Anais XIV Simpósio Brasileiro de Sensoriamento Remoto, Natal, Brasil, 25-30 abril 2009, INPE, p. 5531-5538.

Redes neurais para a seleção de variáveis ambientais no processo de modelagem de distribuição de espécies na região Norte do Brasil

Arimatéa de Carvalho Ximenes[1]
Silvana Amaral[1]
Gustavo Felipe Balué Arcoverde[2]
Antônio Miguel Vieira Monteiro[1]

Instituto Nacional de Pesquisas Espaciais - INPE
Caixa Postal 515 - 12245-970 - São José dos Campos - SP, Brasil

[1]Divisão de Processamento Digital de Imagens - DPI
{arimatea, silvana, miguel}@dpi.inpe.br

[2]Divisão de Sensoriamento Remoto - DSR
gustavo@dsr.inpe.br

Abstract. Predictive species distributions models works over species occurrence records and environmental data to produce a model of the species requirements and a map of its potential geographical distribution. The large number of environmental data available for species distribution modeling requires a rigorous pre-processing analysis: models generated with a few non-correlated variables are more robust and easily interpreted. This work proposes using neural network approach (Self Organizing Maps) to analyze the spatial dependence of climatic and bioclimatic variables commonly used in species distribution modeling. The spatial heterogeneity of 58 variables were prepared based on cells space in approach, provided by TerraView, and analyzed in the neural map Considering the Brazilian North Region Brazilian as area of interest, it was identified some groups by its spatial correlation, and it was observed that latitude, longitude and altitude could substitute most of climatic variables, offering better spatial resolution. This methodological approach can be applied over different regions and data sets offering a spatial criteria for variable selection and optimizing the modeling process.

Key-words: spatial correlation, Self Organizing Maps, neural network, spatial distribution modeling, and climate data.

1. Introdução

A falta de informação sobre a biodiversidade expressa pela distribuição da ocorrência das espécies pode ser contornado com a geração de modelos de distribuição de espécies. Estes modelos baseiam-se no conceito de nicho ecológico, e através de métodos quantitativos, buscam relacionar a evidência da ocorrência da espécie às suas dependências do meio ambiente, gerando uma superfície associada à probabilidade de ocorrência das espécies (Engler et al., 2004).

Os modelos baseados em envelopes bioclimáticos predizem locais com condições climáticas favoráveis a uma determinada espécie para indicar regiões de ocorrência em potencial. Dados climáticos interpolados espacialmente referentes às chamadas, superfícies climáticas (Hijmans et al., 2005) são amplamente usados como variável de entrada nos modelos de distribuição de espécies (Buermann et al., 2008; Loiselle et al., 2008).

Para algumas regiões como o Norte do Brasil, as estações meteorológicas que compõem a base de dados do WordClim não estão regularmente distribuídas, localizam-se em sua maioria próximas a estradas, rios e cidades. Além disto, estudos que utilizam os

dados climáticos do WorldClim raramente discutem a redundância de informação entre as variáveis climáticas.

Dentre as possibilidades para se analisar dados multivariados encontra-se o Mapa Auto-Organizável (SOM - Self-Organizing Map), um tipo de rede neural que permite visualizar a variação de seus atributos em um mapa neural. O SOM, também conhecido como Mapa de Kohonen, aproxima uma função de densidade de probabilidade aos dados de entrada. Tem sido usado com freqüência para clusterização, visualização multivariada, e redução da dimensionalidade (Kohonen, 2001; Park et al., 2003). Depois de seu surgimento em 1982, o Mapa Auto-Organizável (Kohonen, 1982) tem sido aplicado em problemas de diversas áreas do conhecimento como ordenação de dados ecológicos (Foody, 1999; Park et al., 2007), classificação de imagens de satélites (Konishi et al., 2007), análise de dados geoespaciais multivariados (Silva, 2004), mapeamento de ecorregiões (Ximenes, 2008) entre outros.

