Redes Bayesianas no mapeamento de culturas de verão no Estado do Paraná

Alexsandro Cândido de Oliveira Silva[1]
Leila Maria Garcia Fonseca[1]
Marcio Pupin Mello[2]

[1]Instituto Nacional de Pesquisas Espaciais - INPE
Caixa Postal 515 - 12227-010 - São José dos Campos - SP, Brasil
{acos,leila}@dpi.inpe.br

[2]Boeing Pesquisa & Tecnologia – Brasil (BR&TB)
Estrada Dr Altino Bondesan 500 – 12247-016 – São José dos Campos, SP – Brasil
marcio.p.mello@boeing.com

Abstract: In Brazil, the methodologies employed to obtain official agricultural statistics are subjective, and take a long time to be realized. Remote sensing technologies, combined with artificial intelligence, allow quick and accurate outcomes, which may help these methodologies to be more efficient. This paper aims at proposing the use of BayNeRD (Bayesian Network for Raster Data) algorithm to map summer crops areas (soybean and maize) in Paraná State – Brazil. BayNeRD is a computer-aided Bayesian Network method that is able to incorporate expert's knowledge to handle with raster data. The main outcome of BayNeRD is a probability image, wherein each pixel contains the probability of occurrence of target under study. Based on observations of a vegetation index, terrain slope, soil aptitude and other variables, BayNeRD was able to map soybean and maize plantations in Paraná State with 82% of sensitivity and 85% of specificity. Moreover, the probability image showed strong adherence to the reference data used for accuracy assessment and to the literature, denoting BayNeRD's potential to be applicable for agricultural inference through remote sensoring and ancillary data.

Palavras-chave: Bayesian Networks, inference, raster data, Redes Bayesianas, inferência, dados raster

1. Introdução

Atualmente, as previsões de safras são feitas por órgãos oficiais como o Instituto Brasileiro de Geografia e Estatística – IBGE e a Companhia Nacional de Abastecimento – CONAB. São realizadas pesquisas com produtores rurais e cooperativas, com dados de financiamento agrícola e dados históricos. Esta metodologia torna-se cara, demanda tempo para realizá-la e a análise de erros e incertezas envolvidas neste processo não é executável (EPIPHANIO et al., 2010). Portanto, vê-se a necessidade de incorporar novos métodos para complementá-la.

Produtos de Sensoriamento Remoto permitem resultados rápidos, precisos e de baixo custo, além de fornecerem a localização e quantificação das áreas plantadas, o que facilita e direciona o trabalho de agentes envolvidos no segmento agrícola (EPIPHANIO et al., 2010; ESQUERDO et al., 2011; MELLO et al., 2013).

Alguns fenômenos sob estudo envolvem a análise conjunta de múltiplas variáveis. A complexidade das interações entre estas variáveis pode dificultar tratá-las por métodos convencionais. O uso de técnicas de inteligência artificial como a modelagem por Redes Bayesianas (RB) permitem o entendimento e inferência sobre tais fenômenos baseando-se em observações (MELLO et al., 2013).

As Redes Bayesianas (RB) usam a probabilidade como uma medida de incerteza. Uma RB pode ser definida como uma rede esquematizada graficamente por um Grafo Acíclico Direcionado (DAG, do inglês, *Directed Acyclic Graph*), no qual os nós representam as variáveis e os arcos as suas relações de (in)dependência condicional. O conteúdo de cada variável é representado por distribuições de probabilidade (AGUILERA et al., 2011; NEAPOLITAN, 2003; UUSITALO, 2007).

A utilidade das RBs reside na possibilidade de poder calcular, através do Teorema de Bayes, tanto a distribuição de probabilidade de nós filhos que recebam valores dos pais, como também a distribuição dos pais dado os valores dos filhos (UUSITALO, 2007). Isto é, a RB nos permite conhecer os efeitos dado as causas e as causas dados os efeitos (AGUILERA et al., 2011).

