Metodologia para Definição das Atividades do Processo de Modelagem de Distribuição de Espécies

Jorge L. D. Pinaya[1], Pedro L. P. Corrêa[1]

[1]Departamento de Engenharia de Computação e Sistemas Digitais da Escola Politécnica da Universidade de São Paulo (USP) – São Paulo, SP – Brasil

{jorge.pinaya,pedro.correa}@usp.br

Abstract. The area of research called biodiversity informatics, or bioinformatics, has to face the challenge of meeting the demand for technologies to support the conservation of biodiversity. The models of distribution of species have a fundamental implication for understanding the biodiversity and conservation decision making. The objective of this research is to present the process of species distribution modeling with emphasis on the activities of pre-analysis and activities selection of the predictor variables, such as to favor its repeatability and reproducibility by other researchers. The process of modeling species distribution proposed is validated on a case study of modeling distribution of pollinator species Centris hyptidis and oilseed Angelonia campestris and Angelonia cornigera. In this case study we can observe one of the main contributions of this work: the application of statistical techniques for data exploration in the pre-analysis of species distribution modeling process, with improved capacity for evaluation and selection of points of occurrence essential to the performance of the predictive model.

Resumo. *A área de pesquisa científica, denominada computação e biodiversidade, têm por desafio suprir a demanda por tecnologias de apoio à conservação da biodiversidade. Os modelos de distribuição de espécies têm uma importante contribuição para o entendimento da biodiversidade e no apoio para a tomada de decisão em conservação dos recursos de biodiversidade. O objetivo desta pesquisa é apresentar o processo de modelagem de distribuição de espécie com destaque para as atividades de pré-análise e atividades de seleção das variáveis preditoras. O processo de modelagem de distribuição de espécies proposto é avaliado por meio de estudo de caso de Modelagem de Distribuição de Espécies do polinizador Centris hyptidis e das plantas oleaginosas Angelonia campestris e Angelonia cornigera. Neste estudo de caso pode-se verificar que a aplicação de técnicas estatísticas exploratórias de dados na etapa de pré-análise do processo de modelagem distribuição de espécies permite a avaliação da qualidade dos pontos de ocorrência, essenciais para o desempenho preditivo do modelo final.*

1. Introdução

As técnicas de modelagem computacional para distribuição de espécies são críticas para a tarefa de identificar áreas com alto risco de perda de biodivesidade. Estas ferramentas podem auxiliar na conservação da biodiversidade, no planejamento do uso de regiões não habitadas, nas estimativas de risco de invasão de espécies, na proposta de reintrodução de espécies e mesmo no planejamento de proteção de espécies ameaçadas (Corrêa et al., 2011).

Segundo Peterson et al. (2006) e Elith et al. (2006), na modelagem de distribuição de espécies, além dos algoritmos de modelagem e as variáveis preditoras, também estão envolvidos neste processo diversos tipos de dados e formatos de dados, e técnicas de análise de dados, como por exemplo, o formato Darwin Core do Taxonomic Database Working Group (TDWG) (www.tdwg.org), e o protocolo TDWG Acess Protocol for Information Retrieval (TAPIR) (www.tdwg.org/activities/tapir/).

Assim, considera-se que o Processo de Modelagem de Distribuição de Espécie utilizado na criação de um Modelo é complexo. O Processo de Modelagem de Distribuição de Espécie é composto de várias etapas, como pré-análise, análise e pós-análise, e as atividades específicas de cada etapa ,como data clearing e o teste do desempenho preditivo do modelo (Corrêa et al., 2011; Elith; Leathwick, 2009).

Nas sessões a seguir são apresentados os conceitos básicos de Processo de modelagem de distribuição de espécies e a sua aplicação em um estudo de caso de modelagem de distribuição de espécies.

2. Processo de Modelagem de Distribuição de Espécies

A Figura 1 apresenta uma metodologia que combina os diversos Fatores que influênciam a Modelagem de Distribuição de Espécies, usa a notação Profile UML com a Extensão de Negócios Eriksson-Penker (PRESSMAN, 2010; ERIKSSON, PENKER, 1999).