Este trabalho tem como objetivo analisar a dependência espacial entre as variáveis climáticas e bioclimáticas da base de dados do WorldClim. Estes resultados deverão contribuir como um critério de pré-seleção das variáveis ambientais para a geração de modelos de distribuição de espécies da região norte do Brasil. A pré-seleção das variáveis climáticas reduz a dimensionalidade dos dados, o custo computacional e facilita a análise das atividades de modelagem.

2. Metodologia

2.1. Área de estudo

A área de estudo está inserida no domínio do bioma Amazônia e encontra-se no limite político da região norte do Brasil (**Figura 1**).

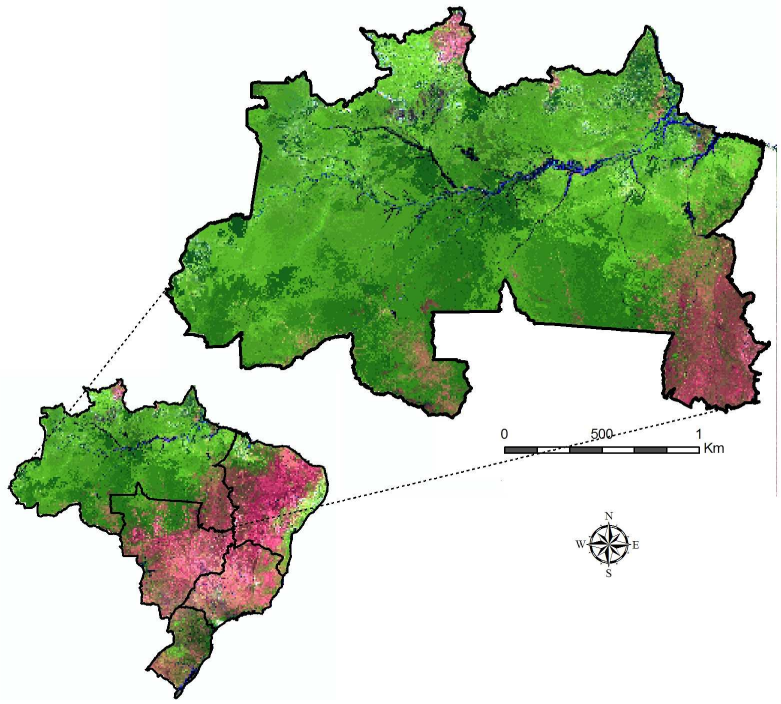

Figura 1. Área de estudo - região Norte do Brasil.

2.2. Variáveis analisadas

Os dados climáticos utilizados foram acessados na base de dados do WorldClim, versão 1.4 (Hijmans et al., 2005). A resolução é de aproximadamente 18,5 km no equador

(10 arc-minuto). O banco de dados mundial do WorldClim consiste em 47.554 estações de medição para a precipitação e 14.835 para a temperatura mínima e máxima. As superfícies climáticas foram calculadas por meio da média de um período de aproximadamente 50 anos (1950/2000), período que pode mudar conforme a disponibilidade de dados das estações meteorológicas para uma dada região.

Para este trabalho todas as variáveis climáticas do WorldClim com 10 arc-minuto de resolução foram analisadas, sendo elas: bioclimáticas, precipitação acumulada, temperatura máxima e mínima.

As variáveis bioclimáticas são derivadas das variáveis de temperatura e precipitação, sendo 11 variáveis derivadas da temperatura (BIO1 a BIO11) e 8 variáveis derivadas da pluviosidade (BIO12 a BIO19).

Tabela 1. Variáveis bioclimáticas fornecidas pelo WorldClim derivadas das variáveis de temperatura e pluviosidade.

