

Inteligência Artificial como ferramenta para direcionar a expansão sustentável da cana-de-açúcar no Estado de São Paulo

Alexsandro Cândido de Oliveira Silva¹

Marcio Pupin Mello²

Leila Maria Garcia Fonseca¹

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

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

Abstract. Predictive models have been used to understand several phenomena in the field of Earth sciences. Bayesian networks have become increasingly popular due to its potential to model such phenomena and, through graphical representation, state the relationship among variables with probabilistic models associated. This paper presents an improved algorithm of Bayesian networks, implemented in R software, able of handling raster data for remote sensing applications: e-BayNeRD (enhanced Bayesian Network for Raster Data). A case study was used to describe the main changes and test the enhanced version of the algorithm. Based on observed values for terrain slope, soil and fertility, edaphoclimatic aptitude and the Agri-environmental zoning, suitable areas for sustainable expansion of sugarcane in São Paulo State were mapped.

Palavras-chave: Bayesian Networks, inference, remote sensing, Redes Bayesianas, inferência, sensoriamento remoto.

1. Introdução

A variedade e complexidade dos fenômenos no domínio das ciências da Terra dificultam o uso de uma única abordagem para analisá-los. Assim, novas técnicas, como redes neurais e árvores de decisão, têm sido aplicadas para na modelagem de alguns desses fenômenos a partir do conhecimento prévio das variáveis envolvidas. Contudo, estes modelos possuem uma abordagem “caixa preta”, enquanto que as estruturas das Redes Bayesianas (RBs) representam uma estrutura lógica de relações de interdependência entre as variáveis (QIN et al., 2006; PEARL e RUSSELL, 2011).

As RBs são modelos gráficos cada vez mais populares devido à rigorosa base probabilística para tratar incertezas, o que as tornam aplicáveis na modelagem de uma ampla gama de fenômenos de interesse (AGUILERA et al., 2011; PEARL e RUSSELL, 2011). De fato, as RBs têm sido usadas em diferentes aplicações: para diagnóstico médico, no reconhecimento de fala e gestos, em ciências ambientais, etc. No entanto, apesar do considerável aumento de aplicações nos últimos anos, poucas delas têm sido voltadas à Ciência de Observação da Terra (QIN et al., 2006. AGUILERA et al., 2011).

Neste contexto, o presente trabalho visa apresentar a implementação de um método de RBs aprimorado capaz de incorporar o conhecimento do especialista para análise de dados *raster*, com foco em aplicações em sensoriamento remoto. O algoritmo, nomeado e-BayNeRD (*enhanced Bayesian Network for Raster Data* (SILVA et al., 2014)), é uma versão aprimorada do algoritmo BayNeRD apresentado por Mello et al., (2013). As principais melhorias implementadas serão brevemente apresentadas utilizando-se um estudo

de caso de mapeamento de áreas aptas à expansão sustentável da cana-de-açúcar no estado de São Paulo.

2. Materiais e métodos

2.1 Redes Bayesianas

RBs são modelos matemáticos graficamente representadas por um DAG (*Directed Acyclic Graph*), com nós correspondendo às variáveis e arcs às suas relações condicionais de dependência (Figura 1); além de distribuições de probabilidades associadas a cada variável (NEAPOLITAN, 2003; UUSITALO, 2007).

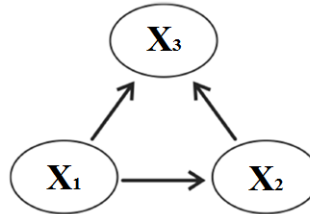


Figura 1: Exemplo de Rede Bayesiana

Se x_i é o valor instanciado para a variável X_i e pa_i é o conjunto de valores para os pais de X_i , então $P(x_i|pa_i)$ denota a distribuição condicional de X_i . A distribuição conjunta das n variáveis em uma RB é dada pelo produto:

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

De acordo com a RB exemplificada na Figura 1:

$$P(x_1, x_2, x_3) = P(x_1) * P(x_2|x_1) * P(x_3|x_1, x_2) \quad (2)$$

2.2 Algoritmo aprimorado

O algoritmo denominado e-BayNeRD (*enhanced Bayesian Network for Raster Data*), desenvolvido por Silva et al. (2014), foi implementado no software R (R CORE TEAM, 2014), mesmo ambiente de programação no qual foi desenvolvida a versão original – BayNeRD (MELLO et al., 2013).