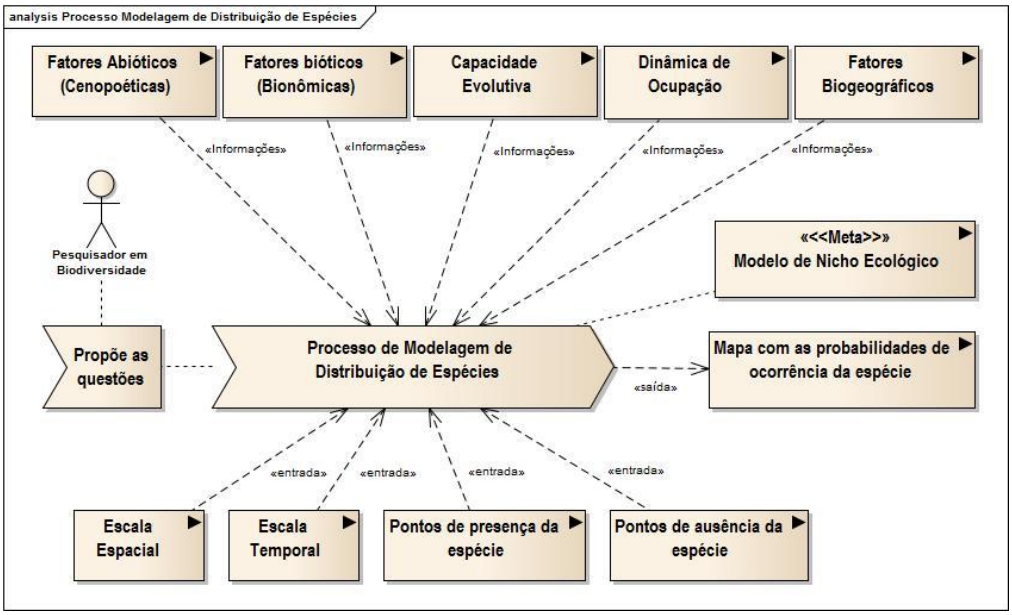

Figure 1. Metodologia para Modelagem de Distribuição de Espécies

O Pesquisador em Biodiversidade define as questões ecológicas que a modelagem de distribuição de espécies deve ajudar a responder, conforme Santana et al. (2008) e Giannini (2011). Conforme Soberón e Peterson (2005) e Hortal et al. (2010), o pesquisador em biodiversidade usa informações sobre:

- Fatores Abióticos ou variáveis Cenopoéticas, que são as características ambientais, por exemplo, temperatura e precipitação média anual;
- Fatores Bióticos ou variaveis Bionômicas, que correspondem às interações entre a espécie e outros que ocupam o mesmo ambiente e a dinâmica de consumo de recursos, por exemplo, predadores e competidores;
- Dinâmica de Ocupação e Dispersão, que são os movimentos espaciais dos indivíduos e populações, por exemplo, com base na configuração da paisagem;
- Fatores Biogeográficos, que determinam o tamanho e forma dos espaços na distribuição de espécies, por exemplo, o Gradiente Latitudinal.
- Capacidade Evolutiva, onde as espécies habitam os ambientes mais produtivos, por exemplo, restrições evolutivas limitam a ocupação de regiões mais frias.

O Pesquisador em biodiversidade deve ainda selecionar a escala espacial e temporal desejada para a pesquisa ecológica. Estas escalas determinam a importância relativa dos fatores e quais aspectos moldariam a distribuição nessa escala, conforme Hortal et al. (2010) e Peterson et al. (2011).

Considera-se neste Framework teórico, que o Processo de Modelagem de Distribuição de Espécies pode então ser definido em cinco Fases: 1) Construção da hipótese científica; 2) Pré-Análise dos dados; 3) Modelagem; 4) Predição; 5) Validação da hipótese científica.

A seqüência de Fases do Processo de Modelagem de Distribuição de Espécies é representada por um Diagrama de Atividade UML (Pressman, 2010; Booch, Rumbaugh, Jacobson, 2000), apresentado na Figura 2.