Variável	Descrição
BIO1	Temperatura média anual
BIO2	Intervalo médio diurno (Média mensal (max temp - min temp))
BIO3	Isotermalidade
BIO4	Sazonalidade de Temperatura (desvio padrão *100)
BIO5	Temperatura máxima do mês mais quente
BIO6	Temperatura mínima do mês mais frio
BIO7	Intervalo da temperatura anual
BIO8	Média do quarto de ano mais úmido
BIO9	Média do quarto de ano mais seco
BIO10	Média do quarto de ano mais quente
BIO11	Média do quarto de ano mais frio
BIO12	Precipitação anual
BIO13	Precipitação do mês mais frio
BIO14	Precipitação do mês mais seco
BIO15	Sazonalidade de precipitação (Coeficiente de variação)
BIO16	Precipitação do quadrimestre mais úmido
BIO17	Precipitação do quadrimestre mais seco
BIO18	Precipitação do quadrimestre mais quente
BIO19	Precipitação do quadrimestre mais frio

Variáveis independentes utilizadas pelo WorldClim

Os dados climáticos do projeto WorldClim foram compilados de diferentes fontes e interpolados espacialmente, utilizando as variáveis independentes de altitude, latitude e longitude para gerar grades regulares chamadas de “superfícies climáticas” (Hijmans et al., 2005).

As coordenadas de latitude e longitude foram obtidas a partir dos centróides das células e armazenadas na tabela como atributos. A altitude foi obtida a partir do mosaico de imagens SRTM reamostradas para 1 km e foi extraída a partir do *plugin* de preenchimento de células por meio da operação de média.

As variáveis de latitude, longitude e altitude foram utilizadas neste trabalho para verificar a influencia que elas exercem sobre as variáveis climáticas.

2.3. Métodos

Para agregar informações com diferentes resoluções numa mesma base espacial utilizou-se a abordagem de espaços celulares. Os atributos de cada célula são armazenados em tabela onde as linhas representam as células ou vetores de entrada, e as colunas correspondem às variáveis, ou atributos.

Para recobrir a área de estudo foram criadas 10.181 células regulares de 20 km. O valor médio de cada variável foi obtido para cada célula, utilizando-se a ferramenta de

preenchimento de células do sistema TerraView versão 3.2. A tabela resultante compõe o conjunto de dados de entrada para o algoritmo de redes neurais do tipo SOM.

2.4. Mapa Auto-Organizável

O Somtoolbox, desenvolvido em Matlab pelo Centro de Pesquisas em Redes Neurais da Universidade Tecnológica de Helsinki na Finlândia, possibilita a execução de algoritmos de treinamento da rede neural do tipo Mapa Auto-organizável (Vesanto et al., 2000). O Somtoolbox dispõe de rotinas específicas para visualização dos dados de entrada e das métricas de qualidade do treinamento, e tem sido utilizado amplamente em diversas aplicações científicas (Park et al., 2006).

Os procedimentos para treinamento do SOM compreendem: normalizar os dados de entrada, estruturar o mapa neural, definir o tipo de inicialização dos pesos sinápticos, apresentar as heurísticas utilizadas para iniciar o algoritmo de treinamento por lote e escolha da forma de rotulação dos neurônios. Antes de inicializar o treinamento foi preciso determinar o número de neurônios e as dimensões do mapa neural. O número de neurônios foi determinado pelo método exploratório com base na melhora visual do mapa neural. As dimensões do mapa neural são calculadas a partir da razão entre os dois maiores autovalores da matriz de covariância dos dados de entrada, e ajustadas de modo que seu produto seja próximo do número desejado de neurônios do mapa (Vesanto et al., 2000).

Depois de definir as dimensões do mapa neural, os vetores de peso sinápticos são inicializados de forma aleatória com a semente de números aleatórios do Matlab fixada em 10. Esta heurística permite que os resultados tornem-se parecidos a cada simulação, contando com a mesma configuração da rede neural.

