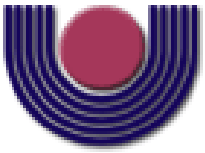


UNIOESTE – Universidade Estadual do Oeste do Paraná

CENTRO DE CIÊNCIAS EXATAS E TECNOLÓGICAS

Colegiado de Ciência da Computação

Curso de Bacharelado em Ciência da Computação

O método Best-Subset incorporado ao sistema SAHGA SDM

Bruno Borguesan

CASCAVEL

2013

BRUNO BORGUESAN

O MÉTODO BEST-SUBSET INCORPORADO AO SISTEMA SAHGA SDM

Monografia apresentada como requisito parcial para obtenção do grau de Bacharel em Ciência da Computação, do Centro de Ciências Exatas e Tecnológicas da Universidade Estadual do Oeste do Paraná - Campus de Cascavel

Orientador: Prof. Dr. Adair Santa Catarina

CASCAVEL

2013

BRUNO BORGUESAN

O MÉTODO BEST-SUBSET INCORPORADO AO SISTEMA SAHGA SDM

Monografia apresentada como requisito parcial para obtenção do Título de Bacharel em Ciência da Computação, pela Universidade Estadual do Oeste do Paraná, Campus de Cascavel, aprovada pela Comissão formada pelos professores:

Prof. Dr. Adair Santa Catarina (Orientador)
Colegiado de Ciência da Computação,
UNIOESTE

Prof. M.Eng. Carlos Jose Maria Olguín
Colegiado de Ciência da Computação,
UNIOESTE

Prof. M.Eng. Josué Pereira de Castro
Colegiado de Ciência da Computação,
UNIOESTE

Cascavel, 22 de Novembro de 2013.

AGRADECIMENTOS

A toda minha família, mas principalmente os meus pais, Cerineu e Cleuza, por tudo que batalharam para me proporcionar a melhor educação possível e nunca deixaram de me apoiar e incentivar em todos esses anos.

Em especial, ao meu irmão Adriano, que no tempo em que esteve comigo me ajudou muito na formação da pessoa que eu sou hoje.

A todos os meus amigos e amigas que de alguma forma me ajudaram, em especial a família Arapuca, que juntamente com a Balalaika tornaram minha vida acadêmica muito mais fácil e agradável.

Ao meu orientador, Adair Santa Catarina, pela grande ajuda e paciência nas atividades de orientação.

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas e Siglas

SIG	Sistemas de Informações Geográficas
AG	Algoritmos Genéticos
SAHGA	*Spatially Aware Hybrid Genetic Algorithm*
MPG	Matriz de Proximidade Generalizada
SDM	*Species Distribution Models*
CCM	Coeficiente de Correlação de Matthews
ROC	*Receiver Operating Characteristic*
AUC	*Area Under the Curve*
FAPESP	Fundação de Amparo à Pesquisa do Estado de São Paulo
CRIA	Centro de Referência em Informação Ambiental
Poli-USP	Escola Politécnica da USP
INPE	Instituto Nacional de Pesquisas Espaciais
GARP	*Genetic Algorithm for Rule-set Prediction*

Sumário

Resumo

Sistemas para geração de Modelos de Distribuição de Espécies (SDM) têm se mostrado ferramentas eficazes no estudo de diversas situações do cotidiano ecológico. O efeito da interferência humana em um nicho realizado, a predação entre espécies e a luta pela preservação de espécies em extinção são apenas alguns exemplos da utilização destes modeladores. Para gerar os SDM foi criado o sistema SAHGA SDM que utiliza um Algoritmo Genético (AG) em seu núcleo de otimização e incorpora relacionamentos geoespaciais representados através de uma Matriz de Proximidade Generalizada (MPG). Este sistema executa transições probabilísticas durante sua execução, gerando resultados variáveis. Um método para tornar o sistema mais robusto e reduzir a dependência dos dados de entrada é o *best-subset*. Este método consiste em criar "n" modelos e escolher os "m" melhores baseados em parâmetros algorítmicos, computando um modelo médio como resultado final. Para implantar esse método foram necessárias algumas alterações no sistema, como a criação de métodos para cópia de objetos dinâmicos, e o uso racional da memória do computador. Depois de efetuadas as alterações foram desenvolvidas duas abordagens para gerar os SDM: na primeira delas um único conjunto de pontos de pseudoausência é gerado para ajustar os modelos; na segunda um novo conjunto de pontos de pseudoausência é gerado para cada modelo ajustado. Para avaliação dos modelos foram criados dois métodos: um calcula as medidas de desempenho pela média das matrizes de confusão dos "m" melhores modelos; o outro gera um novo conjunto de pontos para avaliar os modelos. Após analisar os resultados dessas abordagens e métodos, concluiu-se que gerar novos pontos de pseudoausência a cada execução e fazer a avaliação desses modelos pela média das matrizes de confusão, dos "m" melhores modelos, apresentaram os resultados mais estáveis. Com isso o sistema SAHGA SDM *best-subset* mostrou-se mais robusto. Isso não quer dizer que os modelos serão mais precisos, mas sim que serão mais confiáveis. Os SDM gerados pelo sistema com o *best-subset* apresentam menor variação nos resultados que aqueles gerados sem o método best-subset. Para a espécie em estudo, a *Thalurania furcata boliviana*, o sistema gerou SDM tão bons quanto os produzidos pelo algoritmo GARP *best-subset*, que é referencia na área de modelagem de distribuição de espécies.

Palavras-chave: Modelos de Distribuição de Espécies, SAHGA SDM, Matriz de Proximidade Generalizada, *best-subset*.

Capítulo 1

Introdução

A capacidade para produzir, armazenar e recuperar dados geográficos espaço-temporais cresceu significativamente nos últimos anos. A busca pelo conhecimento contido nestes dados tornou necessário o desenvolvimento de sistemas computacionais para seu processamento, os Sistemas de Informações Geográficas (SIG). Com eles foi possível incrementar os processos de análise, mas ainda assim se tem a necessidade de integrar novos métodos de exploração e análise para esses tipos de dados (Openshaw & Openshaw, 1997; Openshaw & Abrahart, 2000; Santa Catarina, 2009).

Com o surgimento dos SIG também foi necessário o desenvolvimento de um conjunto de algoritmos "inteligentes" para explorar dados geoespaciais; entre eles estão as redes neurais, a busca heurística e os autômatos celulares. Dentre os algoritmos de busca heurística temos os Algoritmos Genéticos (AGs) (Openshaw & Openshaw, 1997; Santa Catarina, 2009).

Para incorporar relacionamentos de dados geoespaciais aos AGs desenvolveu-se o SAHGA - *Spatially Aware Hybrid Genetic Algorithm,* um algoritmo heurístico híbrido adaptável para usos múltiplos, onde os relacionamentos espaciais são representados explicitamente através de uma Matriz de Proximidade Generalizada (MPG). Com a MPG também é possível inserir conhecimento prévio sobre os elementos naturais e artificiais que compõem a região e que, na perspectiva do especialista, afetam o fenômeno em estudo (Santa Catarina, 2009).

A maioria dos organismos tem alguma capacidade de deslocar-se do seu local de nascimento para novos locais (Iwashita, 2007), seja por acontecimentos naturais (enchentes, mudanças geográficas), mudanças climáticas globais (Vivo e Carmignotto, 2004) ou, mais recentemente, pela ação humana. Esse deslocamento normalmente favorece essas espécies quando há uma competição intraespecífica ou quando a qualidade do ambiente natal sofreu algum tipo de degradação, pois locais diferentes, nestes casos, tendem a ser mais benéficos ao desenvolvimento do indivíduo (Daubenmirre, 1968; Iwashita, 2007). Desta maneira, através

deste deslocamento, a espécie procura por um novo habitat, de acordo com suas necessidades básicas de sobrevivência (Finamore, 2010).