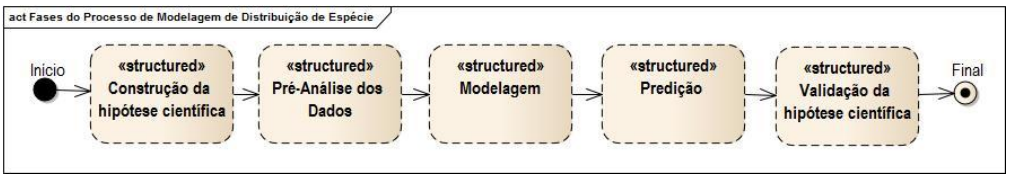

Figura 2: Fases do Processo de Modelagem de Distribuição de Espécies

As Etapas do Processo de Modelagem de Distribuição de Espécies foram decompostas em atividades, com a finalidade de ressaltar os relacionamentos entre as diversas atividades, apresentadas em um Diagrama de Atividades UML na Figura 3.

3. Estudo de Caso do Processo de Modelagem de Distribuição de Espécies

O Processo de Modelagem de Distribuição de Espécie proposto é aplicado na modelagem de distribuição geográfica de espécie da Centris hyptidis, na região nordeste brasileira.

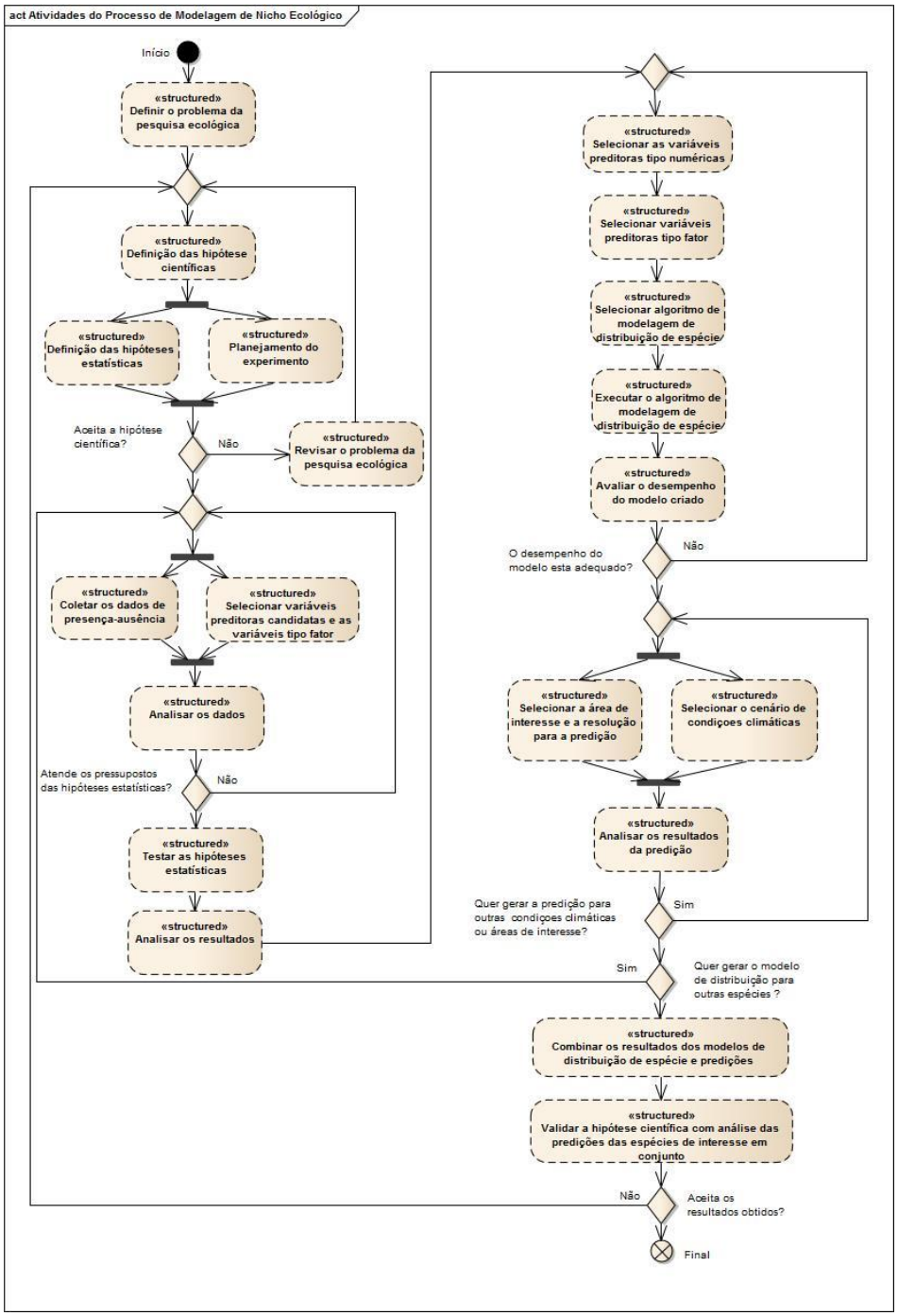

Figura 3: Diagrama de atividades UML, com os relacionamentos entre as atividades do Processo de Modelagem de Distribuição de Espécies.

3.1 Construção da Hipótese cientifica inicial

Nesta etapa do processo de distribuição de espécie proposto, foram coletados os pontos de ocorrência, na área de caatinga na região Nordeste do Brasil, para a modelagem de distribuição de espécies de plantas Angelonia campestris e Angelonia cornigera e o polinizador Centris hypitidis.

As interações entre as abelhas e as flores em um sistema de restinga, na Bahia, são apresentadas na pesquisa de Gimenes, Oliveira e Almeida (2002). Ainda neste sentido, a pesquisa de Machado, Vogel e Lopes (2002) apresenta a polinização da Angelonia cornigera por abelhas coletoras de óleo no nordeste do Brasil, e a pesquisa de Aguiar et al. (2003) apresenta as plantas visitadas pela abelha Centris.

3.2 Pré-análise dos dados

Nesta Etapa deve ser observada que a representatividade da amostra não é o mesmo que amostra aleatória (Hurlbert, 1984). Uma amostra aleatória é aquela em que todo membro de toda uma população tem chance igual e independente de ter sido amostrado.