Apesar da emergência das RBs como método de análise de incertezas, Aguilera et al. (2011) apontaram que raramente têm sido aplicadas às Ciências de Observação da Terra. Neste contexto, este trabalho visa espacializar áreas com culturas de verão (soja e milho 1ª safra) no estado do Paraná através da modelagem Bayesiana como uma forma de auxiliar e direcionar as pesquisas de previsões de safras.

2. Materiais e métodos

O BayNeRD (*Bayesian Network for Raster Data*) é um algoritmo implementado no software R (http://www.r-project.org) e desenvolvido por Mello et al. (2013). O BayNeRD lê arquivos no formato GeoTiff, os quais correspondem às variáveis envolvidas. A variável que denota ao fenômeno de interesse é chamada de *variável alvo*, as demais são chamadas de *variáveis de contexto* (MELLO et al, 2013).

A distribuição conjunta de *n* variáveis de uma RB é calculada como um produto da distribuição de probabilidade de cada uma das *n* variáveis, conforme segue:

$$P(X_1 = x_1, \dots, X_n = x_n) = \prod_{i=1}^{n} P(x_i | pa_i) \quad (1)$$

Em (1), x_i é o valor observado para a variável X_i e pa_i é o conjunto de valores observados para os pais de X_i. Deste modo, a probabilidade conjunta de cada instanciação para as *n* variáveis da RB pode ser calculada como

$$P(X_1 | X_2, X_3, \dots, X_n) \quad (2)$$

Os dados utilizados como referência para treinamento da RB foram fornecidos pela CONAB e estão especializados na Figura 1. Tais dados correspondem às áreas de cultivo da soja e do milho 1ª safra no Estado do Paraná para o ano safra 2012/2013. Representando a *variável alvo* SM (soja-milho), os dados, originalmente vetoriais, foram rasterizados com pixels de 250x250 metros. Afim de evitar a correlação espacial, janelas de 4x4 pixels foram distribuídas aleatoriamente sobre o raster SM de modo que aproximadamente 66% dos pixels fossem designados para treinamento da RB e o restante para validação.

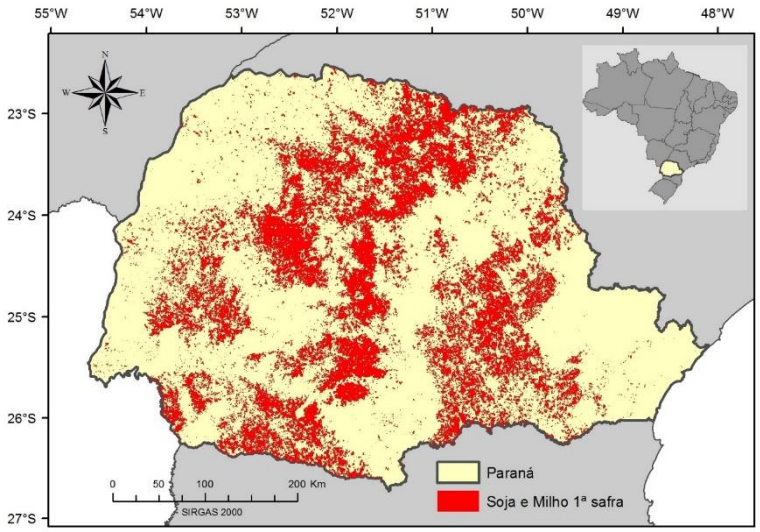

Figura1: Área de estudo e espacialização das plantações de soja e milho 1ª safra

Certas culturas tendem a seguir alguns padrões de cultivo como o plantio em baixa declividade, a distância de corpos d'água e rodovias, conforme observado por Mello et al. (2013) no mapeamento da soja no estado do Mato Grosso. Portanto, a ocorrência do fenômeno sob estudo pode ser representada através das relações de probabilidade entre a *variável alvo* e as *variáveis de contexto* envolvidas

Tabela 1: Variáveis de contexto

Variáveis	**Descrição**
C	Índice CEI no atual ano safra 12/13
CA	Índice CEI no ano safra anterior 11/12
D	Declividade
AS	Aptidão do solo
DU	Distância da área urbana mais próxima
DA	Distância do corpo d'água mais próximo
DR	Distância da rodovia mais próxima