No software R, já existem pacotes implementados para manipulação de dados espaciais, como o pacote *raster* (HIJMANS et al., 2014) e o pacote *rgdal* (BIVAND et al., 2014). Considerando as vantagens do primeiro sobre o segundo, o pacote *raster* foi incorporado ao e-BayNeRD. O algoritmo lê arquivos no formato GeoTiff que correspondem às variáveis envolvidas. A variável que representa o fenômeno de interesse é chamada de *variável alvo*, enquanto que as demais são nomeadas *variáveis de contexto*. Um dos principais benefícios do pacote *raster* no e-BayNeRD é a redução da etapa de pré-processamento dos dados, visto que o algoritmo transforma as variáveis para que tenham os mesmos parâmetros (resolução; extensão; sistemas de coordenadas).

Após a leitura das variáveis, o especialista deve definir as relações de dependência entre as mesmas através da construção do modelo gráfico da RB. Por meio dos pacotes *deal* (BOTTCHEER e DETHLEFSEN., 2013) e *bnlearn* (SCUTARI, 2009) foi possível implementar uma maneira mais eficiente de definir o modelo gráfico da RB. O primeiro permite o usuário especificar o DAG através de uma interface interativa clicando sobre as

variáveis. Enquanto que o pacote *bnlearn* é usado nas análises bayesianas e inferências, posteriormente. Assim, a tarefa de construção do modelo gráfico da RB se tornou mais fácil e rápido comparada à versão original do algoritmo.

Após a definição do modelo gráfico, o algoritmo calcula a probabilidade de ocorrência das classes da *variável alvo* com base na discretização das *variáveis de contexto* e na contagem de pixel (MELLO et al., 2010). O e-BayNeRD terá como resultado uma camada para cada classe da *variável alvo*. Estas camadas foram denominadas Bandas de Probabilidade (BP), que quando agrupadas formam a Imagem de Probabilidade (IP). Os pixels de cada BP conterão a probabilidade relacionada à cada classe da *variável alvo*, dados os valores observados para os mesmos pixels nas *variáveis de contexto*.

2.3 Estudo de caso

O uso de biocombustíveis tem se mostrado como uma opção viável na mitigação das emissões de gases de efeito estufa. A partir de 2003, quando os veículos *flex fuels* entraram no mercado brasileiro, houve uma enorme expansão das áreas de cultivo de cana-de-açúcar para produção do etanol (RUDORFF et al., 2010). Há também uma eminente demanda pelo uso dos biocombustíveis em outros setores de transportes, o que justifica estudos para encontrar caminhos para expansão sustentável do plantio das culturas energéticas. Considerando este fenômeno, o estudo de caso envolve a identificação de áreas aptas ao cultivo sustentável da cana-de-açúcar no estado de São Paulo (Figura 2), responsável por mais de 50% da produção nacional (IBGE, 2014).

A partir da necessidade estratégica de espacializar áreas com potencial de expansão sustentável da cana-de-açúcar, foi realizado o Zoneamento Agroecológico – ZAE-cana no território nacional (MANZATTO et al., 2009). A distribuição espacial destas áreas encontra-se na Figura 2, agrupadas em três classes de uso da terra: Agricultura, Agropecuária e Pecuária; dados fornecidos pela Embrapa Meio Ambiente (<http://geo.cnpma.embrapa.br/>).