O estudo desses deslocamentos tornou necessária a modelagem da distribuição de espécies (SDM - *Species Distribution Models*) para, por exemplo, saber onde essas espécies podem ou não se desenvolver. Os SDM, de um modo geral, podem ser utilizados para predizer e avaliar o impacto do uso acelerado da terra e da mudança do meio-ambiente na distribuição de organismos (Guisan & Zimmerman, 2000).

A utilização de sistemas de SDM tem se mostrado uma ferramenta eficaz para diversas situações do cotidiano ecológico. O efeito da interferência humana em um nicho realizado (onde realmente ocorre a espécie), a predição dos efeitos das mudanças climáticas sobre espécies e a luta pela preservação de espécies em extinção são apenas alguns exemplos da utilização destes modeladores.

Alguns destes sistemas necessitam de um dado particularmente precioso e escasso para predizer as áreas de ocorrência da espécie em estudo, os pontos de ausência. Ao inserir pontos de ausência mais informações são consideradas no processo de modelagem, contribuindo para a construção de modelos mais úteis e precisos (Santa Catarina, 2009; Finamore, 2010).

Esses pontos correspondem a locais onde os pesquisadores procuraram por indivíduos da espécie estudada, mas não a encontraram, ou seja, a espécie estava ausente (Engler *et al.*, 2004). Dados de ausência são mais difíceis de obter, pois em um dado local pode ser registrada a ausência da espécie por diferentes motivos: a) a espécie não pode ser detectada, embora presente; b) por razões históricas a espécie está ausente, embora as condições ambientais sejam adequadas; c) as condições ambientais são realmente inadequadas para a espécie (Phillips *et al.*, 2006).

A alternativa para suplantar a ausência destes pontos é gerá-los automaticamente. Pontos de pseudoausência são uma boa alternativa para os pontos de ausência, pois produzem resultados similares na SDM (Engler *et al.*, 2004).

Um algoritmo utilizado para predizer a distribuição de espécies é o BIOCLIM; esse modelo implementa o conceito de envelope bioclimático (Nix, 1986). O BIOCLIM constrói um envelope bioclimático, com base nos valores mínimo e máximo para cada variável empregada na modelagem, buscando no espaço geográfico locais onde aquelas condições se repetem, atribuindo então uma presença potencial da espécie naqueles locais "adequados" (Cassemiro *et al.*, 2012). Desta forma, regiões externas ao envelope bioclimático podem ser utilizadas para simular pontos de pseudoausência.

Para um melhor entendimento é apresentado na Figura 1.1 o resultado visual de um SDM, materializado num mapa da distribuição prevista, neste caso da espécie *Thalurania furcata boliviana*. Este mapa foi gerado a partir de 65 pontos de presença da espécie, disponíveis no conjunto de exemplos do software OpenModeller Desktop (CRIA *et al.*, 2008). A região em vermelho corresponde à área de distribuição potencial da espécie, segundo o algoritmo BIOCLIM, enquanto a região em azul representa a área onde a espécie não teria condições de existência.

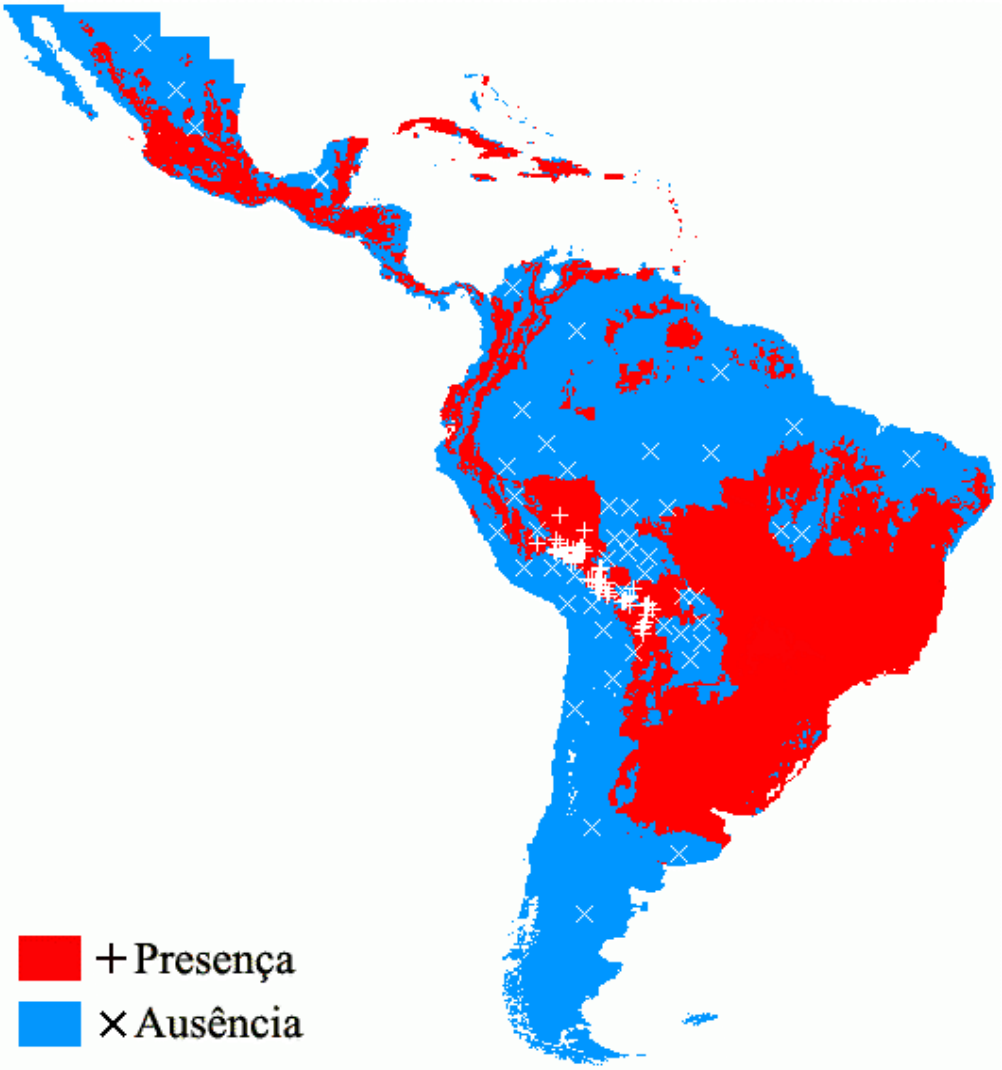


Figura 1.1 Distribuição potencial da espécie *Thalurania furcata boliviana* (BIOCLIM)
Fonte: Santa Catarina (2009)

Considerando os limites da área de distribuição potencial da espécie, outros 50 pontos de pseudoausência foram escolhidos aleatoriamente. Na Figura 1.1 o símbolo "+" representa um

ponto onde foi registrada a presença da espécie enquanto um "X" representa um ponto de pseudoausência.

O sistema SAHGA SDM, para ajustar um SDM, necessita de pontos de ausência ou pseudoausência. Seus dados de entrada são: pontos amostrais de presença e ausência da espécie, com seus relacionamentos espaciais (MPG), e o conjunto de *layers* geográficos que representam as variáveis climáticas e ambientais que podem delimitar a sobrevivência da espécie (Santa Catarina, 2009).

Para atestar a eficiência dos SDM, geralmente se utiliza a matriz de confusão de acertos e erros associados à previsão dos modelos. A matriz de confusão é utilizada para quantificar a qualidade do modelo ajustado (Santa Catarina, 2009).