O arranjo hexagonal foi utilizado para a vizinhança dos neurônios, pois de acordo com Kohonen (2001), é mais adequado para visualizar o mapa neural. A função de vizinhança escolhida foi a gaussiana que permite que o algoritmo SOM convirja mais rapidamente que uma vizinhança topológica retangular (Lo et al., 1991). Para o treinamento da rede neural foi adotado o algoritmo por lote, que apresenta maior rapidez no processamento quando comparado com o algoritmo seqüencial (Vesanto et al., 2000). A independência do resultado quanto à ordem de apresentação dos vetores de entrada e a ausência da taxa de aprendizagem (Silva, 2004; Vesanto et al., 2000) foram fatores decisivos para a escolha deste algoritmo.

Para analisar visualmente a contribuição de cada variável de entrada e identificar o sentido em que os atributos variam no mapa neural foram utilizados os planos de componentes. A relação entre cada variável pode ser identificada através da análise visual dos padrões formados pelos valores altos e baixos dos planos de componentes (Vesanto, 2002). Variáveis que apresentam distribuição semelhante nos planos de componentes podem fornecer o indício de dependência espacial, ou seja, o grau com que a variação dos atributos segue a variação na localização espacial (Silva, 2004).

3. Resultados e discussão

O mapa neural foi gerado com 150 neurônios com dimensões de 15 x 10. O raio de vizinhança inicial foi de 4 e o raio final, o próprio neurônio vencedor. Na fase de ordenamento o número de épocas foi 1 e na fase de convergência 2. Os erros foram considerados baixos com 0,39 de erro de quantização e 0,05 de erro topológico. Esses parâmetros geraram os melhores resultados, segundo a forma de visualização dos planos de componentes e o baixo valor dos erros de treinamento.

As correlações entre as variáveis climáticas foram identificadas por meio dos planos de componentes de cada variável. As análises foram realizadas separadamente para cada conjunto de variáveis, como precipitação, bioclimáticas, temperatura máxima e mínima.

Precipitação

Segundo a análise visual dos planos de componentes (**Figura 3**), os meses correlacionados de precipitação podem ser separados em três grupos. O primeiro grupo é composto pelos meses de janeiro a abril, o segundo maio a agosto e o terceiro grupo de setembro a dezembro. Os meses de janeiro e dezembro também são correlacionados. Os meses de abril e setembro apresentaram menor correlação e o mês de janeiro pode ser usado para caracterizar a estação chuvosa. Ao comparar os planos de componentes da precipitação com as variáveis independentes utilizadas pelo WorldClim percebe-se a influência da longitude e latitude.

Temperatura máxima

As variáveis de temperatura máxima podem ser separadas em quatro conjuntos de meses correlacionados. Os meses de janeiro a março, abril a junho, julho a setembro e outubro a dezembro. Os planos de componentes indicam que a correlação é maior entre meses vizinhos de temperatura máxima e conforme a distância entre os meses aumenta a correlação diminui (**Figura 3**).

Temperatura mínima

Segundo a análise dos planos de componentes, todos os meses de temperatura mínima são correlacionados. Os meses de junho a agosto apresentam maior homogeneidade em comparação com os outros meses (**Figura 3**).

As estações meteorológicas que medem a temperatura mínima encontram-se em menor número quando comparado com as variáveis de precipitação, temperatura média e máxima (Hijmans et al., 2005). Possivelmente, uma rede mais densa de estações meteorológicas, cobrindo mais efetivamente a região norte do pais, conduziria a diferentes resultados de correlação entre os meses para a variável de temperatura mínima.

As variáveis de temperatura máxima e mínima são influenciadas principalmente pela altitude que foi utilizada como variável independente para gerar a superfície climática.

Bioclimáticas

A variável de temperatura mínima do mês mais frio (BIO 6) possui a mesma distribuição dos atributos no plano de componente que as variáveis de temperatura mínima dos meses de junho a agosto, indicando que estes meses são os mais frios do ano na região Norte (**Figura 3**). Dentre as variáveis bioclimáticas, o intervalo médio de temperatura diurna (BIO 2) possui a maior homogeneidade e com valores altos para quase toda a região Norte (**Figura 3**).