Segundo Underwood (1998) e Hulbert (1984), muitos trabalhos ecológicos têm demonstrado a existência de gradientes e o efeito de patchiness, distribuição horizontal heterogênea, na distribuição de animais e plantas. O problema da não representatividade das amostras é o grau em que a amostra esta enviesada, isto é, o grau que ela superestima ou subestima o parâmetro que está sendo investigado.

No planejamento de um novo experimento de modelagem, os problemas de representatividade das amostras podem ser reduzidos com a aplicação das técnicas de design de experimento para segmentar uma área heterogênea de estudo em unidade amostrais (Hulrbert, 1984). A amostragem aleatória é apropriada quando não existe alternativa ou quando puder ser evitada uma amostra enviesada (Underwood, 1998).

3.2.1 Coleta de dados de ocorrências de espécies

Neste estudo de caso, foram levantados 41 pontos de ocorrência da espécie *Centris hypitidis*, 90 pontos de ocorrência da espécie *Angelonia campestris* e 384 pontos de ocorrência da espécie *Angelonia cornigera*.

A distribuição potencial da espécie foi obtida com base nas 19 variáveis bioclimáticas do projeto WorldClim (Hijmans et al., 2005), disponível online (http://www.worldclim.org). O recorte espacial para a região de interesse foi estabelecido de acordo com a coordenadas limites do Brasil, a longitude de 73º 59’ 16,5” W a 32º 23’ 16,5 W e a latitude de 5º 16’ 16,5” N a 33º 45’ 13,5 S, estimada em uma resolução espacial de 30 segundos de arco ,aproximadamente 1km no equador.

O recorte espacial das camadas bioclimáticas foi obtido no site AMBDATA (www.dpi.inpe.br/ambdata/index.php) - Variáveis ambientais para modelagem de distribuição de espécies do Grupo de Modelagem para Estudos de Biodiversidade da Divisão de Processamento de Imagens do INPE. As informações ambientais utilizadas neste experimento são grades projetadas no sistema de coordenadas geodésicas de projeção Latitude/Longitude, Datum WGS-84, com resolução espacial de 30 arc-segundos, ou aproximadamente 1 km.