Essa matriz de acertos e erros, associados à previsão dos modelos, mostra dois tipos de erros: os erros de comissão que são quando o modelo faz a predição do ponto como presente, mas a amostra está como ausente, e os erros de omissão que são quando o modelo faz a predição de ausente onde a amostra está presente, estes são considerados erros graves.

Como o método de escolha de pontos de pseudoausência utiliza processos aleatórios, os SDM ajustados pelo sistema SAHGA SDM podem apresentar variações em seus resultados, gerando modelos com taxas de erros maiores que outros.

A fim de aumentar a confiabilidade dos modelos gerados, viu-se a oportunidade de aprimorar esse sistema adicionando um método para gerar "n" modelos e considerar somente aqueles que possuem uma baixa taxa de erros de omissão e comissão, tendo como resultado final um mapa médio dos melhores resultados. Para isso utilizou-se o método *best-subset*.

O método *best-subset* consiste em gerar "n" SDM usando os mesmos dados iniciais e parâmetros algorítmicos. Destes "n" modelos serão selecionados apenas aqueles que apresentem taxas de erro de omissão inferiores a um limite máximo informado pelo usuário. Os modelos selecionados serão ordenados pelo índice de comissão e, a partir deles, serão escolhidos os "m" modelos mais próximos do valor mediano deste índice.

O resultado final deste processo é um mapa de distribuição da espécie, que consiste na média dos "m" mapas selecionados pelo método *best-subset*, juntamente com sua avaliação pela matriz de confusão.

Os modelos criados com o método *best-subset* foram analisados e comparados com os modelos gerados sem a aplicação do método, para verificar se há diferença significativa nos SDM gerados. Outra comparação foi realizada com o algoritmo "GARP – *best-subset*", que também gera SDM aplicando o método *best-subset*, mas sem considerar os relacionamentos

espaciais na geração dos resultados.

Para implementar o método *best-subset* no sistema SAHGA SDM, foi utilizado o Kit de Desenvolvimento de Sistemas Qt Creator, por ser multiplataforma e, principalmente, por haver uma versão prévia do sistema implementada nesta plataforma.

1.1 Objetivos

O objetivo geral deste trabalho é a implementação do método *best-subset* no sistema SAHGA SDM, com a perspectiva de que os modelos gerados a partir deste método sejam mais precisos que aqueles gerados a partir de uma única execução do sistema.

Para cumprir esse objetivo foram definidas duas abordagens. Na primeira são gerados novos pontos de pseudoausência para cada SDM ajustado, enquanto na outra um único conjunto de pontos de pseudoausência é criado, entretanto para cada SDM ajustado são aleatorizados novos conjuntos de treino e teste. As duas abordagens foram comparadas para se verificar qual delas é capaz de ajustar SDM mais precisos e robustos.

Como objetivo específico deste trabalho pode-se destacar a correção de alguns detalhes na interface do sistema SAHGA SDM a fim de torná-lo mais intuitivo e facilmente utilizável pelo usuário, bem como a correção de erros presente no sistema.

1.2 Motivações

A motivação para este trabalho, implantar e avaliar o método *best-subset* no sistema SAHGA SDM, deve-se ao fato do mesmo utilizar algoritmos aleatórios para gerar os pontos de pseudoausência e os conjuntos de treino e teste necessários para ajustar e avaliar os SDM; com isso, um SDM pode apresentar diferenças significativas quando comparado a outro ajustado com os mesmos dados e parâmetros de entrada.

Como o método *best-subset* seleciona os melhores modelos gerados, a perspectiva que temos é a de eliminar os modelos que apresentam altas taxas de erros, influenciadas pela aleatoriedade nos processos, no intuito de aumentar a robustez e confiabilidade dos modelos gerados.

1.3 Organização do Texto

No capítulo 2 tem-se o referencial teórico, que visa apresentar conceitos necessários à compreensão do trabalho desenvolvido. O capítulo foi dividido em duas partes: a primeira traz os fundamentos teóricos no que diz respeito aos conceitos de SDM e suas formas de avaliação; a segunda aborda o sistema SAHGA SDM, juntamente com outros conceitos que foram usados, como o algoritmo BIOCLIM e a interface gráfica do sistema.

O capítulo 3 descreve as correções efetuadas no sistema SAHGA SDM e faz um aprofundamento sobre o método *best-subset* juntamente com a descrição de sua implementação no sistema SAHGA SDM.

No capítulo 4 é apresentado um estudo de caso, onde SDM gerados pelo sistema SAHGA SDM com e sem a utilização do método *best-subset* são comparados. Outra comparação realizada foi com o SDM gerado pelo algoritmo GARP *best-subset* implementado no sistema OpenModeller Desktop.

Finalmente, no quinto e último capítulo são apresentadas as conclusões obtidas ao término do trabalho, bem como as sugestões para trabalhos futuros.

Capítulo 2

Referencial Teórico

Este capítulo tem como objetivo apresentar os conceitos necessários à compreensão do trabalho desenvolvido.

A primeira seção do capítulo apresenta os SDM, sua estrutura e utilização. Também são apresentados os mecanismos para avaliação dos SDM utilizados nesse trabalho, a matriz de confusão, curvas ROC e o índice AUC.

A segunda seção do capítulo terá como enfoque a apresentação do sistema SAHGA SDM; nela será detalhado o algoritmo BIOCLIM, utilizado para gerar os pontos de pseudoausência necessários para modelar os SDM, bem como a interface gráfica do sistema SAHGA SDM, desenvolvida com plataforma Qt.

2.1 Modelos de Distribuição de Espécies - SDM

Diversos fatores levaram os pesquisadores a estudar diferentes métodos para distribuir espécies, tais como a luta pela preservação de uma espécie em extinção, o controle de espécies invasoras, a predação entre espécies e a interferência humana diminuindo os nichos potenciais (Finamore, 2010).

A distribuição das espécies ainda é uma área muito complexa para podermos prever precisamente em todos os aspectos, de tempo e espaço, com um modelo (Guisan e Thuiller, 2005), mas com algumas técnicas podemos melhorar essas modelagens a fim de deixa-las mais próximas da realidade possível.

Cada vez mais os Modelos de Distribuição de Espécies (SDM) estão sendo usados para tentar prever a ocorrência de espécies através da geração de superfícies temáticas denominadas *layers*, indicando presença ou ausência (Guisan e Thuiller, 2005). A Figura 2.1 apresenta a estrutura geral de um sistema para a geração de SDM.

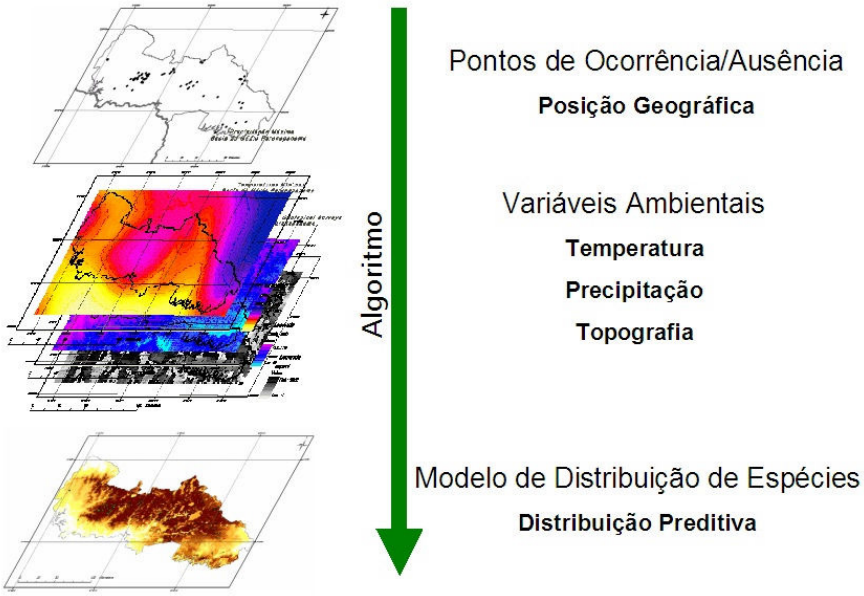


Figura 2.1 – Estrutura geral de um sistema para geração de SDM;
Fonte: Siqueira (2005)

Há três pilares no estudo de modelos matemáticos aplicados à ecologia: generalidade, realidade e precisão (Guisan e Zimmermann, 2000). Levins (1966) formulou o princípio, o qual é aceito até os dias de hoje no mundo da ecologia, que apenas dois desses pilares poderiam ser aplicados simultaneamente, enquanto o terceiro teria que ser sacrificado. Este princípio resultou nos três diferentes grupos de modelos mostrados na figura 2.2, cada um com seu objetivo principal (Guisan e Zimmermann, 2000).