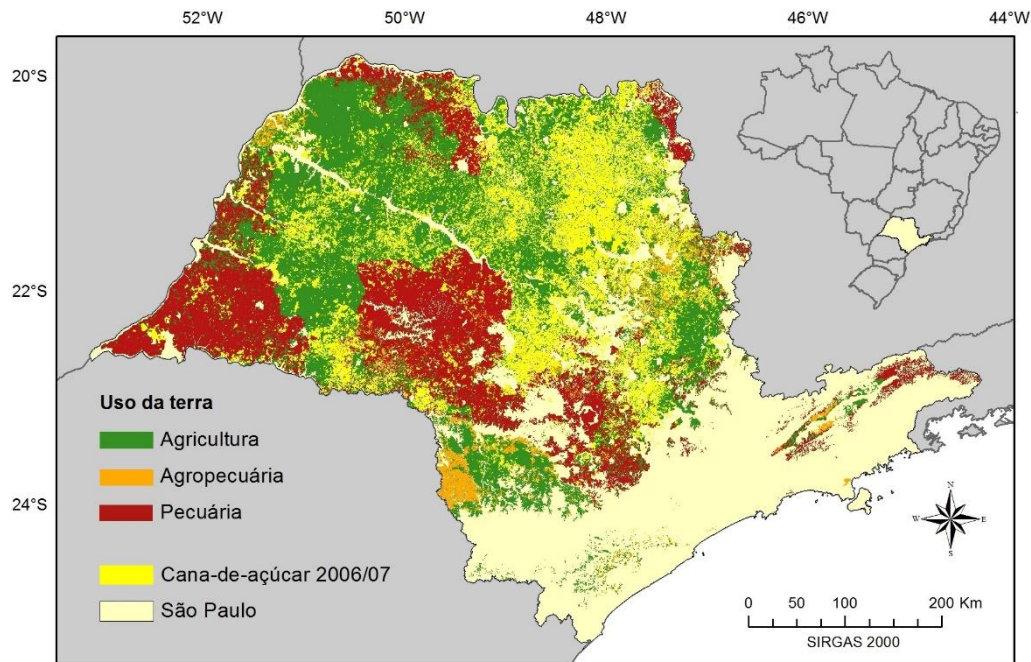


Figura 2: Localização da área de estudo; Zoneamento Agroecológico da Cana-de-açúcar; Polígonos de área de cana-de-açúcar.

A elaboração do ZAE-cana teve como base características dos solos, riscos climáticos, o potencial de produção agrícola sustentável. Adicionalmente algumas regiões foram excluídas: declividade superior a 12%, áreas de preservação, áreas urbanas, além das áreas com cultivo de cana-de-açúcar (ano-safra 2006/07), as quais também são apresentadas na Figura 2 (Projeto Canasat – www.dsr.inpe.br/canasat). Os polígonos de cana-de-açúcar foram agregados à classe Agricultura do ZAE-cana para comporem a *variável alvo* (dados de referência). Os polígonos foram rasterizados com pixels de 250x250 metros.

Para este trabalho foram consideradas quatro *variáveis de contexto* que, assume-se, tem relação com o fenômeno sob estudo (i.e., aptidão para o cultivo da cana-de-açúcar). São elas: Declividade (D), Solo/Fertilidade (SF), Zoneamento Agroambiental (ZAA) e Aptidão Edafoclimática (AEC).

Os dados de declividade foram obtidos no Banco de Dados Geomorfométricos do Brasil – TOPODATA (www.dsr.inpe.br/topodata). O mosaico de todas as folhas que cobrem o Estado de São Paulo foi reamostrado para pixels de 250x250 metros. Os dados de solo/fertilidade foram conseguidos juntamente com o Ministério do Meio Ambiente (<http://mapas.mma.gov.br/i3geo/datadownload.htm#>).

Através de uma parceria entre a Secretaria do Meio Ambiente e a Secretaria da Agricultura e Abastecimento, ambas do estado de São Paulo, também foram elaborados o Zoneamento Agroambiental da Cana-de-açúcar e a Aptidão Edafoclimática (www.ambiente.sp.gov.br/etanolverde/zoneamento-agroambiental/).