3.2.1 Análise Espacial dos dados de ocorrências de espécies

A estimativa de média e variância de uma população é a preocupação de muitos estudos de Biologia e Ecologia. Entretando, outros parâmetros de frequência de distribuição são úteis sobre o processo ecológico, conforme Underwood (1998). Os Sistemas de Informações Geograficas (GIS) permitem realizar a análise espacial, as operações

permitem estabelecer zonas de influencia, criação de novos polígonos e geração de novos atributos a partir de dois ou mais mapas existentes no sistema (Vitti, 2006).

Além disso, a área de estudos pode ser sintetizada pelas medidas estatísticas de posição e de dispersão das variáveis preditoras e pontos de registro de ocorrência (Sokal; Rohlf, 1995). Os dados podem ser sintetizados com o cálculo das medidas de posição e dispersão. As medidas de posição foram calculadas para cada variável preditora em relação à área de estudo, por exemplo, as medidas de tendência central, como a Média Amostral, a Moda e a Mediana; as medidas de dispersão ou variabilidade, como o Desvio Padrão e a Variância.

As medidas de posição e dispersão podem também ser exibidas em gráficos, como exemplo, o Gráfico Box-plot. O gráfico oferece a medida da posição central dos dados através da mediana. O Gráfico Box-plot também dá uma ideia da dispersão, ou contrariamente, da concentração dos valores, através da distância interquartílica (75% (Q3)-25% (Q1)). No Gráfico Box-plot, o comprimento das caudas é dado pelas linhas contínuas que vão da "caixa" (retângulo) aos valores mais afastados que não sejam outliers. Os outliers, ou valores atípicos, são representados por pequenos círculos vazios. O gráfico oferece também a medida da posição central dos dados através da mediana. (SOKAL; ROHLF, 1995).

O gráfico Box-plot pode também ser utilizado para uma comparação preliminar das distribuições das espécies em estudo, por exemplo, podemos observar os gráficos Box-plot, das espécies Angelonia campestris, Angelonia cornigera e Centris hyptidis em relação à Precipitação Anual (Bio12), apresentado na Figura 4.

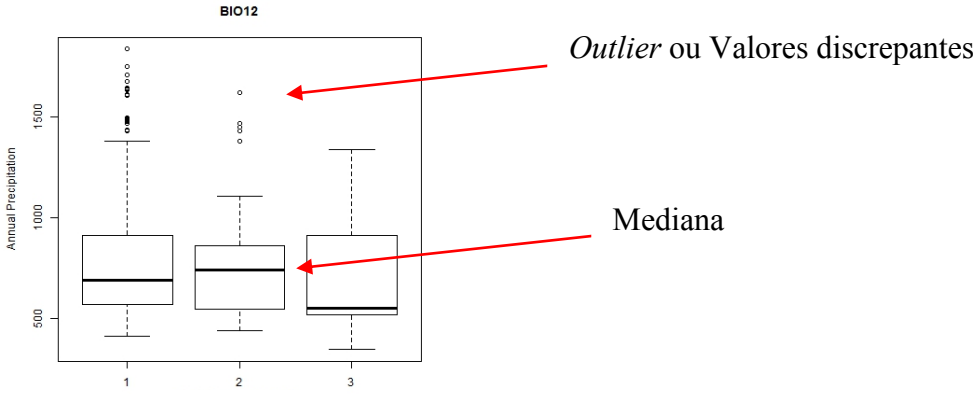

Figura 4: Gráfico Box-plot das espécies: 1- *Angelonia campestris*, 2- *Angelonia cornigera* e 3- *Centris hyptidis* em relação à Precipitação Média Anual (Bio12).

Segundo Phillips et al. (2009) e Hurlbert (1984), registros de ocorrência podem apresentar viés espacial quando a distribuição das espécies só foi amostrada num subconjunto não aleatório da área total onde a espécie ocorre, tais como apenas áreas que são facilmente acessíveis. Este viés pode resultar em má calibração do modelo.

Segundo Peterson et al. (2011) e Hijmans (2012), a autocorrelação espacial prejudica a separação em grupos independentes de pontos de registro de ocorrência para treinamento e teste do modelo de distribuição de espécie.

3.3 Modelagem

De acordo com Austin (2002) e Elith et al. (2006), os dados ambientais ou variáveis preditoras, utilizados no Processo de Modelagem de Distribuição de Espécies de cada região de interesse, devem ser selecionados pela sua relevância para a espécie a ser modelada.