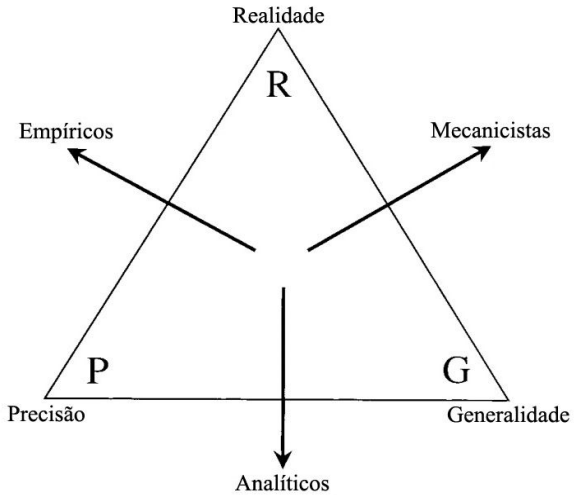


Figura 2.2 – Classificação dos modelos matemáticos aplicados em ecologia, por Levins (1966);
Fonte: Adaptado de Guisan e Zimmermann, (2000)

O primeiro grupo de modelos diz respeito à generalidade e precisão, estes são chamados de analíticos e são designados a predizer uma resposta precisa utilizando uma realidade simplificada e limitada. O segundo grupo de modelos foi feito para ser realista e geral sendo chamados de mecanicistas e suas previsões baseiam-se em relacionamentos reais de causa e efeito. O terceiro grupo de modelos especializa-se na precisão e realidade, para sacrificar a generalidade, e é chamado de empírico ou fenomenológico e, como o próprio nome diz, é baseada na observação e amostragem do mundo (Guisan e Zimmermann, 2000).

Os SDM são geralmente modelos empíricos, pois são baseados em amostras de campo (realidade) e são aplicados especificamente para modelar a ocorrência de uma espécie numa determinada área de estudo, através de métodos estatísticos e/ou computacionais (Guisan e Zimmermann, 2000; Guisan e Thuiller, 2005).

As variáveis ambientais podem exercer efeitos diretamente ou indiretamente sobre as espécies, e devem ser escolhidas de modo a representar os principais fatores que influenciam as espécies (Austin, 2002).

Todos os estudos que envolvem SDM possuem três componentes básicos (Figura 2.3): a) um conjunto de dados descrevendo a incidência ou abundância de espécies e outro conjunto contendo as variáveis explicativas; b) um modelo matemático que relaciona a espécie com a variável explicativa; c) a avaliação da utilidade do modelo através de validação ou por modelos de robustez (Guisan e Zimmermann, 2000; Iwashita, 2007).

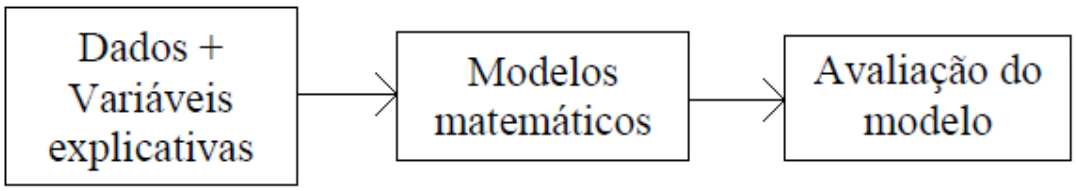


Figura 2.3: Elementos essenciais na modelagem de distribuição de espécies;
Fonte: Iwashita (2007)

Elith *et al*. (2006) classificam os SDM em dois grandes grupos baseados nos tipos de dados que alimentam os modelos. No primeiro grupo estão os modelos que utilizam apenas registros de presença (envelopes climáticos, por exemplo). No segundo grupo estão os modelos que empregam dados de presença e ausência da espécie alvo, de modo a limitar as áreas de ocorrência, diminuindo erros de falsos positivos. O segundo grupo pode ser dividido em dois subgrupos, modelos que utilizam dados de apenas uma espécie e os modelos que descrevem a presença da espécie alvo através de dados de presença de outras espécies, isto é, da comunidade (Iwashita, 2007).

Um conceito importante é o de registro zero, ou ausência, locais onde os pesquisadores procuraram por indivíduos da espécie estudada, mas não a encontraram, ou seja, a espécie está ausente (Engler *et al.*, 2004). Como comentado anteriormente, ainda na introdução deste trabalho, dados de ausência são mais difíceis de obter, pois em um dado local pode ser registrada a ausência da espécie por diferentes motivos. Devido à escassez deste tipo de dado, alguns autores vêm contornando esse problema utilizando dados de pseudoausência simulados para a modelagem (Engler *et al.*, 2004; Santa Catarina, 2009).

2.1.1 Matriz de Confusão para Avaliação dos SDM

O método de avaliação mais utilizado em SDM é a matriz de confusão de acertos e erros associados à previsão dos modelos (Tabela 2.1). O item "a" são os verdadeiros positivos, onde a predição foi presente e a amostra era presente; o item "d" são os verdadeiros negativos, onde a predição foi ausente e a amostra era ausente.

Os possíveis erros dos modelos são os falsos positivos ou erro de Comissão e falsos negativos ou erros de Omissão, itens "b" e "c" respectivamente (Iwashita, 2007).

Tabela 2.1. Matriz de Confusão

	Amostras	
Predição (Modelo)	Presente	Ausente
Presente	**a**	**b**
Ausente	**c**	**d**

Fonte: Adaptado de Meyer (2005)

- **a** e **d** = Predições corretas;
- **b** = Erro por Comissão ou superestimativa. Pode ou não estar errado, portanto não são considerados como erros graves:
  - a área é capaz de ter a espécie, porém ainda não foi registrada;
  - a área é capaz de ter a espécie, mas fatores biológicos ou históricos tem impedido que a espécie ocupasse a área, ou já foi extinta da área; ou
  - a área realmente não é capaz de ter a espécie.
- **c** = Erro por Omissão. Geralmente é considerado um erro grave, pois prediz a ausência da espécie onde ela realmente existe.

A figura 2.4 mostra como ocorrem os erros de omissão e comissão. Cada ponto representa uma presença. A área amarela corresponde à predição de presença pelo SDM enquanto a área verde corresponde à distribuição geográfica real da espécie.