A sazonalidade de temperatura (BIO 4) não é correlacionada com nenhuma bioclimática, porém mostra-se influenciada pela variável de altitude (**Figura 3**).

As variáveis correlacionadas de temperatura média do quarto de ano mais úmido (BIO 8), mais seco (BIO 9), mais quente (BIO 10) e mais frio (BIO 11) e a temperatura média anual (BIO 1) possuem a mesma distribuição no mapa neural.

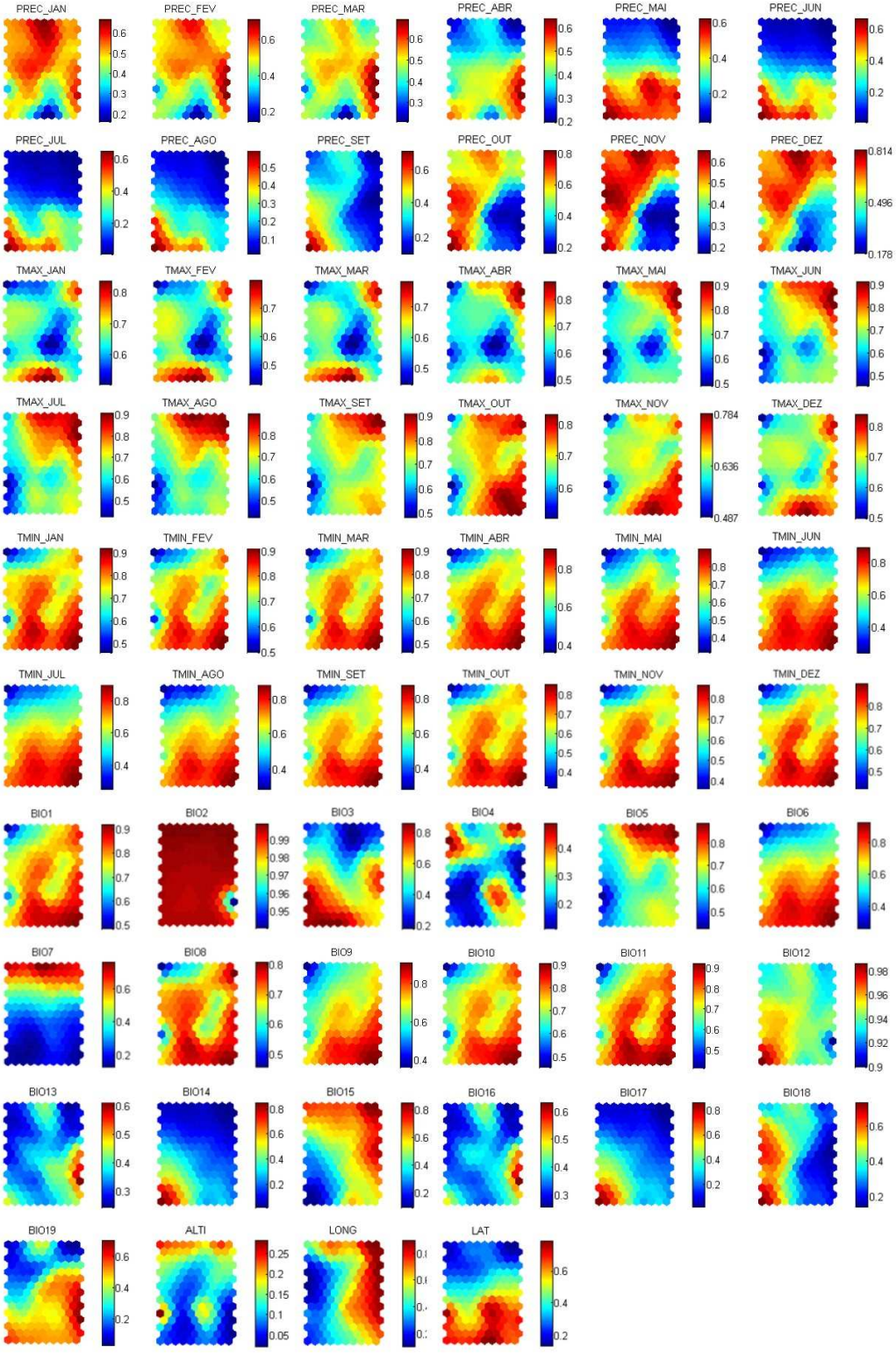

Figura 3. Planos de componentes das variáveis climáticas e das variáveis independentes utilizadas para gerar a superfície climática. PREC - precipitação, TMAX - temperatura máxima, TMIN - temperatura mínima, BIO - bioclimática, ALT - altitude, LONG - longitude, LAT - latitude. Ao lado dos planos de componentes estão os valores normalizados referentes a cada variável. Os níveis de cinza indicam a posição dos valores no mapa neural. Valores altos encontram-se próximos à cor vermelha escura e valores baixos próximos ao azul escuro.

As variáveis de precipitação do mês mais seco (BIO 14) e a média de precipitação do quarto de ano mais seco (BIO 17) são correlacionadas, sendo a sazonalidade de precipitação (BIO 15) negativamente correlacionada com essas variáveis. A isotermalidade (BIO 3) e a sazonalidade de precipitação (BIO 15) também apresentam correlação negativa. A precipitação do quarto de ano mais úmido (BIO 16) e a precipitação do mês mais frio (BIO 13) possuem a distribuição dos atributos no mapa neural idênticos (**Figura 3**).

4. Conclusão

Os resultados sugerem que as variáveis independentes utilizadas pelo WorldClim para gerar as superfícies climáticas, altitude, latitude e longitude, podem substituir a maioria das variáveis climáticas proporcionando resolução espacial mais detalhada.

