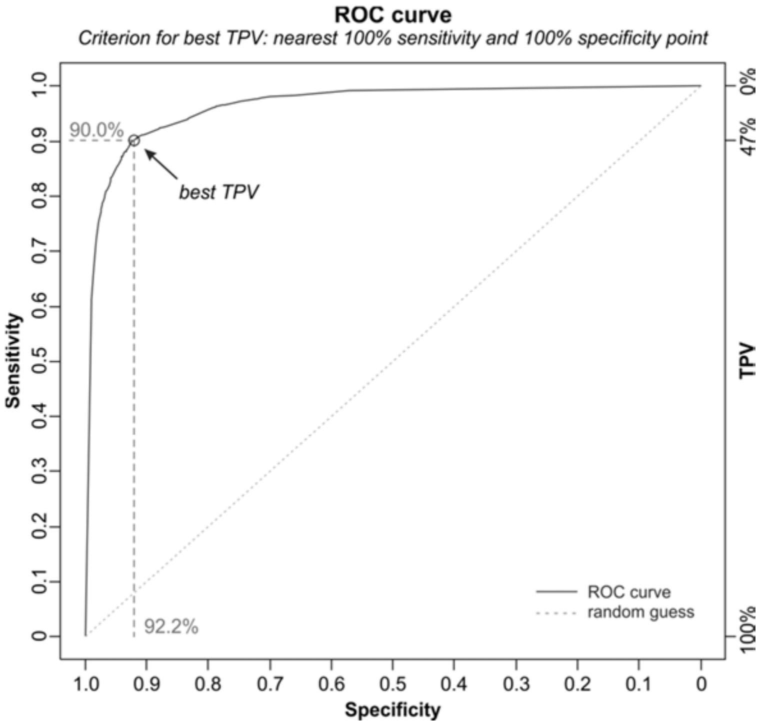


Figure 3.10 – Receiver Operating Characteristic (ROC) curve, depicting sensitivity and specificity indices associated with thematic maps generated from the Probability Image (PI) by varying the Target Probability Value (TPV) from 0 to 100%. The circle points out the best TPV according to the chosen criterion.

In the ROC curve presented in Fig. 3.10 all points plotted above the diagonal (random guess) represent a strong classification result (i.e. better than random) (HANLEY; MCNEIL, 1982). This indicates that the PI is an accurate representation of the phenomenon (in this case, soybean occurrence). According to the *nearest 100%*

sensitivity and 100% specificity point criterion, the best TPV should be 47%, resulting in a thematic map with sensitivity of 90.0% and specificity of 92.2%.Moreover, the overall accuracy of 91.1% and a kappa value of 0.82 corroborated the fact that this best TPV produced an accurate thematic map of soybean areas, based on the *reference data for testing*. Fig. 3.11 shows the accuracy indices for the PI-derived thematic maps generated by varying TPV from 0 to 100%.

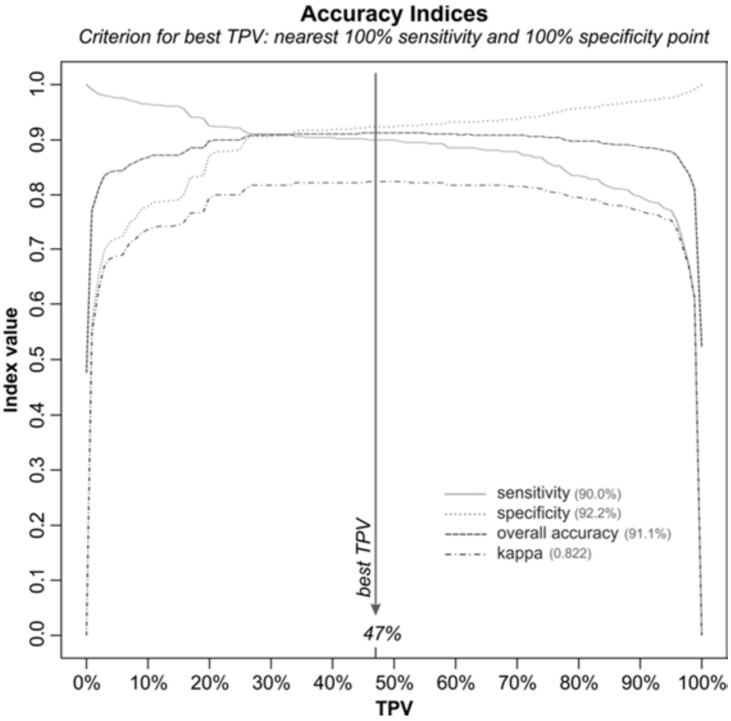


Figure 3.11 – Accuracy indices associated with thematic maps generated from the Probability Image (PI) by varying the Target Probability Value (TPV) from 0 to 100%. The vertical line identifies the best TPV, according to the chosen criterion, highlighting the accuracy achieved according to each index (described in the legend).

A TPV can be defined to be more or less restricted in terms of associating a degree of belief, represented by a probability value, in which a pixel can be associated to the target thematic class, prioritizing either sensitivity or specificity. If the aim is that the total soybean area of the final thematic map closely matches the official statistics, the TPV can also be selected accordingly. For example, the thematic map generated with a TPV equal to 84% is more restrictive in terms of labelling a pixel as soybean but best matched the official soybean acreage for the 2005/2006 crop year in Mato Grosso.

Indeed this thematic map presented 6.1 Mha of soybean – only 0.8% higher than the official data published by IBGE (2012b).

Similar to mapping soybean using remote sensing and environmental variables, Krug *et al.* (2013) used various environmental observations such as sea surface temperature and wind velocity in BNs to investigate coral bleaching along the Bahia State coast, Brazil. They also pointed out that BNs could be used as a prediction tool, incorporating evidence from a large data set of environmental observations, as we demonstrated here.

In fact, BayNeRD could be used to infer knowledge about a variety of phenomena based on observations of variables that are somehow related to the phenomena. For example, it may be used to identify forested areas susceptible to burning based on observations of forcing variables such as selective logging, deforestation, rainfall, distance to roads and land use type of surrounding areas (ARAGÃO *et al.*, 2008; SILVESTRINI *et al.*, 2011). Detecting landslide susceptibility based on observations made upon variables such as slope, soil, lithological classes, terrain curvature, land cover and rainfall represents another possible application of BayNeRD (FELL *et al.*, 2008). BayNERD could also enable inference about the occurrence of certain fish species based on data such as sea surface temperature, chlorophyll concentration and sea surface winds (OLIVEIRA *et al.*, 2010).

3.6. Conclusion to BayNeRD

This paper described the basis functioning and implementation of a computer aided BN method for raster data analysis: Bayesian Networks for Raster Data (BayNeRD). BayNeRD provides a new computer-aided method to characterise phenomena through plausible reasoning inferences based on observations of several variables. The number of variables is not limited and the sole conditions are an accurate match of raster cells and the availability of a suitable reference data set.

The case study of mapping soybean areas in Mato Grosso State, Brazil, showed BayNeRD's capability to model environmental phenomena. Based on observations made upon Crop Enhanced Index (CEI) values for the current and last crop years, soil type, terrain slope and distance to the nearest road and water body, the resulting

Probability Image (PI) from BayNeRD presented a spatial distribution of soybean areas consistent with expert knowledge and official statistical data. Starting from the PI, a thematic map could be produced depicting the spatial distribution of soybean and non-soybean areas with overall accuracy greater than 91%.