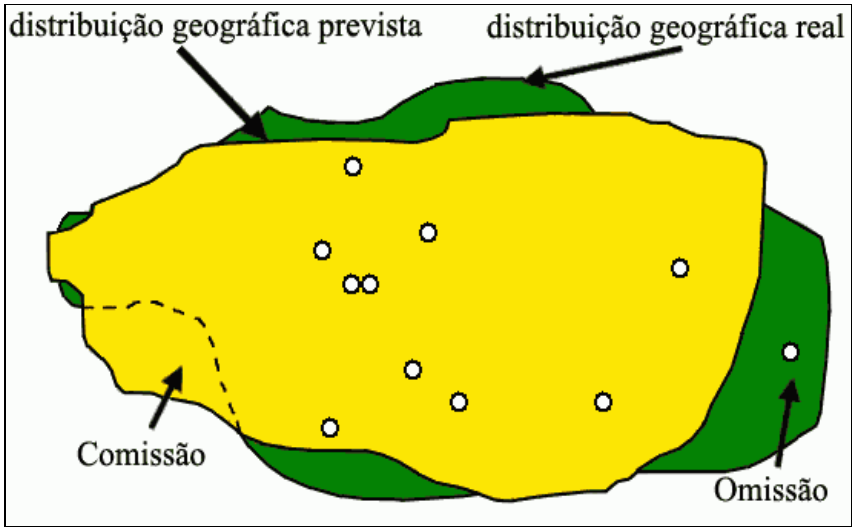


Figura 2.4 – Representação dos erros de omissão e comissão;
Fonte: Modificado de Siqueira (2005).

A partir da matriz de confusão é possível obter várias medidas para avaliação de desempenho de SDM (Fielding e Bell, 1997). Essas medidas são apresentadas na Tabela 2.2.

Tabela 2.2 – Medidas derivadas da matriz de confusão de resultados dos SDM.

Medida	Fórmula
Acurácia	$(a + d)(a + b + c + d)$
Prevalência	$(a + c) / N$
Poder de diagnóstico global	$(b + d) / N$
Taxa de classificação correta	$(a + d) / N$
Sensibilidade	$a / (a + c)$
Especificidade	$d / (b + d)$
Taxa de falso positivo (comissão)	$b / (b + d)$
Taxa de falso negativo (omissão)	$c / (a + c)$
Coeficiente de correlação de Matthews	$\frac{(a * d - b * c)}{((a + b)(a + c)(d + b)(d + c))^{1/2}}$

Fonte: Fielding e Bell (1997); Iwashita (2007); Santa Catarina (2009).

Dentre estas medidas a acurácia, a prevalência, a sensibilidade e a especificidade são as mais usadas. A acurácia mede o acerto global do modelo. Prevalência é o total (%) da área de estudo em que a espécie realmente ocorre. Sensibilidade é uma medida que descreve a probabilidade de um ponto ser corretamente classificado como ocorrência. Especificidade é a

probabilidade de um ponto ser corretamente classificado como ausência (Fielding e Bell, 1997; Guisan e Zimmermann, 2000; Segurado e Araújo, 2004; Iwashita, 2007; Santa Catarina, 2009).

O coeficiente de correlação de Matthews (CCM) é utilizado em aprendizagem de máquina como uma medida de qualidade em classificações binárias (Matthews, 1975). Este coeficiente considera todas as informações advindas da matriz de confusão e é uma medida balanceada que pode ser utilizada mesmo que as classes possuam diferentes tamanhos (Santa Catarina, 2009). Apesar de não existir medida perfeita para descrever a matriz de confusão com um único número, o CCM é considerado uma das melhores medidas com este objetivo (Baldi *et al.*, 2000).

2.1.2 Curvas ROC e o Índice AUC para Avaliação dos SDM

Outro método utilizado na avaliação de SDM é a curva ROC (*Receiver Operating Characteristic)*. Esse método consiste em fazer um gráfico entre as taxas de Sensibilidade e Especificidade mostrados na Tabela 2.2. Em outras palavras, este teste descreve a relação entre a proporção da presença observada, estimada corretamente, e a proporção das ausências observadas estimadas incorretamente. O resultado ideal para essas taxas seriam Sensibilidade com valor 1 e Especificidade com valor 0 (Pearson *et al.*, 2007). A Figura 2.5 abaixo mostra um exemplo da curva ROC.

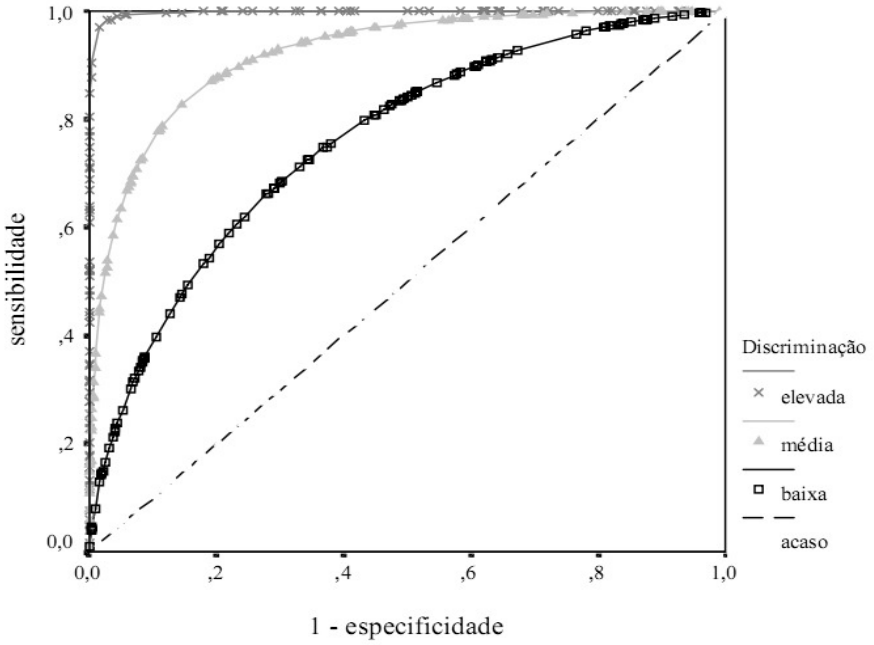


Figura 2.5 – Curvas ROC para três graus de capacidade de discriminação;
Fonte: Adaptado de Braga (2000).

A AUC (*Area Under the Curve*) caracteriza o desempenho do modelo em todos os limites possíveis, representada por um único valor calculado a partir da soma da área sob a curva ROC (Phillips *et al.*, 2006). A vantagem de usar a AUC é que ela fornece uma única medida do desempenho do modelo e da capacidade de prever corretamente a presença e a ausência. Os valores de AUC variam de 0,5 (predição aleatória) a 1,0 (predição ideal). Quando o teste é utilizado apenas com dados de presença e pseudoausências, a AUC é uma medida válida da capacidade ao modelo de classificar corretamente a presença com mais precisão do que uma previsão aleatória, em vez de distinguir a presença da ausência (Phillips *et al.*, 2006; Solorzano, 2011). Este método é bastante utilizado porque é uma medida global de desempenho independente de limites de corte, geralmente empregados na construção da matriz de confusão (Deleo, 1993; Santa Catarina, 2009).

2.2 SAHGA SDM

O sistema SAHGA SDM é um sistema que emprega o algoritmo SAHGA para modelar a distribuição potencial de espécies. O SAHGA utiliza em seu núcleo de otimização um AG híbrido, maiores detalhes sobre o núcleo do SAHGA são encontrados no trabalho de Santa Catarina (2009).

O diferencial deste sistema está na sua capacidade de construir SDM que considerem os relacionamentos espaciais presentes nos dados de entrada, representando-os através de uma MPG (Santa Catarina, 2009).