3.3.1 Selecionar as variáveis preditoras

A redução da quantidade de variáveis preditoras, na prática, impede que padrões aleatórios espúrios sejam incorporados ao modelo construído. È também comum o interesse em se analisar o comportamento de duas ou mais variáveis quantitativas, para obter uma medida estatística que indique se existe ou não relação entre as duas variáveis; e se existe relação e qual a sua magnitude e sinal.

O coeficiente linear de Pearson é utilizado para quantificar a correlação entre duas variáveis quantitativas. Índica o quanto a nuvem de pontos aproxima-se de uma reta ascendente ou descendente (Zar, 2010; Legendre, Legendre, 1998).

3.3.2 Selecionar e executar o algoritmo de modelagem de distribuição de espécie

O algoritmo selecionado para este experimento foi o algoritmo de Entropia Máxima (Phillips et al., 2006). A versão do algoritmo de Entropia Máxima utilizada foi a 3.3.3k implementada no software MaxEnt (Phillps; Dudik; Schapire, 2004), disponível online (HTTP://www.cs.princenton.edu/~schapire/maxent). O algoritmo de Entropia Máxima do MaxEnt, disponível no Pacote "Dismo" (Hijmans, 2011), foi executado a partir de scripts desenvolvidos na linguagem R (R Development Core Team, 2009). A escolha foi devido ao seu melhor desempenho em estudos comparativos (Elith et al., 2011).

O desempenho preditivo do modelo é analisado pelo teste AUC (*Area Under the receiver operating characteristic Curve*), que mede a habilidade do modelo de discriminar entre a omissão de áreas com registros e a sobreprevisão de áreas não ocupadas (Elith et al., 2006). Os erros de comissão e omissão são avaliados com o Gráfico da Matriz de Confusão. O teste Jackknife foi utilizado para medir a importância das variáveis, o ganho quando a variável é aplicada isolada e a perda quando é omitida.

3.4 Predição

Segundo Peterson (2011), os modelos de nicho oferecem o potencial para prever áreas potenciais de distribuição de espécies em condições climáticas correntes ou em condições especificas, por exemplo, cenários futuros de mudanças climáticas, baseados nos *Special Report on Emissions Scenarios* (SRES) do *Intergovernmental Panel on Climate Change* (IPCC).

O resultado é um Mapa com a estimativa de Distribuição Potencial da Espécie para uma área de interesse, com valores em escala logística entre 0 e 1, onde 1 seria a máxima adequabilidade do habitat às condições de ocorrência da espécie. O Mapa de Distribuição Potencial da espécie Centris hyptidis obtido a partir do algoritmo de Entropia Máxima é apresentado na Figura 5.

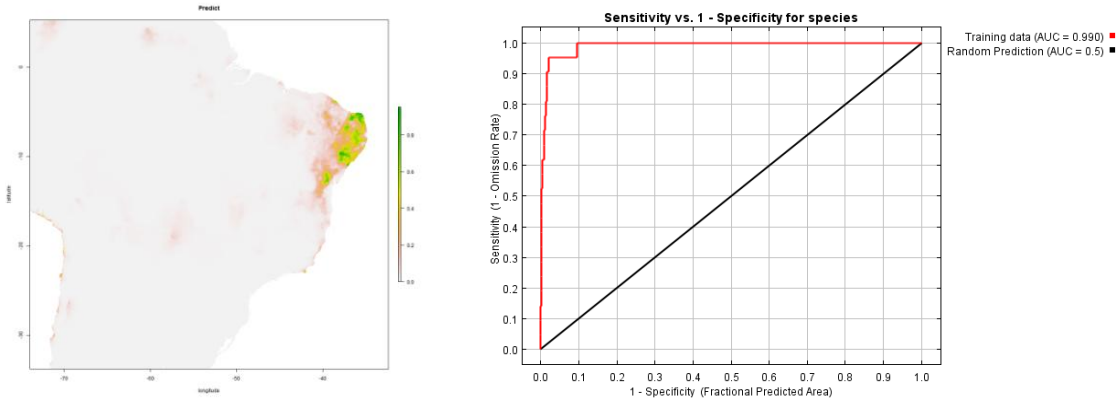

Figura 5: Mapa da Modelagem de Distribuição Potencial da espécie *Centris hyptidis* com o algoritmo de Entropia Máxima e o resultado do AUC (0,990).