Advantages of BayNeRD include that it incorporates expert’s knowledge into the process; it models the (in)dependence relationships among several observed variables; it outputs variable importance information, through the Kullback-Leibler divergence; it can accommodate different forms of data (numerical and categorical); it can handle incomplete data; it allows computation of probability functions from the data; and it is a user-friendly implementation in a free software ready to handle raster data sets.

The BayNeRD algorithm has been implemented in R software and can be found on internet.

Acknowledgment

The authors thank Dr Nikolay Balov, from the University of Rochester Medical Center, for assistance with the catnet R package; MSc RuyDalla Valle Epiphanio for sharing part of the dataset used in this work; and the Brazilian Research Councils CNPq (*ConselhoNacional do Desenvolvimento Científico e Tecnológico* – 142845/2011-6), CAPES (*Coordenação de Aperfeiçoamento de Pessoal de Nível Superior* – 33010013005P0) and Fapesp (*Fundação de Amparo à Pesquisa do Estado de São Paulo* – 2008/56252-0), and the German Exchange Service DAAD (*Deutscher Akademischer Austausch Dienst* – 12/A/72898) for financial support.

4 Final remarks

This thesis presented two methods, which represent an advance to the development and implementation of methods for remotely sensed data analysis focused on cropland mapping applications. These methods were described in full and tested using case studies of sugarcane harvest and soybean mapping.

Chapter 2 presented the Spectral-Temporal Analysis by Response Surface (STARS) method in full and showed how STARS may be efficiently used to monitor sugarcane harvest in Brazil. We tested two different response surface models [i.e., Polynomial Trend Surface (PTS) and Collocation Surface (CS)] and two types of coefficients (i.e., orthonormal and regular) for the description of a multitemporal-multispectral Landsat dataset of 11 images (six spectral bands). With an overall accuracy of 99%, STARS performed well when used as input features in classifications aiming to map sugarcane fields harvested with or without straw burning, and sugarcane fields not harvested by the end of the crop harvest season. Although tested as input for classifiers, STARS is a robust method for modelling spectral-temporal changes of agricultural targets on Earth’s surface. It reduces noise and dimensionality (e.g., PTS model) and may deal with images acquired at irregular time intervals, by different sensors with multispectral bands. Additionally, STARS can be used in a range of applications.

Complimentary to the thesis objective, chapter 3 described the Bayesian Network for Raster Data (BayNeRD), which allows the modelling of complex phenomena integrating variables into a model to make inferences using plausible reasoning from observations. This chapter briefly introduced Bayesian Networks (BN) theory and described how it was used to develop BayNeRD. The case study of soybean mapping in Mato Grosso State was used to test and evaluate BayNeRD. We integrated two years of remotely sensed (represented by a crop index named Crop Enhancement Index – CEI) and ancillary (i.e., topography, soil, roads and water bodies) data into a BN model which encoded the dependence relationship among these variables and between each one of them with soybean occurrence in Mato Grosso. The Probability Image (PI) that resulted from BayNeRD showed strong adherence to the official agricultural statistics from IBGE. Moreover, the thematic map generated from PI presented more than 91% of overall accuracy. Although ancillary data proved to increase accuracy of classifications, we found that remotely sensed data had the strongest influence, as evidenced by the calculated probability of soybean presence. This result demonstrated the potential of remote sensing as a source of data for agricultural monitoring. BayNeRD allowed the expert to model the soybean occurrence phenomenon, outputted variables' importance information, and handled incomplete and different sort of data. Indeed, BayNeRD showed potential for use in several applications such as for the Soy Moratorium context.

The two methods developed and tested confirm our hypothesis that remotely sensed (and ancillary) data analysis can be automated through computer aided methods to model a range of cropland phenomena for agriculture applications, maintaining consistency and accuracy. Both methods were entirely implemented in R software.

REFERENCES

ABDEL-RAHMAN, E. M.; AHMED, F. B. The application of remote sensing techniques to sugarcane (*Saccharum spp. hybrid*) production: a review of the literature. **International Journal of Remote Sensing**, v. 29, n. 13, p. 3753–3767, jul. 2008.

ABEND, K.; HARLEY JR., T. J. Comments on “On the mean accuracy of statistical pattern recognizers” by Hughes, G. F. **IEEE Transactions on Information Theory**, v. 15, n. 3, p. 420–423, maio 1969.

ADAMI, M.; MELLO, M. P.; AGUIAR, D. A.; RUDORFF, B. F. T.; SOUZA, A. F. A Web platform development to perform thematic accuracy assessment of sugarcane mapping in south-central Brazil. **Remote Sensing**, v. 4, n. 10, p. 3201–3214, 2012a.

ADAMI, M.; RUDORFF, B. F. T.; FREITAS, R. M.; AGUIAR, D. A.; SUGAWARA, L. M.; MELLO, M. P. Remote sensing time series to evaluate direct land use change of recent expanded sugarcane crop in Brazil. **Sustainability**, v. 4, n. 4, p. 574–585, 2012b.

AGUIAR, D. A.; RUDORFF, B. F. T.; ADAMI, M.; SHIMABUKURO, Y. E. Imagens de sensoriamento remoto no monitoramento da colheita da cana-de-açúcar. **Engenharia Agrícola**, v. 29, n. 3, p. 440–451, set. 2009.

AGUIAR, D. A.; RUDORFF, B. F. T.; SILVA, W. F.; ADAMI, M.; MELLO, M. P. Remote sensing images in support of environmental protocol: monitoring the sugarcane harvest in São Paulo State, Brazil. **Remote Sensing**, v. 3, n. 12, p. 2682-2703, 2011.

AGUILERA, P. A.; FERNÁNDEZ, A.; FERNÁNDEZ, R.; RUMÍ, R.; SALMERÓN, A. Bayesian networks in environmental modelling. **Environmental Modelling & Software**, v. 26, n. 12, p. 1376–1388, 2011.

AHA, D. W.; KIBLER, D.; ALBERT, M. K. Instance-based learning algorithms. **Machine Learning**, v. 6, n. 1, p. 37–66, 1991.

ALBERT, J. **Bayesian Computation with R**. 2. ed. New York, NY, USA: Springer New York, 2009. 298 p.

ALTMAN, D. G.; BLAND, J. M. Statistics notes: diagnostic tests 1: sensitivity and specificity. **BMJ**, v. 308, n. 6943, p. 1552–1552, 1994.

ANEEL. **Sistema de Informações Georeferenciadas do Setor Elétrico (SIGEO)**. Available at: <http://sigel.aneel.gov.br>. Accessed on: 20 Apr., 2012.

APLIN, P. On scales and dynamics in observing the environment. **International Journal of Remote Sensing**, v. 27, n. 11, p. 2123–2140, 2006.

ARAGÃO, L. E. O. C.; MALHI, Y.; BARBIER, N.; LIMA, A.; SHIMABUKURO, Y.; ANDERSON, L.; SAATCHI, S. Interactions between rainfall, deforestation and fires during recent years in the Brazilian Amazonia. **Philosophical transactions of the Royal Society of London - Series B, Biological sciences**, v. 363, n. 1498, p. 1779–85, 2008.