O índice CEI – Crop Enhancement Index (RIZZI et al., 2009) foi desenvolvido para ter sensibilidade ao calendário agrícola da região e cultura sob estudo. Ele considera o comportamento temporal do EVI – Enhanced Vegetation Index (HUETE et al., 1997) no monitoramento da cultura ao longo do seu ciclo de desenvolvimento, especificamente em dois momentos: período de mínimo EVI, quando há o preparo do solo para plantio ou a planta está em fase inicial; e o período de máximo EVI, quando a cultura atinge seu vigor vegetativo.

Considerando o calendário agrícola do Paraná, fornecido pela Secretaria da Agricultura e do Abastecimento (SEAB-DERAL, 2013), não é possível discriminar a soja e o milho de primeira safra em todo o estado pela análise temporal. As épocas de plantio (agosto e setembro) e máximo vigor vegetativo (dezembro e janeiro) de ambas as culturas se sobrepõem. Para compor as imagens de mínimo e máximo EVI foram selecionadas imagens no Banco de Produtos MODIS da Embrapa (http://www.modis.cnptia.embrapa.br).

Para obter os valores de CEI da safra 2012/2013 (C), o mínimo EVI foi calculado com imagens entre 28 de agosto de 2012 e 18 de dezembro de 2012; enquanto o máximo EVI foi calculado com imagens entre 16 de novembro de 2012 e 06 de março de 2013. Os valores de CEI da safra 2011/2012 (CA) foram obtidos com imagens EVI em um período semelhante ao anterior. Apenas os pixels com boa qualidade foram selecionados em cada imagem EVI. Uma máscara foi criada sobre as imagens considerando a imagem de confiabilidade dos pixels também disponível no Banco de Produtos MODIS da Embrapa.

Os dados de declividade (D) foram obtidos no Banco de Dados Geomorfométricos do Brasil – TOPODATA (http://www.dsr.inpe.br/topodata/). Fez-se um mosaico das folhas que cobriam todo o estado do Paraná o qual foi reamostrado para pixels de 250x250 metros.

A imagem de aptidão do solo (AS) do estado Paraná foi construída considerando o tipo de solo, a capacidade de retenção de água, a fertilidade e a textura. Os dados foram obtidos junto ao Instituto de Terras, Cartografia e Geociências (http://www.itcg.pr.gov.br/). Como resultado obteve-se um raster com três classes: alta, média e baixa aptidão.

Por fim, foram gerados três dados rasters cujos pixels, respectivamente, continham a distância da área urbana mais próxima (DU), distância do corpo d'água mais próximo (DA) e distância da estrada mais próxima (DR). Para isso, os dados vetoriais de Manchas Urbanas, de Rede Hidrográfica e de Malha Rodoviária foram adquiridos junto ao Sistema Nacional de Informações das Cidades (http://www.brasilemcidades.gov.br).

O próximo passo, após a inserção das variáveis, é a definição do modelo gráfico da RB. O BayNeRD interage com o usuário para defini-lo através da construção das relações de (in)dependência entre as variáveis. A Figura 2 mostra o modelo gráfico definido para esta aplicação.

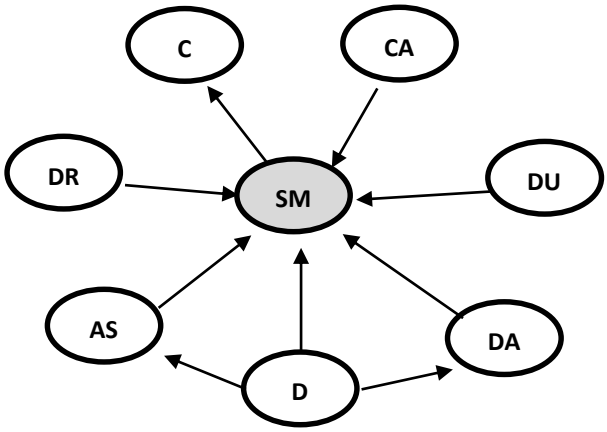

Figura 2: Modelo gráfico da Rede Bayesiana.