A MPG, ou matriz de vizinhança generalizada, é uma variação da matriz de proximidade. Os pesos são calculados a partir de relações espaciais no espaço absoluto como distância euclidiana e adjacência, ou com base em relações espaciais no espaço relativo, que levam em conta a conectividade de objetos em uma rede de transporte ou de comunicação, por exemplo (Aguiar *et al.*,2003; Pedrosa, 2003).

Uma MPG é composta por um conjunto de objetos geoespaciais “O” que são representados por células regulares ou polígonos, de acordo coma a representação utilizada; um grafo “G” que é constituído por um conjunto de nós e arcos, onde cada nó representa um objeto e os arcos representam os relacionamentos de vizinhança entre dois nós; e uma matriz de proximidade “V” que indica o quão próximo dois objetos “O” estão, geralmente representada em termos de adjacência ou distância euclidiana (Santa Catarina, 2009).

A figura 2.6 auxilia na compreensão da regra de proximidade utilizada no sistema SAHGA SDM. Se um objeto estiver até uma distancia "X" informada de outro objeto, a relação de proximidade entre estes recebe o peso 1; caso estejam a uma distância entre X e 2X ela recebe peso 0,5; para distâncias maiores que 2X o peso associado à relação de proximidade é 0.

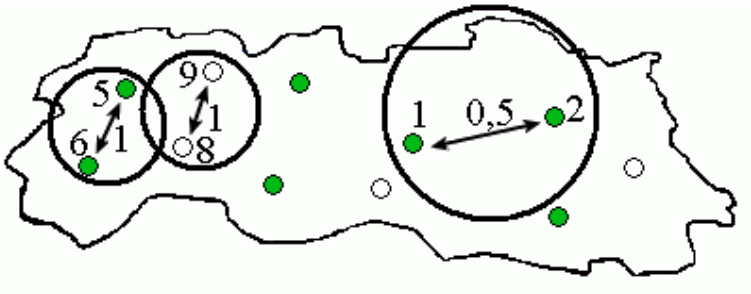


Figura 2.6 – Regra de Proximidade empregada na Construção da MPG;
Fonte: Santa Catarina (2009).

A MPG atende a dois objetivos: incorporar os relacionamentos espaciais existentes entre os pontos e a representação do conhecimento pré-existente sobre os elementos naturais e artificiais presentes no espaço, cujos efeitos são significativos na distribuição potencial da espécie modelada, como estradas, rios, cadeias de montanhas, etc. (Santa Catarina, 2009).

A estrutura geral do sistema SAHGA SDM é apresentada na Figura 2.7.

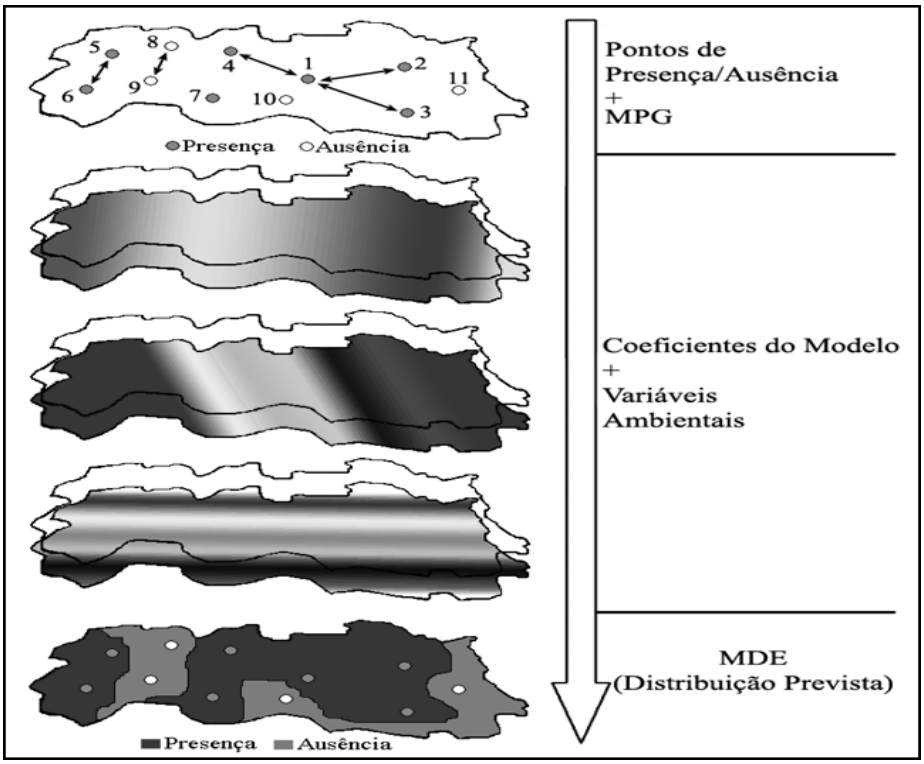


Figura 2.7 – Estrutura geral do sistema SAHGA SDM
Fonte: Santa Catarina (2009).

Como entrada, o SAHGA SDM utiliza pontos amostrais de presença e ausência da espécie, com seus relacionamentos espaciais (MPG), e o conjunto de *layers* geográficos que representam as variáveis ambientais que podem delimitar a sobrevivência da espécie (Santa Catarina, 2009).

Os coeficientes associados às variáveis ambientais, calculados para o modelo ajustado pelo sistema SAHGA SDM, quantificam o efeito destas variáveis sobre a distribuição prevista da espécie. Quanto maior este coeficiente, maior a importância daquela variável ambiental na distribuição potencial da espécie.

2.2.1 O algoritmo BIOCLIM para geração de pseudoausência

O BIOCLIM representa um dos algoritmos mais empregados para realizar prognósticos sobre os mais diversos cenários de mudanças climáticas globais (Beaumont *et al.*, 2005). Como comentado anteriormente, o BIOCLIM implementa o conceito de envelope bioclimático que é semelhante à simulação do nicho fundamental, quando utiliza os intervalos das condições ambientais necessárias para a existência da espécie sem considerar a influência da competição interespecífica ou da predação por outra espécies, através de técnicas de geoprocessamento (Nix, 1986; Busby, 1991). Visualmente o BIOCLIM é o que modela melhor o nicho fundamental, quando as amostras não possuem erros de posicionamento quando foram anotados os dados (Iwashita, 2007). O algoritmo calcula a média e o desvio-padrão para cada variável ambiental associada aos pontos de presença da espécie, assumindo uma distribuição normal. Cada variável tem seu próprio envelope representado pelo intervalo $[m - c * s, m + c * s]$, onde m é a média, c é um parâmetro que representa o ponto de corte e s é o desvio padrão. A implementação do algoritmo BIOCLIM, disponível no OpenModeller Desktop v1.0.6 (CRIA *et al.*, 2008) utiliza $c = 0{,}674$ como valor padrão.

Além do envelope, cada variável ambiental possui também os limites superior e inferior correspondentes aos valores mínimo e máximo associados ao conjunto de pontos de ocorrência da espécie (Santa Catarina, 2009).

Sendo assim, segundo Finamore (2010), cada ponto pode ser considerado:

a) Adequado: quando todos os valores das variáveis ambientais encontram-se dentro do envelope estabelecido;
b) Marginal: quando ao menos um dos valores das variáveis encontra-se fora do envelope estabelecido, porém ainda encontra-se dentro dos valores máximos e mínimos em todos os envelopes;

c) Inadequado: quando ao menos um valor das variáveis ambientais encontra-se fora de seus limites mínimo e máximo.

Segundo Finamore (2010) o algoritmo utilizado para gerar pontos de pseudoausência seleciona aleatoriamente um ponto no espaço; então se verifica se o ponto selecionado encontra-se dentro de uma região permitida; caso esteja em uma região válida, verifica-se se está contido no envelope bioclimático estipulado pelo BIOCLIM.