ARVOR, D.; JONATHAN, M.; MEIRELLES, M. S. P.; DUBREUIL, V.; DURIEUX, L. 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.

ASNER, G. P. Cloud cover in Landsat observations of the Brazilian Amazon. **International Journal of Remote Sensing**, v. 22, n. 18, p. 3855–3862, 2001.

ATZBERGER, C. Advances in remote sensing of agriculture: context description, existing operational monitoring systems and major information needs. **Remote Sensing**, v. 5, n. 2, p. 949–981, 2013.

BALOV, N.; SALZMAN, P. **catnet:** Categorical Bayesian Network Inference. 2011. Available at: <http://cran.r-project.org/package=catnet>. Accessed on: 8 May, 2011

BARGIEL, D.; HERRMANN, S. Multi-temporal land-cover classification of agricultural areas in two European regions with high resolution spotlight TerraSAR-X data. **Remote Sensing**, v. 3, n. 5, p. 859–877, 2011.

BATISTA, G. T.; MENDONÇA, F. J.; LEE, D. C. L.; TARDIN, A. T.; CHEN, S. C.; NOVAES, R. A. **Uso de sensores remotos a bordo de satélite e aeronave na identificação e avaliação de áreas de culturas para fins de previsão de safras**. São José dos Campos, SP, Brazil: INPE, 1978. (INPE-1229-NTI/103).

BECKER-RESHEF, I.; JUSTICE, C.; SULLIVAN, M.; VERMOTE, E.; TUCKER, C.; ANYAMBA, A.; SMALL, J.; PAK, E.; MASUOKA, E.; SCHMALTZ, J.; HANSEN, M.; PITTMAN, K.; BIRKETT, C.; WILLIAMS, D.; REYNOLDS, C.; DOORN, B. Monitoring global croplands with coarse resolution earth observations: the Global Agriculture Monitoring (GLAM) project. **Remote Sensing**, v. 2, n. 6, p. 1589–1609, 2010.

BIVAND, R. S.; PEBESMA, E. J.; GÓMEZ-RUBIO, V. **Applied spatial data analysis with R**. New York, NY, USA: Springer, 2008. 378 p.

BJÖRCK, Å. Solving linear least squares problems by Gram-Schmidt orthogonalization. **BIT Numerical Mathematics**, v. 7, n. 1, p. 1–21, 1967.

BOVOLO, F.; MARCHESI, S.; BRUZZONE, L. A framework for automatic and unsupervised detection of multiple changes in multitemporal images. **IEEE Transactions on Geoscience and Remote Sensing**, v. 50, n. 6, p. 2196–2212, 2012.

BRASIL. **Resolução da Presidência do IBGE de n° 5 (R.PR-5/02) de 10 de outubro de 2002.** Brasília, DF, Brazil: Diário Oficial da União, 2002.

BRUZZONE, L.; SMITS, P. **Analysis of multi-temporal remote sensing images: proceedings of Multitemp 2001**. Trento, Italy: World Scientific Publishing Company, 2002. v. 2. 456 p.

CANÇADO, J. E. D.; SALDIVA, P. H. N.; PEREIRA, L. A. A.; LARA, L. B. L. S.; ARTAXO, P.; MARTINELLI, L. A.; ARBEX, M. A.; ZANOBETTI, A.; BRAGA, A. L. F. the impact of sugar cane–burning emissions on the respiratory system of children and the elderly. **Environmental Health Perspectives**, v. 114, n. 5, p. 725–729, 2006.

CANTY, M. J.; NIELSEN, A. A. Automatic radiometric normalization of multitemporal satellite imagery with the iteratively re-weighted MAD transformation. **Remote Sensing of Environment**, v. 112, n. 3, p. 1025–1036, 2008.

CANTY, M. J.; NIELSEN, A. A.; SCHMIDT, M. Automatic radiometric normalization of multitemporal satellite imagery. **Remote Sensing of Environment**, v. 91, n. 3-4, p. 441–451, 2004.

CARRÃO, H.; GONÇALVES, P.; CAETANO, M. Contribution of multispectral and multitemporal information from MODIS images to land cover classification. **Remote Sensing of Environment**, v. 112, n. 3, p. 986–997, 2008.

CARTER, G. A. Primary and secondary effects of water content on the spectral reflectance of leaves. **American Journal of Botany**, v. 78, n. 7, p. 916–924, 1991.

CHANDER, G.; MARKHAM, B. L.; HELDER, D. L. Summary of current radiometric calibration coefficients for Landsat MSS, TM, ETM+, and EO-1 ALI sensors. **Remote Sensing of Environment**, v. 113, n. 5, p. 893–903, 2009.

COHEN, J. A coefficient of agreement for nominal scales. **Educational and Psychological Measurement**, v. 20, n. 1, p. 37–46, 1960.

COMPANHIA NACIONAL DE ABASTECIMENTO (CONAB). **Séries históricas relativas às safras 1976/77 a 2011/2012 de área plantada, produtividade e produção.** Available at: <http://www.conab.gov.br/conteudos.php?a=1252&t=>. Accessed on: 21 Mar., 2013.

CONGALTON, R. G.; GREEN, K. **Assessing the accuracy of remotely sensed data: principles and pratices**. 2. ed. Boca Raton, FL, USA: CRC Press, 2009. 183 p.

COOPER, G. F.; HERSKOVITS, E. A Bayesian method for the induction of probabilistic networks from data. **Machine Learning**, v. 9, n. 4, p. 309–347, 1992.

COPPIN, P.; JONCKHEERE, I.; NACKAERTS, K.; MUYS, B.; LAMBIN, E. F. Digital change detection methods in ecosystem monitoring: a review. **International Journal of Remote Sensing**, v. 25, n. 9, p. 1565–1596, 2004.

CRAWLEY, M. J. **The R book**. Chichester, England: John Wiley & Sons, 2007. 950 p.

DAI, X.; KHORRAM, S. The effects of image misregistration on the accuracy of remotely sensed change detection. **IEEE Transactions on Geoscience and Remote Sensing**, v. 36, n. 5, p. 1566–1577, 1998.

DARWICHE, A. **Modeling and reasoning with Bayesian networks**. New York, NY, USA: Cambridge University Press, 2009. 560 p.

DEFRIES, R. S.; BELWARD, A. S. Global and regional land cover characterization from satellite data: an introduction to the Special Issue. **International Journal of Remote Sensing**, v. 21, n. 6, p. 1083–1092, 2000.

DEMIR, B.; BOVOLO, F.; BRUZZONE, L. Detection of land-cover transitions in multitemporal remote sensing images with active-learning-based compound classification. **IEEE Transactions on Geoscience and Remote Sensing**, v. 50, n. 5, p. 1930–1941, 2011a.

DEMIR, B.; PERSELLO, C.; BRUZZONE, L. Batch-mode active-learning methods for the interactive classification of remote sensing images. **IEEE Transactions on Geoscience and Remote Sensing**, v. 49, n. 3, p. 1014–1031, mar. 2011b.

DONNER, R.; BARBOSA, S.; KURTHS, J.; MARWAN, N. Understanding the Earth as a complex system – recent advances in data analysis and modelling in Earth sciences. **The European Physical Journal Special Topics**, v. 174, n. 1, p. 1–9, 2009.