Com a definição do modelo gráfico, o algoritmo é treinado calculando os valores de probabilidades envolvidos com base na discretização das variáveis e pela contagem de pixels (MELLO et al., 2010). O conhecimento do especialista é fundamental no processo de discretização das variáveis. O algoritmo é capaz de incorporar esse conhecimento através da construção do modelo gráfico e da discretização das *variáveis de contexto*. A Tabela 2 sumariza a discretização das variáveis.

Tabela 2: Discretização das variáveis de contexto.

Intervalos	**C**	**CA**	**D**	**AS**	**DU***	**DA***	**DR***
1	-∞ , 0.08	-∞ , 0.05	-∞ , 0.07	alta	-∞ , 3.6	-∞ , 0.25	-∞ , 2.0
2	0.08 , 0.2	0.05 , 0.2	0.07 , 4.5	média	3.6 , 18.5	0.25 , 0.75	2.0 , 5.0
3	0.2 , 0.28	0.2 , 0.3	4.5 , 6.5	baixa	18.5 , ∞	0.75 , 3.5	5.0 , 8.3
4	0.28 , +∞	0.3 , +∞	6.5 , +∞			3.5 , ∞	8.3 , +∞
						* valores em Km	

3. Resultados e Discussão

Após a definição das relações entre as variáveis e a discretização, o algoritmo calcula para cada pixel a probabilidade da presença do alvo dado os valores observados nas *variáveis de contexto*:

$$P(\ SM=1 \mid C=c,\ CA=ca,\ D=d,\ ApS=aps,\ DU=du,\ DA=da,\ DR=dr\ ) \quad (3)$$

em que as letras minúsculas (cc, lc, s, as, du, dw, dr) representam um valor instanciado (observado) das respectivas variáveis (representadas por letras maiúsculas). Quando as probabilidades são calculadas para cada pixel da área de estudo forma-se a Imagem de Probabilidade – IP, apresentada na Figura 3.