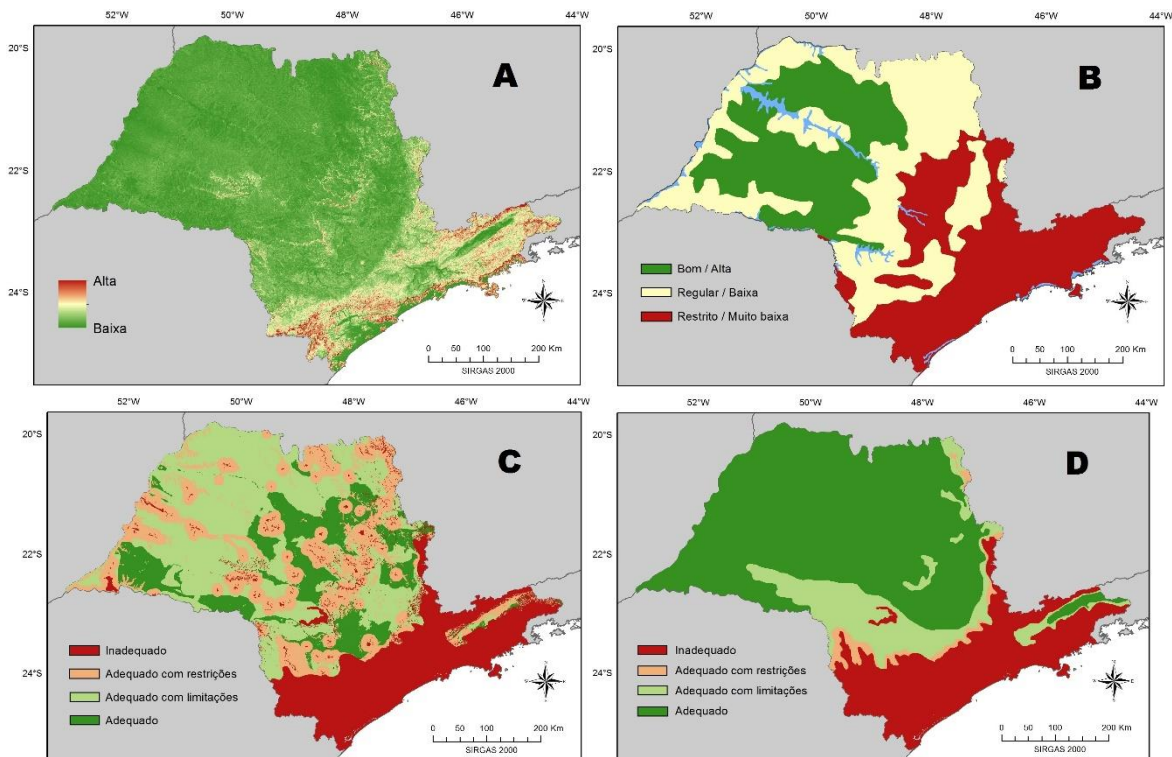


Figura 3: Variáveis de contexto envolvidas: A – Declividade; B – Solo/Fertilidade; C – Zoneamento Agroambiental; e D – Aptidão Edafoclimática.

A seguir definiu-se as relações condicionais entre as variáveis através da construção do modelo gráfico da RB apresentado na Figura 4. A incorporação do conhecimento se dá quando o especialista define quais as relações entre as variáveis e no processo de discretização das mesmas. Durante este processo, os intervalos podem ser definidos manualmente pelo usuário ou por critérios implementados no e-BayNeRD: intervalos equidistantes, frequência ou *clustering*. Assim, uma *variável de contexto* terá n categorias se discretizada em n intervalos. As funções de probabilidade são computadas pela contagem de pixels para cada variável e levando em conta suas relações de dependência. A Tabela 1 resume a discretização das variáveis.