DRAPER, N. R.; SMITH, H. **Applied regression analysis**. 1. ed. New York, USA: John Wiley & Sons, 1966.

EL HAJJ, M.; BÉGUÉ, A.; GUILLAUME, S.; MARTINÉ, J.-F. Integrating SPOT-5 time series, crop growth modeling and expert knowledge for monitoring agricultural practices The case of sugarcane harvest on Reunion Island. **Remote Sensing of Environment**, v. 113, n. 10, p. 2052–2061, 2009.

EPIPHANIO, R. D. V.; FORMAGGIO, A. R.; RUDORFF, B. F. T.; MAEDA, E. E.; LUIZ, A. J. B. Estimating soybean crop areas using spectral-temporal surfaces derived from MODIS images in Mato Grosso, Brazil. **Pesquisa Agropecuária Brasileira**, v. 45, n. 1, p. 72–80, 2010.

FOOD AND AGRICULTURE ORGANIZATION (FAO). **FAOSTAT:** FAO statistical database. Available at: <http://faostat.fao.org>. Accessed on: 9 Apr., 2012.

FARRELL, M. D.; MERSEREAU, R. M. On the impact of pca dimension reduction for hyperspectral detection of difficult targets. **IEEE Geoscience and Remote Sensing Letters**, v. 2, n. 2, p. 192–195, 2005.

FAUSETT, L. V. **Fundamentals of neural networks:** architectures, algorithms and applications. 1. ed. Upper Saddle River, NJ, USA: Prentice Hall, 1993. 461 p.

FEARNSIDE, P. M. Soybean cultivation as a threat to the environment in Brazil. **Environmental Conservation**, v. 28, n. 01, p. 23–38, 2002.

FELL, R.; COROMINAS, J.; BONNARD, C.; CASCINI, L.; LEROI, E.; SAVAGE, W. Z. Guidelines for landslide susceptibility, hazard and risk zoning for land-use planning. **Engineering Geology**, v. 102, n. 3-4, p. 99–111, 2008.

FIGUEIREDO, E. B.; LA SCALA JR., N. Greenhouse gas balance due to the conversion of sugarcane areas from burned to green harvest in Brazil. **Agriculture, Ecosystems & Environment**, v. 141, n. 1-2, p. 75–85, 2011.

FOODY, G. M. Assessing the accuracy of land cover change with imperfect ground reference data. **Remote Sensing of Environment**, v. 114, n. 10, p. 2271–2285, 2010.

FOODY, G. M. Sample size determination for image classification accuracy assessment and comparison. **International Journal of Remote Sensing**, v. 30, n. 20, p. 5273-5291, 2009.

FOODY, G. M. Status of land cover classification accuracy assessment. **Remote Sensing of Environment**, Consagrado paper do Foody sobre exatidão de mapas temáticos, v. 80, n. 1, p. 185–201, 2002.

FORSYTHE, G. E. Generation and use of orthogonal polynomials for data-fitting with a digital computer. **Journal of the Society for Industrial and Applied Mathematics**, v. 5, n. 2, p. 74–88, jun. 1957.

FUNDAÇÃO NACIONAL DO ÍNDIO (FUNAI). **Maps**. Available at: <http://mapas.funai.gov.br>. Accessed on: 20 Jan., 2013.

GALDOS, M. V.; CERRI, C. C.; CERRI, C. E. P. Soil carbon stocks under burned and unburned sugarcane in Brazil. **Geoderma**, v. 153, n. 3-4, p. 347–352, 2009.

GALFORD, G. L.; MUSTARD, J. F.; MELILLO, J.; GENDRIN, A.; CERRI, C. C.; CERRI, C. E. . Wavelet analysis of MODIS time series to detect expansion and intensification of row-crop agriculture in Brazil. **Remote Sensing of Environment**, v. 112, n. 2, p. 576–587, 2008.

GARRETT, R. D.; LAMBIN, E. F.; NAYLOR, R. L. Land institutions and supply chain configurations as determinants of soybean planted area and yields in Brazil. **Land Use Policy**, v. 31, p. 385–396, 2013.

GAUSMAN, H. W.; ALLEN, W. A.; CARDENAS, R. Reflectance of cotton leaves and their structure. **Remote Sensing of Environment**, v. 1, n. 1, p. 19–22, 1969.

GOLDEMBERG, J.; COELHO, S. T.; GUARDABASSI, P. The sustainability of ethanol production from sugarcane. **Energy Policy**, v. 36, n. 6, p. 2086–2097, 2008.

GOLUB, G. H.; VAN LOAN, C. F. **Matrix computations**. 3. ed. Baltimore, MD, USA: Johns Hopkins University Press, 1996.

HANLEY, J. A.; MCNEIL, B. J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. **Radiology**, v. 143, n. 1, p. 29–36, 1982.

HARALICK, R. M.; STERNBERG, S. R.; ZHUANG, X. Image analysis using mathematical morphology. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. PAMI-9, n. 4, p. 532–550, 1987.

HARDY, R. L. multiquadric equations of topography and other irregular surfaces. **Journal of Geophysical Research**, v. 76, n. 8, p. 1905–1915, 1971.

HECKERMAN, D. Bayesian networks for data mining. **Data Mining and Knowledge Discovery**, v. 1, n. 1, p. 79–119, 1997.

HOOGWIJK, M.; FAAIJ, A.; EICKOUT, B.; DEVRIES, B.; TURKENBURG, W. Potential of biomass energy out to 2100, for four IPCC SRES land-use scenarios. **Biomass and Bioenergy**, v. 29, n. 4, p. 225–257, 2005.

HUDSON, W. D. Correct formulation of the Kappa coefficient of agreement. **Photogrammetric Engineering & Remote Sensing**, v. 53, n. 4, p. 421–422, 1987.

HUETE, A. R.; DIDAN, K.; MIURA, T.; RODRIGUEZ, E. .; GAO, X.; FERREIRA, L. G. 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.

HUGHES, G. On the mean accuracy of statistical pattern recognizers. **IEEE Transactions on Information Theory**, v. 14, n. 1, p. 55–63, 1968.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA (IBGE). **Maps**. Available at: <http://mapas.ibge.gov.br/en/>. Accessed on: 29 Apr., 2012a.

______. **Pesquisas agropecuárias**. 2. ed. Rio de Janeiro, RJ, Brazil: IBGE, 2002. 92 p.

______. Sistema IBGE de Recuperação Automática (SIDRA) - Produção Agrícola Municipal (PAM). . Rio de Janeiro, RJ, Brazil: IBGE. Available at: <http://www.sidra.ibge.gov.br/>. Accessed on: 21 Nov., 2012b.

INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS (INPE). **PRODES:** Projeto de monitoramento do desflorestamento na Amazônia Legal. Available at: <http://www.obt.inpe.br/prodes/index.php>. Accessed on: 20 Jan., 2012.

JASINSKI, E.; MORTON, D.; DEFRIES, R.; SHIMABUKURO, Y.; ANDERSON, L.; HANSEN, M. Physical landscape correlates of the expansion of mechanized agriculture in Mato Grosso, Brazil. **Earth Interactions**, v. 9, n. 16, p. 1–18, 2005.

JAYNES, E. T. **Probability theory:** the logic of science. Cambridge, UK: Cambridge University Press, 2003. 727 p.