3.5 Validar a hipótese científica

A hipótese cientifica pode ser analisada por meio da inspeção visual dos Mapas das predições, ou projeções, das Modelagens de Distribuição de Espécie, por Pesquisadores Especialistas nas Espécies, a fim de verificar se os modelos gerados são consistentes com as informações sobre a distribuição empiricamente determinada.

4. Conclusões

Com esta visão detalhada dos requisitos do Processo de Modelagem de Distribuição de Espécies surge à necessidade do desenvolvimento de ferramentas para *Biodiversity Informatics*, que possam compartilhar os dados da modelagem de distribuição de espécie, e permitir a flexibilidade necessária para que o Processo de Modelagem de Distribuição de Espécies possa atender às questões ecológicas complexas.

Entre as principais contribuições deste trabalho foram à apresentação de um Processo de Modelagem de Distribuição de Espécie revisado que permite buscar sistematicamente a melhoria da metodologia utilizada e a análise da Hipótese Científica dentro de um mesmo contexto integrado.

A identificação de técnicas estatísticas que podem ser utilizadas nas diversas Etapas do Processo de Modelagem de Distribuição de Espécies e o detalhamento do relacionamento entre as diversas atividades podem ser utilizados na definição de novos requisitos de software para as ferramentas de Modelagem de Distribuição de Espécies.

5. Agradecimentos

Agradeço a Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) GmbH, Instituto Chico Mendes de Conservação da Biodiversidade (ICMBio) do Ministério do Meio Ambiente (MMA) do Brasil, ao grupo de pesquisa do Laboratório de Automação Agrícola do Departamento de Engenharia de Computação e Sistemas Digitais da Escola Politécnica da Universidade de São Paulo, e do Laboratório de Abelhas (BeeLab) do Departamento de Ecologia do Instituto de Biociências da Universidade de São Paulo (IB-USP).

Referências

Aguiar, C.M.L., Zanella, F. C. V., Martins, C. F., Carvalho, C. A. L. (2003) Plantas visitadas por Centris spp. (Hymenoptera: Apidae) na Caatinga para obtenção de recursos florais. Neotropical Entomology v. 32. p. 247–259.

Austin, M. P. (2002) Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. Ecological Modelling. v. 157 p. 101-118.

Booch, G.; Rumbaugh, J.; Jacobson, I. (2000) "UML, Guia do Usuário: tradução"; Fábio Freitas da Silva, Rio de Janeiro, Ed.Campus.

Corrêa, P. L. P.; Carvalhaes, M. A.; Saraiva, A. M.; Rodrigues, F. A.; Rodrigues, E. S. C.; Rocha, R. L. A. (2011) "Computational techniques for biologic species distribution modeling", p. 308-325, DOI: 10.4018/978-1-61692-871-1.ch015.

Elith, J.; Graham, C. H.; Aanderson, R. P.; Dudik, M.; Ferrier, S.; Guisan, A.; Hijmans, R. J.; Huettmann, F.; Leathwick, J. R.; Lehmann, A.; LI, J.; Lohmann, L. G.; Loiselle, B. A.; Manion, G.; Moritz, C.; Nakamura, M.; Nakazawa, Y.; Overton, J.M.; Peterson, A. T.; Philips, S.J.; Richardson, K.; Scachietti-Pereira, R.; Schapire, R. E.; Soberon, J.; Willians, S.; Wisz, M.S.; Zimmermann, N.E. (2006) "Novel methods improve prediction of species' distributions from occurrence data", Ecography. v. 29 p. 129-151.

Elith, J.; Leathwick, J.R.. (2009) "Conservation prioritization using species distribution models. In Spatial Conservation Prioritization: Quantitative Methods and Computational Tools", ed. A Moilanen, KA Wilson, HP Possingham. Oxford: Oxford Univ. Press. pp 70–93.

Elith,J.; Philips, S. J.; Hastie, T; Dudik, M.; Chee, Y. E.; Yates, C. J. (2011) "A statistical explanation of maxent for ecologists." Diversity and Distributions. v. 17, p. 43-57.