As ferramentas de visualização do SOM permitiram avaliar e identificar a dependência espacial entre variáveis a partir da análise dos planos de componentes. Com o método de visualização das variáveis a partir dos planos de componentes pode-se estudar não somente a correlação como também avaliar a dependência espacial entre as variáveis. Como exemplo, a variável BIO 2 (Variação média de temperatura diurna) não é correlacionada com as outras variáveis bioclimáticas, porém sua contribuição é pequena devido à homogeneidade para a região de estudo. A falta de estações meteorológicas bem distribuídas na região Norte do país não permite uma avaliação mais detalhada quanto à redundância de informação e homogeneidade climática.

Este trabalho ao avaliar a correlação entre dados climáticos e ambientais, contribui para a pré-seleção de variáveis climáticas a serem utilizadas no processo de modelagem de distribuição de espécies para a região Norte do Brasil. Desta forma, obtém-se a redução da dimensionalidade das variáveis de entrada, o que por sua vez proporciona coerência na explanação dos resultados dos modelos. Esta metodologia pode avaliar a correlação entre diferentes conjuntos de variáveis ambientais, ou ser aplicada para outras regiões, bem como para todo o Brasil, de acordo com a área de interesse para modelagem.

Agradecimentos

Os autores agradecem à rede GEOMA - Rede Temática de Pesquisa em Modelagem da Amazônia pelo suporte para realização do trabalho.

Referências bibliográficas

Aguiar, A. P. D.; Andrade, P. R.; Ferrari, P. G. **Preenchimento de células**. Relatório técnico, 2008. p. 11. Disponível em: http://www.dpi.inpe.br/~anapaula/plugin_celulas/help.pdf

Buermann, W.; Saatchi, S.; Smith, T. B.; Zutta, B. R.; Chaves, J. A.; Milá, B.; Graham, C. H. Predicting species distributions across the Amazonian and Andean regions using remote sensing data. **Journal of Biogeography**, v. 35, p. 1160-1176, 2008.

Engler, R.; Guisan, A.; Rechsteiner, L. An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data. **Journal of Applied Ecology**, v. 41, n. 2, p. 263-274, 2004.