JENSEN, J. R. **Remote sensing of the environment:** an earth resource perspective. 2. ed. Upper Saddle River, USA: Prentice Hall, 2006. 608 p.

JENSEN, F. V.; NIELSEN, T. D. **Bayesian networks and decision graphs**. 2. ed. New York, NY, USA: Springer, 2007. 447 p.

JEPSON, W. Producing a Modern Agricultural Frontier: Firms and Cooperatives in Eastern Mato Grosso, Brazil. **Economic Geography**, v. 82, n. 3, p. 289–316, 2009.

JIMENEZ, L.; LANDGREBE, D. Hyperspectral data analysis and supervised feature reduction via projection pursuit. **IEEE Transactions on Geoscience and Remote Sensing**, v. 37, n. 6, p. 2653–2667, 1999.

JOLLIFFE, I. T. **Principal component analysis**. 2. ed. New York, NY, USA: Springer-Verlag New York Inc., 2002. 487 p.

JUSTICE, C. .; TOWNSHEND, J. R. G.; VERMOTE, E. .; MASUOKA, E.; WOLFE, R. .; SALEOUS, N.; ROY, D. .; MORISETTE, J. . An overview of MODIS Land data processing and product status. **Remote Sensing of Environment**, v. 83, n. 1-2, p. 3–15, 2002.

KIM, H.; KIM, S.; DALE, B. E. Biofuels, Land Use Change, and Greenhouse Gas Emissions: Some Unexplored Variables. **Environmental Science & Technology**, v. 43, n. 3, p. 961–967, 2009.

KIRCHHOFF, V. W. J. H.; MARINHO, E. V. A.; DIAS, P. L. S.; PEREIRA, E. B.; CALHEIROS, R.; ANDRÉ, R.; VOLPE, C. Enhancements of CO and O3 from burnings in sugar cane fields. **Journal of Atmospheric Chemistry**, v. 12, n. 1, p. 87-102, 1991.

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

KRUG, L. A.; GHERARDI, D. F. M.; STECH, J. L.; LEÃO, Z. M. A. N.; KIKUCHI, R. K. P.; HRUSCHKA, E. R.; SUGGETT, D. J. The construction of causal networks to estimate coral bleaching intensity. **Environmental Modelling & Software**, v. 42, p. 157–167, 2013.

KULLBACK, S.; LEIBLER, R. A. On Information and Sufficiency. **The Annals of Mathematical Statistics**, v. 22, n. 1, p. 79–86, 1951.

KUTNER, M. H.; NACHTSHEIM, C. J.; NETER, J.; LI, W. **Applied linear statistical models**. 5. ed. New York, USA: McGraw-Hill, 2005. 1424 p.

LAMBIN, E. F.; GEIST, H. (Org.). **Land-use and land-cover change**: local processes and global impacts. Berlin Heidelberg, Germany: Springer, 2006. 222 p.

LAMBIN, E. F.; LINDERMAN, M. Time series of remote sensing data for land change science. **IEEE Transactions on Geoscience and Remote Sensing**, v. 44, n. 7, p. 1926-1928, 2006.

LAMBIN, E. F.; STRAHLER, A. H. Indicators of land-cover change for change-vector analysis in multitemporal space at coarse spatial scales. **International Journal of Remote Sensing**, v. 15, n. 10, p. 2099–2119, 1994.

LANDGREBE, D. A. Multispectral land sensing: where from, where to? **IEEE Transactions on Geoscience and Remote Sensing**, v. 43, n. 3, p. 414–421, 2005.

LARA, L. L.; ARTAXO, P.; MARTINELLI, L. A.; CAMARGO, P. B.; VICTORIA, R. L.; FERRAZ, E. S. B. Properties of aerosols from sugar-cane burning emissions in Southeastern Brazil. **Atmospheric Environment**, v. 39, n. 26, p. 4627–4637, 2005.

LEE, C. A.; GASSTER, S. D.; PLAZA, A.; CHANG, C.-I.; HUANG, B. Recent developments in high performance computing for remote sensing: a review. **IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing**, v. 4, n. 3, p. 508–527, 2011.

LEE, J.; ERSOY, O. K. consensual and hierarchical classification of remotely sensed multispectral images. **IEEE Transactions on Geoscience and Remote Sensing**, v. 45, n. 9, p. 2953–2963, 2007.

LEE, S.; ZOU, F.; WRIGHT, F. A. Convergence and prediction of Principal Component scores in high-dimensional settings. **Annals of Statistics**, v. 38, n. 6, p. 3605–3629, 2010.

LEITE, R. C. C.; LEAL, M. R. L. V.; CORTEZ, L. A. B.; GRIFFIN, W. M.; SCANDIFFIO, M. I. G. Can Brazil replace 5% of the 2025 gasoline world demand with ethanol? **Energy**, v. 34, n. 5, p. 655–661, 2009.

LI, Z.; CHEN, J.; BALTSAVIAS, E. (Org.). **Advances in photogrammetry, remote sensing and spatial information sciences:** 2008 ISPRS Congress Book. 1. ed. Trowbridge, Wiltshire, UK: CRC Press, 2008. 546 p.

LIN, H.; CHEN, J.; PEI, Z.; ZHANG, S.; HU, X. Monitoring Sugarcane Growth Using ENVISAT ASAR Data. **IEEE Transactions on Geoscience and Remote Sensing**, v. 47, n. 8, p. 2572–2580, 2009.

LIU, C.; FRAZIER, P.; KUMAR, L. Comparative assessment of the measures of thematic classification accuracy. **Remote Sensing of Environment**, v. 107, n. 4, p. 606–616, 2007.

LOONEY, C. G. **Pattern recognition using neural networks:** theory and algorithms for engineers and scientists. New York, NY, USA: Oxford University Press, USA, 1997. 480 p.

LOVELAND, T. R.; SOHL, T. L.; STEHMAN, S. V.; GALLANT, A. L.; SAYLER, K. L.; NAPTON, D. E. A strategy for estimating the rates of recent United States land-cover changes. **Photogrammetric Engineering & Remote Sensing**, v. 68, n. 10, p. 1091–1099, 2002.

LU, D.; WENG, Q. A survey of image classification methods and techniques for improving classification performance. **International Journal of Remote Sensing**, v. 28, n. 5, p. 823–870, 2007.

LUCON, O.; GOLDEMBERG, J. Sao Paulo – the “other” Brazil: different pathways on climate change for state and federal governments. **The Journal of Environment & Development**, v. 19, n. 3, p. 335–357, 2010.

LUNETTA, R. S.; KNIGHT, J. F.; EDIRIWICKREMA, J.; LYON, J. G.; WORTHY, L. D. Land-cover change detection using multi-temporal MODIS NDVI data. **Remote Sensing of Environment**, v. 105, n. 2, p. 142–154, 2006.

MACEDO, I. C.; SEABRA, J. E. A.; SILVA, J. E. A. R. Green house gases emissions in the production and use of ethanol from sugarcane in Brazil: the 2005/2006 averages and a prediction for 2020. **Biomass and Bioenergy**, v. 32, n. 7, p. 582–595, 2008.

MACEDO, M. N.; DEFRIES, R. S.; MORTON, D. C.; STICKLER, C. M.; GALFORD, G. L.; SHIMABUKURO, Y. E. Decoupling of deforestation and soy production in the southern Amazon during the late 2000s. **Proceedings of the National Academy of Sciences of the United States of America**, v. 109, n. 4, p. 1341–1346, 2012.