- Se o ponto estiver dentro do envelope bioclimático, o ponto é descartado, pois trata-se de uma região correspondente à presença.
- Se estiver fora do envelope bioclimático, mas ainda dentro dos valores máximos e mínimos para todos os *layers*, então o ponto é descartado, por haver uma probabilidade marginal de presença.
- Se o ponto estiver fora do intervalo de mínimo e máximo para qualquer uma das variáveis ambientais, o ponto é considerado como inadequado e, sendo assim, válido como um ponto de pseudoausência.

A figura 2.8 mostra um fluxograma de como foi implementado o processo de geração de um ponto de pseudoausência utilizando o algoritmo do BIOCLIM no sistema SAHGA SDM.

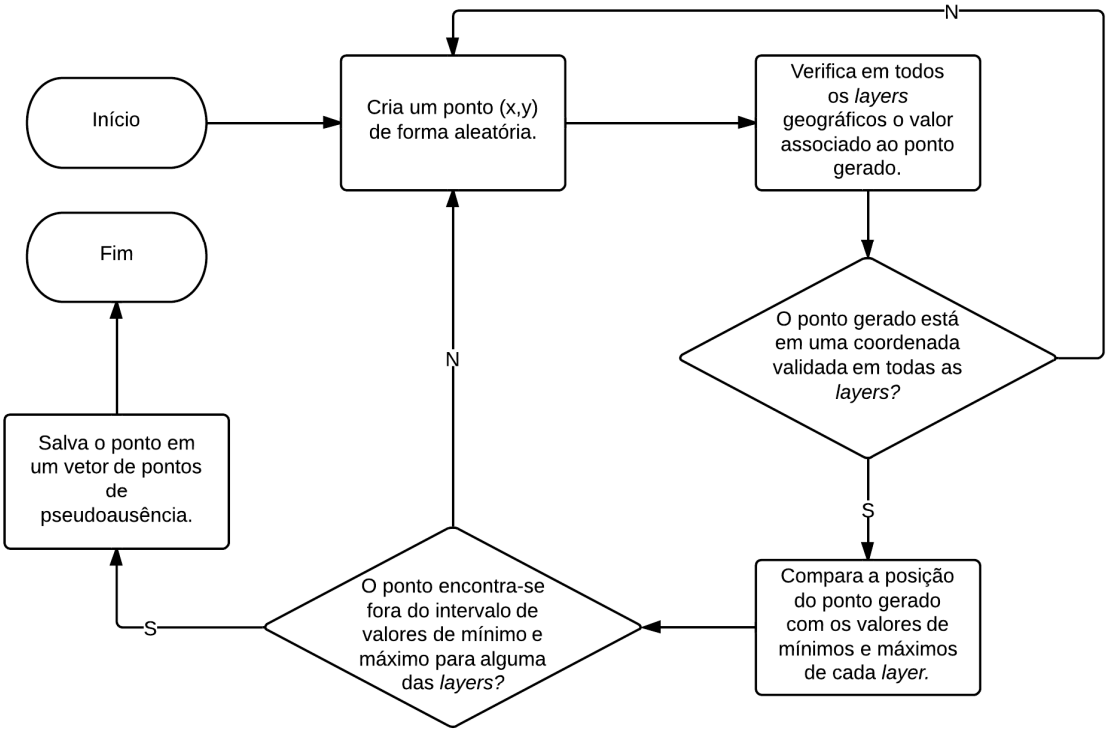


Figura 2.8 - Fluxograma do processo de geração de um ponto de pseudoausência utilizando o algoritmo BIOCLIM

2.2.2 Espécie *Thalurania furcata boliviana* Boucard, 1894

A base de dados *Thalurania furcata boliviana* é um dos conjuntos de dados fornecidos com o instalador do software OpenModeller Desktop. A distribuição potencial desta espécie foi o estudo de caso a ser avaliado neste trabalho. A base contém 65 pontos de presença da referida espécie; também são disponibilizados 8 *layers* geográficos correspondentes às variáveis: precipitação acumulada no trimestre mais úmido, precipitação acumulada no trimestre mais quente, precipitação anual, temperatura média anual, temperatura média no trimestre mais frio, temperatura média no trimestre mais seco, temperatura média no trimestre mais quente e temperatura média no trimestre mais úmido (Santa Catarina, 2009).

2.2.3 Interface do SAHGA SDM

O sistema SAHGA SDM foi desenvolvido por Santa Catarina (2009). Como trabalho posterior adicionou-se uma interface gráfica ao sistema, que inicialmente era executado em modo texto.

A interface do SAHGA SDM foi desenvolvida por Finamore (2010), que tinha como objetivo principal implantar métodos para geração de pontos de pseudoausência e também criar a interface gráfica do sistema, usando a plataforma Qt para o desenvolvimento e o sistema OpenModeller Desktop como modelo para interface. A figura 2.9 mostra algumas das telas criadas para o sistema SAHGA SDM.

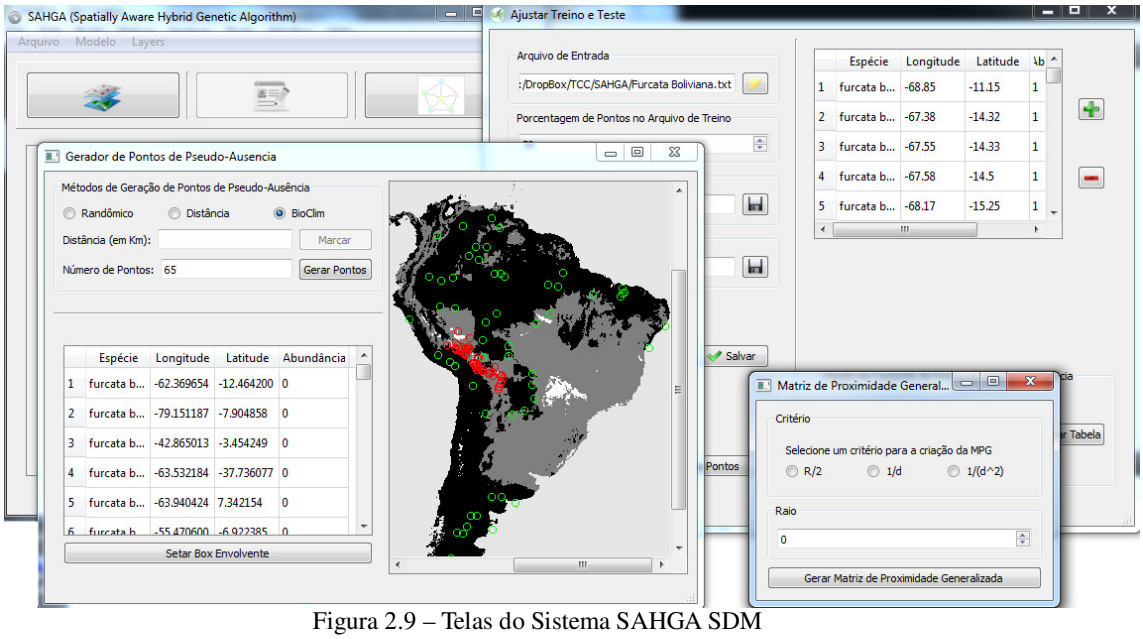


Figura 2.9 – Telas do Sistema SAHGA SDM

O Qt é uma biblioteca multiplataforma para desenvolvimento de interfaces de usuário. Criado pela empresa norueguesa Trolltech em março de 1994 e depois comprado pela Nokia, em 2008, a fim de desenvolver os aplicativos para a sua tecnologia baseada em Symbian, acabou sendo adquirido pela empresa Digia em 2012.