Hijmans, R. J.; Cameron, S. E.; Parra, J. L.; Jones, P. G.; Jarvis, A. Very high resolution interpolated climate surfaces for global land areas. **International Journal of Climatology**, v. 25, p. 1965-1978, 2005.

Hutchinson, G. E. Concluding remarks. **Cold Spring Harbour Symposium on Quantitative Biology**, v. 22, p. 415-427, 1957.

Kohonen, T. Self-organized formation of topologically correct feature maps. **Biological Cybernetics**, v. 43, p. 59–69, 1982.

Kohonen, T. The Self-Organizing Map. **Proceedings of the IEEE**, v. 78, n. 9, p. 1464-1480, 1990.

Kohonen, T. **Self-Organizing Maps**, Springer Series in Information Sciences. Springer, Berlin, Heidelberg, v. 30, ed. 3, 2001.

Konishi, T.; Omatu, S.; Suga, Y. Extraction of rice-planted area using a self-organizing feature map. **Artif. Life Robotics**, v. 11, p. 215-218, 2007
Park, Y. S.; Song, M. Y.; Park, Y. C.; Oh, K. H. Cho, E.; Chon, T. S. Community patterns of benthic macroinvertebrates collected on the national scale in Korea. **Ecological Modelling**, v. 203, p. 26-33, 2007.

Lo, Z.; Fujita, M.; Bavarian, B. Analysis of neighborhood interaction in Kohonen neural networks. **Proceeding of the 6th International Parallel Processing Symposium**, p. 247–249, 1991.

Loiselle, B. A.; Jørgensen, P. M.; Consiglio, T.; Jiménez, I.;. Blake, J. G.; Lohmann, L. G.; Montiel, O. M. Predicting species distributions from herbarium collections: does climate bias in collection sampling influence model outcomes? **Journal of Biogeography**, v. 35, p. 105-116, 2008.

Park, Y. S.; Lek, S.; Scardi, M.; Verdonschot, P. F. M.; Jørgensen, S. E. Patterning exergy of benthic macroinvertebrate communities using self-organizing maps. **Ecological Modelling**, v. 195, p. 106–114, 2006.

Park, Y. S.; Song, M. Y.; Park, Y. C.; Oh, K. H.; Cho, E.; Chon, T. S. Community patterns of benthic macroinvertebrates collected on the national scale in Korea. **Ecological Modelling**, v. 203, p. 26-33, 2007.

Silva, M. A. S. **Mapas auto-organizáveis na análise exploratória de dados geoespaciais multivariados**. 2004-03-08. 117 p. (INPE-12434-TDI/996). Dissertação (Mestrado em Computação Aplicada) - Instituto Nacional de Pesquisas Espaciais, São José dos Campos. 2004. Disponível em: <http://urlib.net/sid.inpe.br/jeferson/2004/03.25.16.40>. Acesso em: 28 out. 2008.

Vesanto, J.; Himberg, J.; Alhoniemi, E.; Parhankangas, J. **SOM Toolbox for Matlab 5**. Technical Report A57, Neural Networks Research Centre, Helsinki University of Technology, Helsinki, Finland, 2000.

Vesanto, J. **Data exploration process based on the Self-Organizing Map**. Dissertation for the degree of Doctor of Technology. Helsinki University of Technology. Espoo, Finland, 2002.

The Mathworks Inc., 2001. MATLAB Version 7.1. The Mathworks, Inc., Massachusetts.

Ximenes, A. C. **Mapas auto-organizáveis para a identificação de ecorregiões do interflúvio Madeira-Purus**: uma abordagem da biogeografia ecológica. 2008-06-05. 155 p. (INPE-15332-TDI/1372). Dissertação (Mestrado em Sensoriamento Remoto) - Instituto Nacional de Pesquisas Espaciais, São José dos Campos. 2008. Disponível em: < http://urlib.net/sid.inpe.br/mtc-m18@80/2008/08.18.14.02 >. Acesso em: 28 out. 2008.