MARKHAM, B. L.; BARKER, J. L. Landsat MSS and TM post-calibration dynamic ranges, exoatmospheric reflectances and at-satellite temperatures. **EOSAT Landsat Data User Notes**, v. 1, n. 1, p. 3–8, 1986.

MATHER, P. M. **Computational methods of multivariate analysis in physical geography**. Chichester, UK: John Wiley & Sons, 1976. 532 p.

MCGRAYNE, S. B. **The theory that would not die:** How Bayes’ rule cracked the enigma code, hunted down russian submarines, and emerged triumphant from two centuries of controversy. New Haven, CT, USA: Yale University Press, 2011. 336 p.

MELESSE, A. M.; WENG, Q.; THENKABAIL, P. S.; SENAY, G. B. Remote sensing sensors and applications in environmental resources mapping and modelling. **Sensors**, v. 7, n. 12, p. 3209–3241, 2007.

MELLO, M. P. **Classificação espectro-temporal de imagens orbitais para o mapeamento da colheita da cana-de-açúcar com queima da palha**. 2009. 130 p. (INPE-16222-TDI/1543). Dissertation (M.Sc. in Remote Sensing) – National Institute for Space Research (INPE), São José dos Campos, SP, Brazil, 2009.

MELLO, M. P.; AGUIAR, D. A.; RUDORFF, B. F. T.; PEBESMA, E.; JONES, J.; SANTOS, N. C. P. Spatial statistic to assess remote sensing acreage estimates: an analysis of sugarcane in São Paulo State, Brazil. In: IEEE INTERNATIONAL GEOSCIENCE AND REMOTE SENSING SYMPOSIUM (IGARSS 2013), 33., 2013, Melbourne, Australia. **Proceedings...** Los Alamitos, USA: IEEE, 2013a. p. 4233-4236.

MELLO, M. P.; RUDORFF, B. F. T.; ADAMI, M.; RIZZI, R.; AGUIAR, D. A.; GUSSO, A.; FONSECA, L. M. G. A simplified Bayesian Network to map soybean plantations. In: IEEE INTERNATIONAL GEOSCIENCE AND REMOTE SENSING SYMPOSIUM (IGARSS 2010), 30., 2010, Honolulu, HI, USA. **Proceedings...** Los Alamitos, USA: IEEE, 2010a. p. 351–354.

MELLO, M. P.; RUDORFF, B. F. T.; VIEIRA, C. A. O.; AGUIAR, D. A. Classificação automática da colheita da cana-de-açúcar utilizando Modelo Linear de Mistura Espectral. **Revista Brasileira de Cartografia**, v. 62, n. 2, p. 181–188, 2010b.

MELLO, M. P.; VIEIRA, C. A. O.; RUDORFF, B. F. T.; APLIN, P.; SANTOS, R. D. C.; AGUIAR, D. A. STARS: a new method for multitemporal remote sensing. **IEEE Transactions on Geoscience and Remote Sensing**, v. 51, n. 4, p. 1897–1913, 2013b.

MINISTÉRIO DO MEIO AMBIENTE (MMA). **Download de dados geográficos**. Available at <http://mapas.mma.gov.br/i3geo/datadownload.htm>. Accessed on: 20 Jan., 2013.

MORTON, D. C.; DEFRIES, R. S.; SHIMABUKURO, Y. E.; ANDERSON, L. O.; ARAI, E.; DEL BON ESPIRITO-SANTO, F.; FREITAS, R.; MORISETTE, J. Cropland expansion changes deforestation dynamics in the southern Brazilian Amazon. **Proceedings of the National Academy of Sciences**, v. 103, n. 39, p. 14637–14641, 2006.

MOUNTRAKIS, G.; IM, J.; OGOLE, C. Support vector machines in remote sensing: a review. **ISPRS Journal of Photogrammetry and Remote Sensing**, v. 66, n. 3, p. 247-259, 2011.

NEAPOLITAN, R. E. **Learning Bayesian networks**. Upper Saddle River, NJ, USA: Prentice Hall, 2003. 674 p.

NELLIS, M. D.; PRICE, K. P.; RUNDQUIST, D. Remote Sensing of Cropland Agriculture. In: WARNER, T. A.; NELLIS, M. D.; FOODY, G. M. (Org.). **The SAGE Handbook of Remote Sensing**. London, England: SAGE Publications Ltd, 2009. p. 368–383.

NIELSEN, A. A.; CANTY, M. J. Multi- and hyperspectral remote sensing change detection with generalized difference images by the IR-MAD method. In: INTERNATIONAL WORKSHOP ON THE ANALYSIS OF MULTI-TEMPORAL REMOTE SENSING IMAGES (Multitemp 2005), 3., 2005, Biloxi, MS, USA. **Proceedings...** Los Alamitos, USA: IEEE, 2005. p. 169–173.

OLIVEIRA, F. S. C.; GHERARDI, D. F. M.; STECH, J. L. The relationship between multi-sensor satellite data and Bayesian estimates for skipjack tuna catches in the South Brazil Bight. **International Journal of Remote Sensing**, v. 31, n. 15, p. 4049–4067, 2010.

OLIVEIRA, L. G. L.; PONZONI, F. J.; MORAES, E. C. Conversão de dados radiométricos orbitais por diferentes metodologias de caracterização atmosférica. **Revista Brasileira de Geofísica**, v. 27, n. 1, p. 121–133, 2009.

PALMIERE, F.; SANTOS, H. G.; GOMES, I. A.; LUMBRERAS, J. F.; AGLIO, M. M. D. The Brazilian soil classification system. In: RICE, T.; ESWARAN, H.; STEWART, B.; AHRENS, R. (Org.). **Soil classification:** a global desk reference. 1. ed. Boca Raton, FL, USA: CRC Press, 2002. p. 127–146.

PARK, S. J.; MCSWEENEY, K.; LOWERY, B. Identification of the spatial distribution of soils using a process-based terrain characterization. **Geoderma**, v. 103, n. 3-4, p. 249–272, 2001.

PEARL, J. **Probabilistic reasoning in intelligent systems:** networks of plausible inference. 1. ed. San Francisco, CA, USA: Morgan Kaufmann, 1988. 552 p.

PINO, F. A. Estatísticas agrícolas para o século XXI. **Agricultura em São Paulo**, v. 46, n. 2, p. 71–105, 1999.

PITARCH, Y.; VINTROU, E.; BADRA, F.; BÉGUÉ, A.; TEISSEIRE, M. Mining sequential patterns from MODIS time series for cultivated area mapping. In: GEERTMAN, S.; REINHARDT, W.; TOPPEN, F. (Org.). **Advancing Geoinformation Science for a Changing World**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011. p. 45–62.

PONZONI, F. J.; SHIMABUKURO, Y. E. **Sensoriamento remoto no estudo da vegetaçao**. 1. ed. São José dos Campos, SP, Brazil: Parêntese, 2007. 144 p.

QUINLAN, J. R. **C4.5:** programs for machine learning. San Mateo, CA, USA: Morgan Kaufmann, 1993. 302 p.

R CORE TEAM. **R**: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Available at: <http://www.r-project.org>. Accessed on: 30 Apr., 2013.