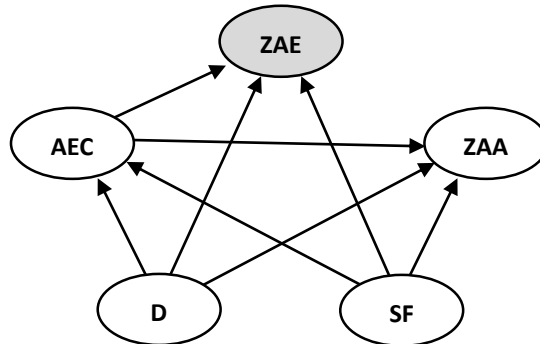


Figura 4: Modelo gráfico da Rede Bayesiana

Tabela 1: Discretização das variáveis de contexto

Intervalos	D	S/F	ZAA	AEC
1	0.01% - 6%	Bom/Alta	Adequado	Adequado
2	6% - 12%	Regular/Baixa	Adequado/limitações	Médio *
3	> 12%	Restrito/ Muito Baixa	Adequado/restrições	Inadequado
4	-	-	Inadequado	-

* Inclui: Adequado com limitações e Adequado com restrições

3. Resultados e Discussão

Outra melhoria do e-BayNeRD é a possibilidade de a *variável alvo* conter mais de duas classes. Isto é, considerando o *raster* ZAE-cana, seus pixels podem ser categorizados como 1 para classe Pecuária, 2 para Agropecuária, 3 para Agricultura e 0 para não-alvo. O que não é possível na primeira versão do algoritmo, no qual a *variável alvo* deve ser instanciada com dois valores: 1 – presença do alvo e 0 – ausência do alvo. Portanto, o número de valores instanciados para a *variável alvo* definirá o número de Bandas de Probabilidade (BP). O e-BayNeRD terá como resultado um dado *raster* com valores de probabilidade para cada classe da *variável alvo*.

Através do pacote *bnlearn*, são realizadas consultas às tabelas de probabilidades condicionais e o algoritmo retorna a probabilidade de um evento dado as evidências observadas. Segue que as BPs são calculadas como:

$$P(ZAE = 1 \mid D = d, AS = aps, ZAA = zaa, AEC = apec) \quad (3)$$

$$P(ZAE = 2 \mid D = d, AS = aps, ZAA = zaa, AEC = apec) \quad (4)$$

$$P(ZAE = 3 \mid D = d, AS = aps, ZAA = zaa, AEC = apec) \quad (5)$$

em que as letras minúsculas (*d, aps, zaa, ap ec*) representam os valores observados para o mesmo pixel nas respectivas *variáveis de contexto*. Quando as probabilidades são calculadas para cada pixel da área de estudo formam-se as BPs. As Figuras 5 e 6 apresentam as BPs referentes às classes Pecuária e Agricultura, respectivamente. Os valores de probabilidade associados à classe Agropecuária foram muito baixos, sendo o valor máximo igual a 0,1. Este resultado era esperado devido ao menor número de dados de referência da classe Agropecuária, frente às demais classes.