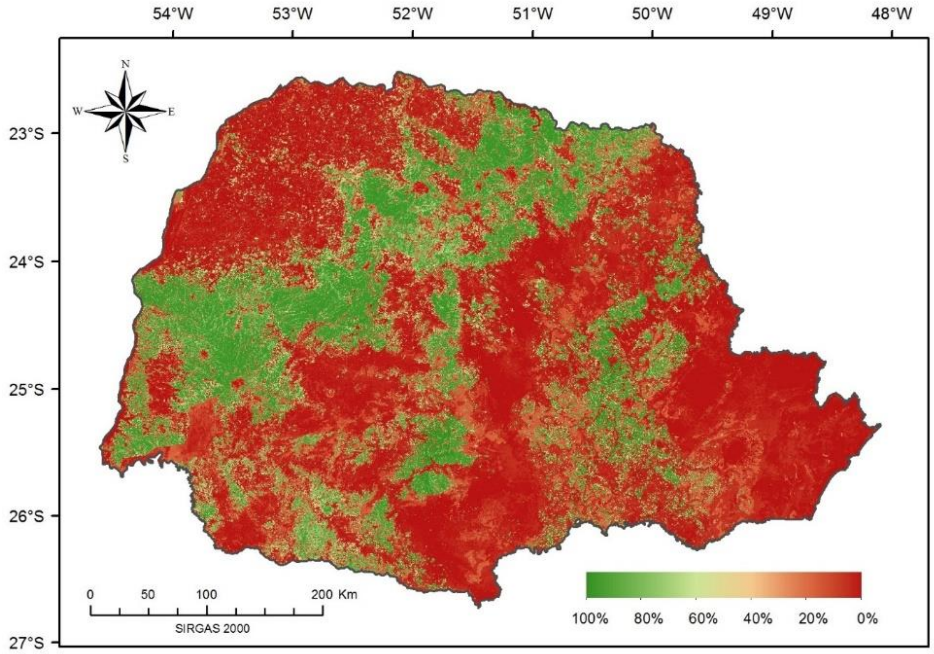

Figura 3: Imagem de Probabilidade

A IP é o principal produto computado pelo algoritmo BayNeRD. A IP mostra a distribuição espacial das plantações de soja e milho 1ª safra no Estado do Paraná no ano-safra 2012/2013. Pixels com tonalidades verdes representam áreas com alta probabilidade da presença do alvo baseado nas observações das variáveis de contexto.

Através da análise entre a IP e os dados de referência (Figura 1), pode-se perceber a diferença de concentração das áreas de soja e milho na região oeste do estado. Johann et al. (2012) mapeou as áreas de ambas as culturas no Paraná para o ano-safra 2007/2008 com imagens multitemporais EVI/MODIS. A Figura 4 mostra uma composição RGB das imagens EVI/MODIS juntamente com a máscara para as culturas. A distribuição espacial do alvo na IP se mostra coerente com resultados de Johann et al. (2012) evidenciando a existência do "cinturão de soja" que ocorre desde a região oeste à região norte do Paraná, além das regiões Centro-Oriental e Centro-Sul.

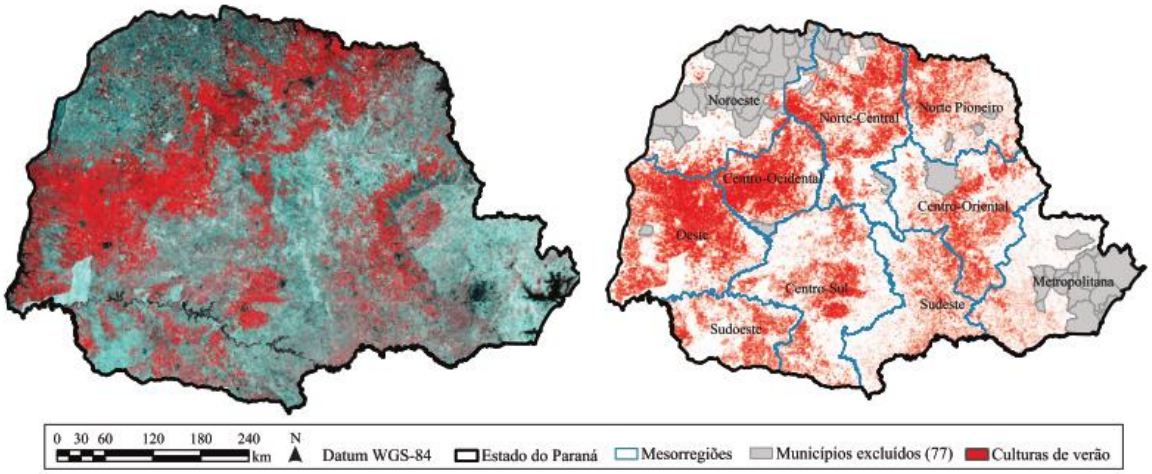

Figura 4: Composição RGB – MaxEVI (R) MinEVI (GB) e Máscara MODIS para as culturas de verão soja e milho safra 2007/2008. Fonte: Johann et al. (2012)

A IP pode ser usada para produzir um mapa temático através da escolha de um limiar de probabilidade. Pixels com valores acima deste limiar são rotulados como presença do alvo e valores abaixo como ausência do alvo. Este limiar foi nomeado TPV, do inglês *Target Probability Value*. O melhor TPV produz o mapa temático mais adequado. No algoritmo BayNeRD há seis critérios para escolha do melhor TPV, dentre eles os índices sensibilidade e especificidade (MELLO et al., 2013).