RABUS, B.; EINEDER, M.; ROTH, A.; BAMLER, R. The shuttle radar topography mission - a new class of digital elevation models acquired by spaceborne radar. **ISPRS Journal of Photogrammetry and Remote Sensing**, v. 57, n. 4, p. 241–262, 2003.

RICHARDS, J. A. Analysis of remotely sensed data: the formative decades and the future. **IEEE Transactions on Geoscience and Remote Sensing**, v. 43, n. 3, p. 422-432, 2005.

RISSO, J. **Diagnóstico espacialmente explícito da expansão da soja no Mato Grosso de 2000 a 2012**. 2013. 110 p. (sid.inpe.br/mtc-m19/2013/02.27.01.54-TDI). Dissertation (M.Sc. in Remote Sensing) – National Institute for Space Research (INPE), São José dos Campos, SP, Brazil, 2013.

RISSO, J.; RIZZI, R.; RUDORFF, B. F. T.; ADAMI, M.; SHIMABUKURO, Y. E.; FORMAGGIO, A. R.; EPIPHANIO, R. D. V. Índices de vegetação Modis aplicados na discriminação de áreas de soja. **Pesquisa Agropecuária Brasileira**, v. 47, n. 9, p. 1317–1326, set. 2012.

RIZZI, R.; RISSO, J.; EPIPHANIO, R. D. V.; RUDORFF, B. F. T.; FORMAGGIO, A. R.; SHIMABUKURO, Y. E.; FERNANDES, S. L. Estimativa da área de soja no Mato Grosso por meio de imagens MODIS. In: BRAZILIAN REMOTE SENSING SYMPOSIUM (SBSR 2009), 14., 2009, Natal, RN, Brazil. **Proceedings...** São José dos Campos, SP, Brazil: INPE, 2009. p. 387–394.

RIZZI, R.; RUDORFF, B. F. T. Estimativa da área de soja no Rio Grande do Sul por meio de imagens Landsat. **Revista Brasileira de Cartografia**, v. 57, n. 3, p. 226–234, dez. 2005.

RIZZI, R.; RUDORFF, B. F. T.; SHIMABUKURO, Y. E.; DORAISWAMY, P. C. Assessment of MODIS LAI retrievals over soybean crop in Southern Brazil. **International Journal of Remote Sensing**, v. 27, n. 19, p. 4091–4100, 2006.

ROUSE JR, J. W.; HAAS, R. H.; SCHELL, J. A.; DEERING, D. W. Monitoring vegetation systems in the great plains with ERTS. In: EARTH RESOURCES TECHNOLOGY SATELLITE-1 SYMPOSIUM, 3., 1973, Washington, DC, USA. **Proceedings...** Washington, DC, USA: NASA, 1973. p. 309–317.

RUDORFF, B. F. T.; ADAMI, M.; AGUIAR, D. A.; MOREIRA, M. A.; MELLO, M. P.; FABIANI, L.; AMARAL, D. F.; PIRES, B. M. The Soy Moratorium in the Amazon biome monitored by remote sensing images. **Remote Sensing**, v. 3, n. 1, p. 185–202, 2011.

RUDORFF, B. F. T.; ADAMI, M.; RISSO, J.; AGUIAR, D. A.; PIRES, B.; AMARAL, D.; FABIANI, L.; CECARELLI, I. Semote sensing images to detect soy plantations in the Amazon biome—The Soy Moratorium Initiative. **Sustainability**, v. 4, n. 12, p. 1074–1088, 2012.

RUDORFF, B. F. T.; AGUIAR, D. A.; SILVA, W. F.; SUGAWARA, L. M.; ADAMI, M.; MOREIRA, M. A. Studies on the rapid expansion of sugarcane for ethanol production in São Paulo State (Brazil) using Landsat data. **Remote Sensing**, v. 2, n. 4, p. 1057–1076, 2010.

RUDORFF, B. F. T.; SHIMABUKURO, Y. E.; CEBALLOS, J. C. (Org.). **O sensor MODIS e suas aplicaçoes ambientais no Brasil**. São José dos Campos, SP, Brazil: Parêntese Editora, 2007. 424 p.

RUDORFF, C. M.; RIZZI, R.; RUDORFF, B. F. T.; SUGAWARA, L. M.; VIEIRA, C. A. O. Superfícies de resposta espectro-temporal de imagens do sensor MODIS para classificação de área de soja no Estado do Rio Grande do Sul. **Ciência Rural**, v. 37, n. 1, p. 118–125, 2007.

SALMON, B. P.; OLIVIER, J. C.; WESSELS, K. J.; KLEYNHANS, W.; VAN DEN BERGH, F.; STEENKAMP, K. C. unsupervised land cover change detection: meaningful sequential time series analysis. **IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing**, v. 4, n. 2, p. 327–335, 2011.

SANO, E. E.; FERREIRA, L. G.; ASNER, G. P.; STEINKE, E. T. Spatial and temporal probabilities of obtaining cloud-free Landsat images over the Brazilian tropical savanna. **International Journal of Remote Sensing**, v. 28, n. 12, p. 2739–2752, 2007.

SANTOS, H. G.; OLIVEIRA, J. B.; LUMBRELAS, J. F.; ANJOS, L. H. C.; COELHO, M. R.; JACOMINE, P. K. T.; CUNHA, T. J. F.; OLIVEIRA, V. Á. (Org.). **Sistema Brasileiro de Classificação de Solos**. 2. ed. Rio de Janeiro, RJ, Brazil: Embrapa Solos, 2006. 306 p.

SANTOS, J. A.; GOSSELIN, P.-H.; PHILIPP-FOLIGUET, S.; DA S. TORRES, R.; FALAO, A. X. Multiscale Classification of Remote Sensing Images. **IEEE Transactions on Geoscience and Remote Sensing**, v. 50, n. 10, p. 3764–3775, 2012.

SCHROEDER, T.; COHEN, W.; SONG, C.; CANTY, M. J.; YANG, Z. Radiometric correction of multi-temporal Landsat data for characterization of early successional forest patterns in western Oregon. **Remote Sensing of Environment**, v. 103, n. 1, p. 16–26, 2006.

SEERUTTUN, S.; CROSSLEY, C. P. Use of digital terrain modelling for farm planning for mechanical harvest of sugar cane in Mauritius. **Computers and Electronics in Agriculture**, v. 18, n. 1, p. 29–42, 1997.

SELLERS, P. J.; MEESON, B. W.; HALL, F. G.; ASRAR, G.; MURPHY, R. E.; SCHIFFER, R. A.; BRETHERTON, F. P.; DICKINSON, R. E.; ELLINGSON, R. G.; FIELD, C. B.; HUEMMRICH, K. F.; JUSTICE, C. O.; MELACK, J. M.; ROULET, N. T.; SCHIMEL, D. S.; TRY, P. D. Remote sensing of the land surface for studies of global change: models - algorithms - experiments. **Remote Sensing of Environment**, v. 51, n. 1, p. 3–26, 1995.

SEPLAN-MT. **Sistema Interoperável de Informações Geoespaciais do Estado do Mato Grosso (SIIGEO)**. Available at: <http://www.siigeo.mt.gov.br/>. Accessed on: 14 Apr., 2012.