Eriksson, H.; Penker, M. (1999) "Business Modeling with UML: Business Patterns at work", Wiley & Sons, Fall.

Giannini, T. C. (2011) "Distribuição geográfica de abelhas e plantas associadas através de modelagem", 140p. Tese (Doutorado) – Instituto de Biociências, Universidade de São Paulo, São Paulo, 2011.

Gimenes, M., Oliveira, P., Almeida, G. F. (2002) Estudo das interações entre as abelhas e as flores em um ecossistema de restinga na Bahia. Anais 5th Encontro sobre Abelhas de Ribeirão Preto. Universidade de São Paulo, Ribeirão Preto, pp. 117-121.

Hijimans, R. J.; Camaron, S.; Parra, J.; Jones, P.G.; Jarvis, A. (2005) Very high-resolution interpolated climate surfaces for global land areas. International Journal of Climatology. v.25 p.1965–1978.

Hijmans, R.J., Phillips, S., Leathwick, J., and ELITH, J. (2011) "dismo", ver. 0.6-3, package for R. http://cran.r-project.org/.

Hijimans, R. J. (2012) Cross-validation of species distribution models: removing spatial sorting bias and calibration with a null model. Ecology. v. 93 p. 679-688.

Hortal, J.; Roura-Pascual, N.; Sanders, N. J.; Rahbek, C. (2010) "Understanding (insect) species distributions across spatial scales", Ecography. v. 33, p. 51-53.

Hurlbert, Stuart H. (1984) Pseudoreplication and the Design of Ecological Field Experiments. Author(s): Source: Ecological Monographs, Vol. 54, No. 2.

Legendre, P.; Legendre, L. (1998) Numerical ecology, Elsevier. 2th Edition. New York. p. 853.

Machado, I. C., Vogel, S., Lopes, A. V. (2002) Pollination of Angelonia cornigera Hook. (Scrophulariaceae) by long-legged oil-collecting bees in NE Brazil. Plant Biology v. 4 p. 352-359.

Peterson, A. T.; Soberón, J.; Pearson, R. G. Anderson, R. P.; Martínez-Meyer, E.; Nakamura, M.; Araújo, M. B. (2006) "Ecological Niches and Geographic Distributions", ISBN 978-0-691-13686-8. PRINCETON UNIVERSITY PRESS. United Kingdom, 328p.

Peterson, A. T.; Soberón, J.; Pearson, R. G. Anderson, R. P.; Martínez-Meyer, E.; Nakamura, M.; Araújo, M. B. (2011) Ecological Niches and Geographic Distributions. ISBN 978-0-691-13686-8. PRINCETON UNIVERSITY PRESS. United Kingdom. 328p.

Philips, S. J.; Dudik, M.; Elith, J.; Grahan, C. H.; Lehmann, A.; Leathwick, J., Ferrier, S. (2009) "Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data." Ecological Applications. v. 19 p. 181-197.

Phillips, S. J.; Anderson, R. P.; Schapire, R. E. (2006) Maximum entropy modeling of species geographic distributions. Ecological Modelling. v. 190, p. 231-259.

Pressman, R. S. (2010) "Software Engineering: A Practitioner's Approach", 7 ed., McGraw Hill.

R Development Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna. http://www.R-project.org. 2009.

Santana, F.S.; Siqueira, M.F.; Saraiva A.M.; Corrêa, P.L.P.(2008) "A reference business process for ecological niche modelling", Ecological Informatics, v.3, p.75-86.

Soberón, J.; Peterson, A. T. (2005) "Interpretation of models of fundamental ecological niches and species' distributional areas", Biodiversity Informatics. v. 2, p. 1-10.

Sokal, R. R.; Rohlf, F. J. (1995) "Biometry: the principles and practice of statistics in biological research." New York Freeman. 3ºEdição. 887 p.

Underwood, A.J. (1997) "Experiments in Ecology: Their logical design and interpretation using analysis of variance." United Kingdom: Cambridge University Press. 504 p.

Vitt, A.C. (2006) "Contribuições à história e à epistemologia da geografia" Bertrand Brasil, RJ, PP. 101-125.

Zar, J. H. (2010) Biostatistical Analysis, Pearson Prentice Hall. 5.ed. USA. 994p.