Foody (2002) apontou estes dois índices complementares para uma classificação binária. Estes índices indicam a habilidade de encontrar verdades positivas (áreas com presença do alvo que foram corretamente classificadas como presença do alvo) e verdades negativas (áreas com ausência do alvo que foram corretamente classificadas como ausência do alvo). O mapa ideal seria um TPV mais próximo de 100% de sensibilidade e 100% de especificidade. A Figura 4 mostra a curva da variação dos índices: com um melhor TPV de 20%, sensibilidade e especificidade de 82,04% e 84,95%, respectivamente. A Figura 5 mostra o mapa temático produzido a partir do limiar selecionado.

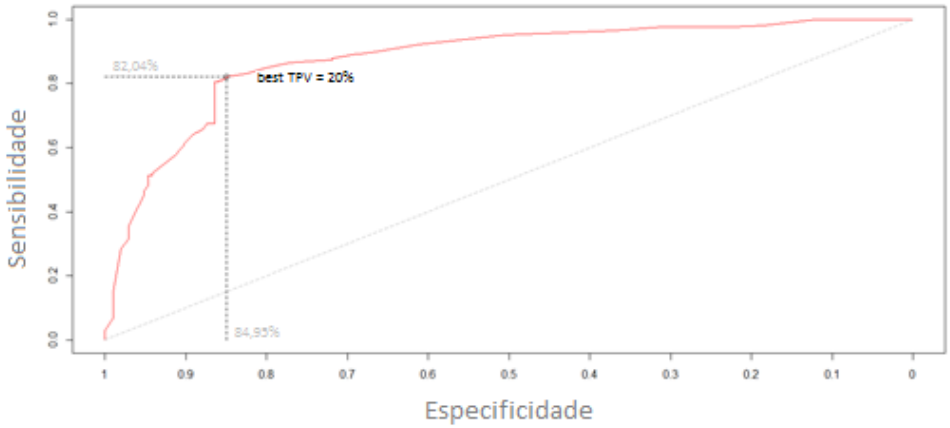

Figura 4: Variação dos índices de sensibilidade e Especificidade

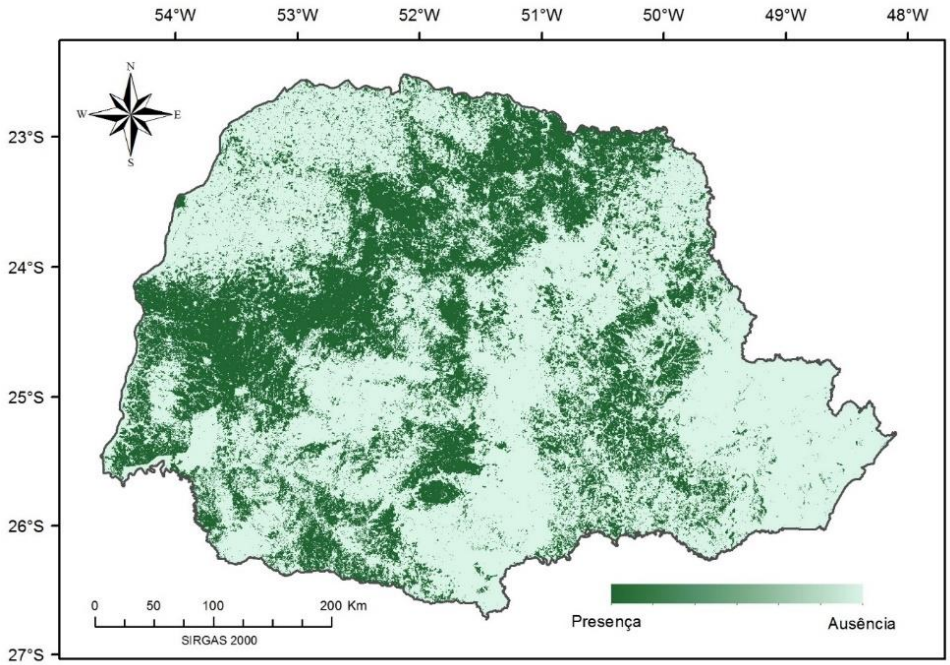

Figura 5: Mapa temático

4. Conclusão

Neste estudo realizou-se o mapeamento de culturas de verão no estado do Paraná utilizando uma abordagem de Redes Bayesianas com o intuito de contribuir com metodologias de previsão de safras. A ferramenta BayNeRD foi implementada de maneira a facilitar a interação com o usuário e a aplicação das RB no processo de inferência se baseando em observações de variáveis relacionadas com o fenômeno sob estudo. Os resultados alcançados foram coerentes e satisfatórios. O algoritmo foi capaz de indicar as áreas com presença de culturas de verão (soja-milho). Isto mostra o grande potencial das Redes Bayesianas em inferir através de observações sobre as variáveis de contexto.

5. Agradecimentos

Os autores agradecem ao CNPq (134400/2013-5) pelo suporte financeiro e às instituições onde os dados foram adquiridos.

6. Referências Bibliográficas

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, dez. 2011.

Araújo, G. K. D.; Rocha, J. V.; Lamparelli, R. A. C.; Rocha, A. M. Mapping of summer crops in the state of Paraná, Brazil, through the 10-day spot vegetation NDVI composites. **Engenharia Agrícola - online**, v. 31, n. 4, p. 760–770, 2011.

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.

Esquerdo, J. C. D. M.; Zullo Júnior, J.; Antunes, J. F. G. Use of NDVI/AVHRR time-series profiles for soybean crop monitoring in Brazil. **International Journal of Remote Sensing**, v. 32, n. 13, p. 3711–3727, 10 jul. 2011.