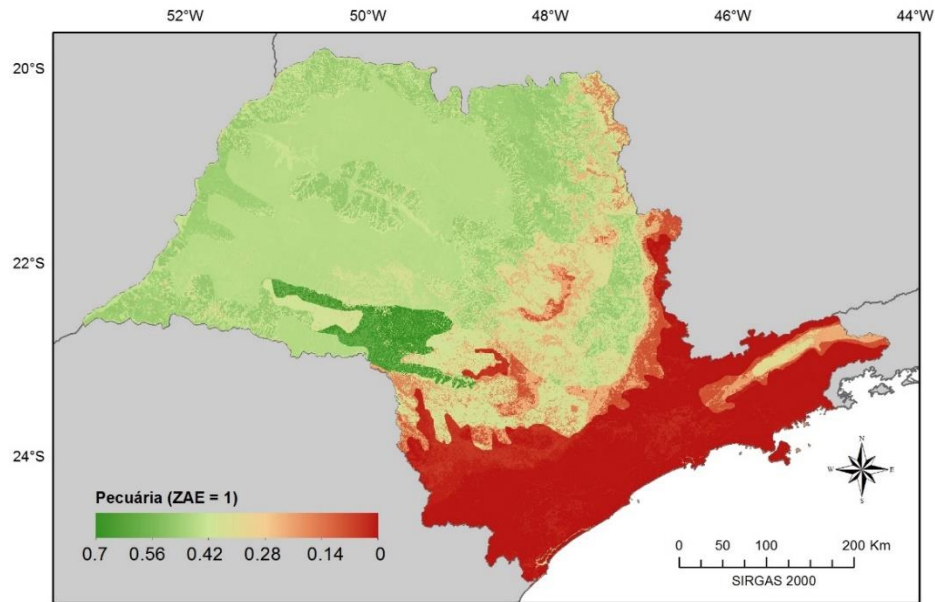


Figura 5: BP1 – valores de probabilidade associados à expansão da cana-de-açúcar sobre áreas de Pecuária ($ZAE = 1$)

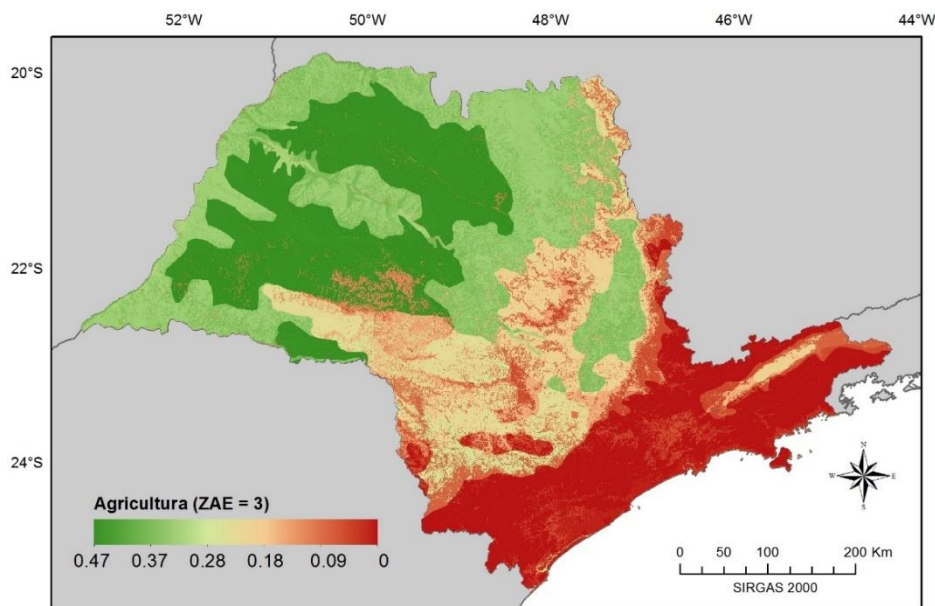


Figura 6: BP3 – valores de probabilidade associados à expansão da cana-de-açúcar sobre áreas de Agricultura ($ZAE = 3$)

Os cenários são formados através da combinação de todas as evidências possíveis (valores observados). Para um determinado cenário, existirão três eventos ($ZAE = 1$, $ZAE = 2$, e $ZAE = 3$) para calcular as probabilidades associadas. Consequentemente, as BPs são calculadas simultaneamente para otimizar o algoritmo. Se alguma *variável de contexto* contem qualquer pixel sem informação (dato ausente), ainda assim a probabilidade para este pixel será calculada. O e-BayNeRD é capaz de calcular $P(ZAE = zae)$ para pixels sem qualquer observação (i.e., com dato ausente para todas as variáveis de contexto) através da probabilidade a priori.

As tonalidades em verde nas Figuras 5 e 6 representam as áreas com maiores probabilidades de serem aptas à expansão da cana-de-açúcar. Nota-se que os valores de probabilidades associados ao evento $ZAE = 1$ (classe Pecuária – Figura 5) são superiores àqueles referentes ao evento $ZAE = 3$ (classe Agricultura – Figura 6). Mesmo considerando as regiões mais esverdeadas de BP3 (Figura 6), estas regiões apresentam valores maiores em BP1 (Figura 5). Segundo Adami et al. (2012), 69,7% das áreas de cana-de-açúcar expandiram-se sobre áreas de pastagem e 25% sobre outras culturas agrícolas (*citrus* e culturas anuais) na última década em toda a região centro-sul do país.