SHAXSON, F. **New concepts and approaches to land management in the tropics with emphasis on steeplands**. Rome, Italy: FAO, 1999. 125 p.

SHIMABUKURO, Y. E.; BATISTA, G. T.; MELLO, E. M. K.; MOREIRA, J. C.; DUARTE, V. Using shade fraction image segmentation to evaluate deforestation in Landsat Thematic Mapper images of the Amazon Region. **International Journal of Remote Sensing**, v. 19, n. 3, p. 535–541, 1998.

SHIMABUKURO, Y. E.; SMITH, J. A. The least-squares mixing models to generate fraction images derived from remote sensing multispectral data. **IEEE Transactions on Geoscience and Remote Sensing**, v. 29, n. 1, p. 16–20, 1991.

SILVA, J. A. A.; NOBRE, A. D.; JOLY, C. A.; NOBRE, C. A.; MANZATTO, C. V.; RECH FILHO, E. L.; SKORUPA, L. A.; MAY, P. H.; CUNHA, M. M. L. C.; RODRIGUES, R. R.; AHRENS, S.; SÁ, T. D. A.; AB'SÁBER, A. N. **Brazil forest code and science:** contributions to the dialogue. 2. ed. São Paulo, SP, Brazil: The Brazilian Society for the Advancement of Science – SBPC, 2012. 147 p.

SILVA, M. P. S.; CAMARA, G.; ESCADA, M. I. S.; SOUZA, R. C. M. Remote-sensing image mining: detecting agents of land-use change in tropical forest areas. **International Journal of Remote Sensing**, v. 29, n. 16, p. 4803–4822, 2008.

SILVESTRINI, R. A.; SOARES-FILHO, B. S.; NEPSTAD, D.; COE, M.; RODRIGUES, H.; ASSUNÇÃO, R. Simulating fire regimes in the Amazon in response to climate change and deforestation. **Ecological applications**, v. 21, n. 5, p. 1573–90, 2011.

SMITS, P.; BRUZZONE, L. (Org.). **Analysis of multi-temporal remote sensing images**: proceedings of Multitemp 2003. Ispra, Italy: World Scientific Publishing Company, 2004. 404 p.

SMITS, P. C.; DELLEPIANE, S. G.; SCHOWENGERDT, R. A. Quality assessment of image classification algorithms for land-cover mapping: A review and a proposal for a cost-based approach. **International Journal of Remote Sensing**, v. 20, n. 8, p. 1461-1486, jan. 1999.

SOARES-FILHO, B. S.; NEPSTAD, D. C.; CURRAN, L. M.; CERQUEIRA, G. C.; GARCIA, R. A.; RAMOS, C. A.; VOLL, E.; MCDONALD, A.; LEFEBVRE, P.; SCHLESINGER, P. Modelling conservation in the Amazon basin. **Nature**, v. 440, n. 7083, p. 520–3, 23 mar. 2006.

SONG, C.; WOODCOCK, C. E.; SETO, K. C.; LENNEY, M. P.; MACOMBER, S. A. Classification and change detection using Landsat TM data: When and how to correct atmospheric effects? **Remote Sensing of Environment**, v. 75, n. 2, p. 230–244, 2001.

SOUTH, S.; QI, J.; LUSCH, D. P. Optimal classification methods for mapping agricultural tillage practices. **Remote Sensing of Environment**, v. 91, n. 1, p. 90–97, 2004.

TAN, P.-N.; STEINBACH, M.; KUMAR, V. **Introduction to data mining**. 1. ed. Boston, MA, USA: Addison Wesley, 2006. 769 p.

TSO, B.; MATHER, P. M. **Classification methods for remotely sensed data**. 2. ed. Boca Raton, FL, USA: CRC, 2009. 356 p.

URIARTE, M.; YACKULIC, C. B.; COOPER, T.; FLYNN, D.; CORTES, M.; CRK, T.; CULLMAN, G.; MCGINTY, M.; SIRCELY, J. Expansion of sugarcane production in São Paulo, Brazil: implications for fire occurrence and respiratory health. **Agriculture, Ecosystems & Environment**, v. 132, n. 1-2, p. 48–56, 2009.

UUSITALO, L. Advantages and challenges of Bayesian networks in environmental modelling. **Ecological Modelling**, v. 203, n. 3-4, p. 312–318, 2007.

VERMOTE, E. F.; TANRE, D.; DEUZE, J. L.; HERMAN, M.; MORCETTE, J.-J. Second Simulation of the Satellite Signal in the Solar Spectrum, 6S: an overview. **IEEE Transactions on Geoscience and Remote Sensing**, v. 35, n. 3, p. 675–686, 1997.

VIEIRA, C. A. O. **Accuracy of remotely sensing classification of agricultural crops:** a comparative study. 2000. 323 p. Thesis (Ph.D. in Geography) - University of Nottingham, Nottingham, England, 2000.

VIEIRA, M. A.; FORMAGGIO, A. R.; RENNÓ, C. D.; ATZBERGER, C.; AGUIAR, D. A.; MELLO, M. P. Object Based Image Analysis and Data Mining applied to a remotely sensed Landsat time-series to map sugarcane over large areas. **Remote Sensing of Environment**, v. 123, p. 553–562, 2012.

WALTER, A.; ROSILLO-CALLE, F.; DOLZAN, P.; PIACENTE, E.; CUNHA, K. B. Perspectives on fuel ethanol consumption and trade. **Biomass and Bioenergy**, v. 32, n. 8, p. 730–748, 2008.

WARDLOW, B. D.; EGBERT, S. L.; KASTENS, J. H. Analysis of time-series MODIS 250 m vegetation index data for crop classification in the U.S. Central Great Plains. **Remote Sensing of Environment**, v. 108, n. 3, p. 290–310, 2007.

WATSON, D. F. **Contouring:** a guide to the analysis and display of spatial data - with programs on diskette. 1. ed. Oxford, UK: Pergamon, 1992. 321 p.

WILKINSON, G. G. Results and implications of a study of fifteen years of satellite image classification experiments. **IEEE Transactions on Geoscience and Remote Sensing**, v. 43, n. 3, p. 433–440, 2005.

WITTEN, I. H.; FRANK, E.; HALL, M. A. **Data mining**: practical machine learning tools and techniques. 3. ed. San Francisco, CA, USA: Morgan Kaufmann, 2011. 664 p.

XAVIER, A. C.; RUDORFF, B. F. T.; SHIMABUKURO, Y. E.; BERKA, L. M. S.; MOREIRA, M. A. Multi-temporal analysis of MODIS data to classify sugarcane crop. **International Journal of Remote Sensing**, v. 27, n. 4, p. 755–768, 2006.

YANG, C.; LU, L.; LIN, H.; GUAN, R.; SHI, X.; LIANG, Y. A Fuzzy-Statistics-based Principal Component Analysis (FS-PCA) method for multispectral image enhancement and display. **IEEE Transactions on Geoscience and Remote Sensing**, v. 46, n. 11, p. 3937–3947, 2008.

YATA, K.; AOSHIMA, M. PCA consistency for non-Gaussian data in high dimension, low sample size context. **Communications in Statistics - Theory and Methods**, v. 38, n. 16-17, p. 2634–2652, 2009.

ZWEIG, M. H.; CAMPBELL, G. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. **Clinical Chemistry**, v. 39, n. 8, p. 561–577, 1993.