Foody, G. M. Status of land cover classification accuracy assessment. **Remote Sensing of Environment**, v. 80, n. 1, p. 185–201, abr. 2002.

Huete, A. R.; Liu, H. Q.; Batchily, K.; Leeuwen, W. Van. A Comparison of Vegetation Indices over a Global Set of TM Images for EOS-MODIS. **Remote Sensing of Environment**, v. 59, p. 440–451, 1997.

Johann, J. A.; Rocha, J. V.; Duft, D. G.; Lamparelli, R. A. C. Estimativa de áreas com culturas de verão no Paraná , por meio de imagens multitemporais EVI / Modis. **Pesquisa Agropecuária Brasileira**, v. 47, n. 1, p. 1295–1306, set. 2012.

Mello, M. P.; Risso, J.; Atzberger, C.; Aplin, P.; Pebesma, E.; Vieira, C. A. O.; Rudorff, B. F. T. Bayesian Networks for Raster Data (BayNeRD): Plausible Reasoning from Observations. **Remote Sensing**, v. 5, n. 11, p. 5999–6025, 15 nov. 2013.

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, 2010, Honolulu, HI, USA: **Anais…** IEEE, 2010. p. 351–354. E-ISBN:978-1-4244-9564-1. Disponível em: <http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=5651814>. Acesso em: 10 dez 2013.

Neapolitan, R. E. Learning **Bayesian Networks**. [S.l.]: Pearson Prentice Hall, 2003. p. 674

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: Simpósio Brasileiro de Sensoriamento Remoto (SBSR), 2009, Natal. Anais... São José dos Campos, 2009. Artigos. p. 387–394.

Disponível em: < http://marte.dpi.inpe.br/col/dpi.inpe.br/sbsr@80/2008/11.16.18.50.57/doc/387-394.pdf >. Acesso em 12 dez. 2013.

SEAB-DERAL, Secretaria de Estado da Agricultura e do Abastecimento do Paraná - Departamento de Economia Rural. **Calendário Agrícola - Evolução de plantio, colheita e comercialização**. . [S.l: s.n.], 2013. Disponível em: <http://www.agricultura.pr.gov.br/arquivos/File/deral/pss.xls>. Acesso em: 10 dez. 2013.

Sugawara, L. M.; Rudorff, B. F. T.; Adami, M. Viabilidade de uso de imagens do Landsat em mapeamento de área cultivada com soja no Estado do Paraná. **Pesquisa Agropecuária Brasileira**, v. 43, n. 12, p. 1777–1783, dez. 2008.

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