4. Conclusão

Este trabalho apresentou a nova versão de um algoritmo para aplicações de redes bayesianas para análise de dados *rasters*: *enhanced Bayesian Network for Raster Data* (e-BayNeRD). Um estudo de caso foi utilizado para descrever as principais mudanças e testar a nova versão, no qual foram mapeadas, a partir do Zoneamento Agroecológico da cana-de-açúcar, áreas aptas à expansão desta cultura no estado de São Paulo.

Para que uma área seja apta ao plantio sustentável de determinada cultura, diversas variáveis devem ser analisadas. Neste trabalho foram consideradas a declividade, a aptidão do solo, a aptidão edafoclimática e o zoneamento agroambiental. Baseado nas observações sobre essas variáveis o e-BayNeRD foi capaz de espacializar as áreas aptas ao plantio, atestando o potencial da ferramenta. Contudo, várias outras variáveis (e.g., econômicas, sociais, etc.) devem ser incluídas e uma modelagem mais refinada pode melhorar os resultados.

5. Agradecimentos

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

6. Referências Bibliográficas

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. 12, p. 574–585, 2 abr.

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.

Bivand, R.; Keitt, T.; Rowlingson, B. **rgdal: Bindings for the Geospatial Data Abstraction Library**. . [S.l.: s.n.]. Disponível em: <<http://cran.r-project.org/package=rgdal>>.2014

Bottcher, S. G.; Dethlefsen., C. **deal: Learning Bayesian Networks with Mixed Variables**. . [S.l: s.n.]. Disponível em: <<http://cran.r-project.org/package=deal>>.2013

Hijmans, R. J.; Etten, J. Van; Mattiuzzi, M.; Sumner, M.; Greenberg, J. A.; Lamigueiro, O. P.; Bevan, A.; Racine, E. B.; Shortridge, A. **raster: Geographic data analysis and modeling. R package version 2.2-12**. [S.l: s.n.]. Disponível em: <<http://cran.r-project.org/web/packages/raster/>>. 2014

IBGE, Instituto Brasileiro de Geografia e Estatística. **Indicadores IBGE - Estatística da Produção Agrícola**. . [S.l: s.n.], 2014.

Manzatto, C. V.; Assad, E. D.; Bacca, J. F. M.; Zaroni, M. J.; Pereira, S. E. M. **Zoneamento Agroecológico da Cana-de Açúcar - Expandir a produção , preservar a vida , garantir o futuro**. (C. V. Manzatto, E. D. Assad, J. F. M. Bacca, M. J. Zaroni & S. E. M. Pereira, Org.). Rio de Janeiro: EMBRAPA SOLOS, 2009. Disponível em: <http://www.cnps.embrapa.br/zoneamento_cana_de_acucar/ZonCana.pdf>.

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

Pearl, J.; Russell, S. **Bayesian Networks**. In: Department of Statistics Papers, 2011. Califórnia: Department of Statistics, UCLA. Disponível em: <<https://escholarship.org/uc/item/53n4f34m>>. Acesso em: 08 ago. 2014.

Qin, D.; Jianwen, M.; Yun, O. Y. Remote sensing data change detection based on the CI test of Bayesian networks. **Computers & Geosciences**, v. 32, n. 2, p. 195–202, mar. 2006.

R Core Team. **R: A Language and Environment for Statistical Computing**. . Vienna, Austria: R Foundation for Statistical Computing. Disponível em: <<http://www.r-project.org/>>. , 2014

Rudorff, B. F. T.; De Aguiar, D. A.; Da 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, 9 abr. 2010.

Scutari, M. Learning Bayesian Networks with the bnlearn R Package. **Journal of Statistical Software**, v. 35, n. 3, p. 22, 26 ago. 2009.

Silva, A. C. O.; Mello, M. P.; Fonseca, L. M. G. Enhancements to the Bayesian Network for Raster Data (BayNeRD). 2014, Campos do Jordão: Proceedings of the XV Symposium on GeoInformatics - GEOINFO, 2014. p. No prelo.

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