O Qt é amplamente utilizado para o desenvolvimento de interfaces gráficas de usuário (KDE, Skype, VirtualBox, VLC Media Player, etc.) (Finamore, 2010).

O sistema OpenModeller Desktop é um sistema de modelagem de distribuição de espécies financiado pela Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), desenvolvido pelo Centro de Referência em Informação Ambiental (CRIA), Escola Politécnica da USP (Poli-USP) e o Instituto Nacional de Pesquisas Espaciais (INPE) (CRIA *et al.*, 2008).

Capítulo 3

Metodologia

O SAHGA SDM tem como dados de entrada um conjunto de pontos de presença e ausência da espécie a ser modelada e um conjunto de *layers* geográficos associados às variáveis ambientais. A espécie *Thalurania furcata boliviana*, que foi utilizada como estudo de caso, possui um conjunto com 65 pontos de presença e 8 *layers* geográficos.

Depois de inseridas estas informações no sistema SAHGA SDM, escolhe-se um algoritmo, como o BIOCLIM, para geração de um número de pontos de pseudoausência, número este também definido pelo usuário. O conjunto de pontos de presença e pseudoausência é dividido aleatoriamente em dois conjuntos, treino e teste, cada um com 50% do total dos pontos. Posteriormente o modelo de distribuição de espécie é gerado, em função de parâmetros escolhidos pelo usuário, e projetado para toda a área geográfica do estudo.

Na última etapa o modelo é avaliado e o sistema fornece um relatório com os valores da avaliação da matriz de confusão apresentados na Tabela 2.2, bem como os dados necessários para construção da curva ROC e o índice AUC. A figura 3.1 mostra essa sequência de passos.

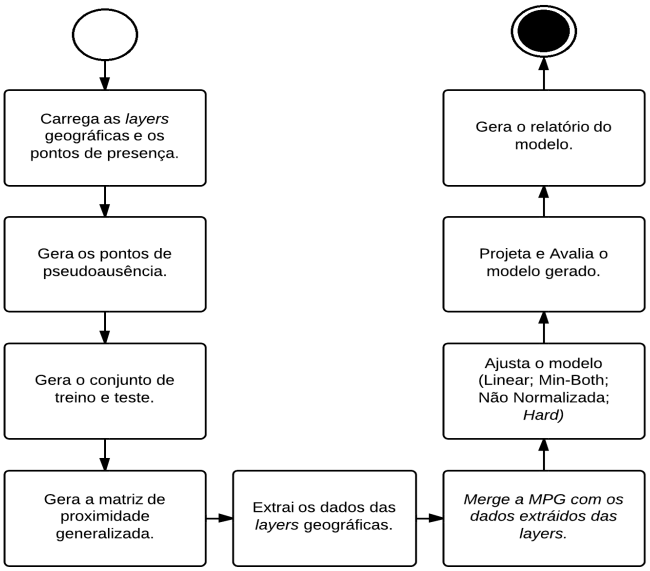


Figura 3.1 – Diagrama de atividades do sistema SAHGA SDM

Como mencionado anteriormente o objetivo principal desse trabalho é a implantação do método *best-subset* no sistema SAHGA SDM. Para atingir este objetivo foi necessário estudar e compreender o algoritmo; como consequência deste estudo foram detectados e corrigidos alguns erros no sistema. Este capítulo focará na descrição das correções realizadas, na explicação do método *best-subset* e sua implementação.

3.1 Correções no sistema SAHGA SDM

Para execução desta tarefa, inicialmente estudou-se a primeira versão do sistema SAHGA SDM, produzido por Santa Catarina (2009), e a segunda versão com interface gráfica e métodos para geração dos pontos de pseudoausência, desenvolvida por Finamore (2010). As alterações realizadas na segunda versão visavam corrigir os relatórios de saída do sistema, para ficarem em consonância com os apresentados pelo sistema SAHGA SDM em sua primeira versão. Para essas correções utilizou-se o *debug* do Qt Creator, o que possibilitou identificar e corrigir os erros.

O erro mais grave identificado na segunda versão do sistema SAHGA SDM refere-se à cópia de objetos com atributos dinâmicos. A atribuição direta entre estes objetos não realiza a cópia destes atributos; ela copia apenas o endereço (referência) dos mesmos. Em uma única execução este erro não é percebido, pois estes objetos são alterados uma única vez; o erro foi detectado quando se fez necessária a geração de "n" modelos no método *best-subset* e estes objetos precisaram ser reutilizados. Para corrigir esse problema foram criados métodos para atribuição destes objetos, que copiam todos os atributos internos, inclusive aqueles criados dinamicamente.

Outro problema identificado, na segunda versão do sistema, diz respeito ao uso inconsequente da memória RAM. Todos os objetos alocados dinamicamente não eram removidos da memória, causando estouro de memória na execução do método *best-subset*. Para corrigir esse problema foi analisado e adicionado corretamente os destrutores desses objetos.

3.2 *Best-subset*

Como o nome do método já diz ele tem como objetivo selecionar o melhor subconjunto de resultados baseado em alguns parâmetros pré-definidos.

Como base para o desenvolvimento desses parâmetros foi usado o modelo de *best-subset* de Anderson *et al.* (2003), implantado na plataforma OpenModeller Desktop. Na Tabela 3.1 são mostrados alguns desses parâmetros (Iwashita, 2007).

Tabela 3.1 - Parâmetros específicos do GARP *best-subset* utilizados no trabalho.

Índice	**Valores Exemplo**
Número de Execuções (n)	100
Limite de Omissão (*hard*)	10%
Número Modelos Abaixo do Limite de Omissão	30
Limite de Comissão	50%
Número de Modelos (m)	20

Fonte: Iwashita (2007).

- Número de Execuções: número de amostras utilizadas para calcular os erros de comissão;
- Limite de Omissão (*hard*): limite da taxa de erros de omissão para os modelos a serem analisados;
- Número Modelos Abaixo do Limite de Omissão: número de modelos para ordenar por comissão desde que estejam abaixo do limite de omissão "hard", caso não tenha modelos suficiente abaixo do limite de omissão o sistema utiliza somente os modelos que estão abaixo, desconsiderando esse parâmetro;
- Limite de Comissão: indica a posição relativa a ser utilizada para selecionar os modelos de acordo com os erros de comissão, onde 50% representa a mediana do conjunto de modelos pré-selecionados;
- Número de Modelos: número de modelos utilizados para se obter o resultado médio.

Para aplicar o método nesse trabalho primeiramente foi preciso gerar "n" modelos de distribuição de espécies, todos gerados usando os mesmos dados iniciais e parâmetros algorítmicos informados pelo usuário, esses parâmetros serão detalhados no capítulo seguinte.

A partir destes "n" modelos aplicou-se o método *best-subset*, que consiste nos cinco passos descritos a seguir.

1. Ordenar, de modo crescente, os "n" modelos pela sua taxa de erros de omissão;

2. Selecionar apenas aqueles que apresentam taxas de erros de omissão inferiores ao limite máximo admitido nos parâmetros, evitando que modelos com altas taxas de erro grave sejam escolhidos;
3. Os modelos selecionados no passo anterior são ordenados pela sua taxa de erros de comissão;
4. Selecionar os "m" modelos mais próximos do valor mediano das taxas de erro de omissão, evitando modelos com superajuste, que é quando o modelo prediz precisamente cada ponto de presença, e modelos com superpredição, que é quando o modelo prediz área de presença em quase todo o mapa;
5. Construção de um mapa médio baseado nos "m" melhores modelos gerados. Este mapa médio corresponde ao SDM gerado através do método *best-subset*.

A figura 3.2 apresenta um exemplo do método *best-subset* para distribuição de espécies

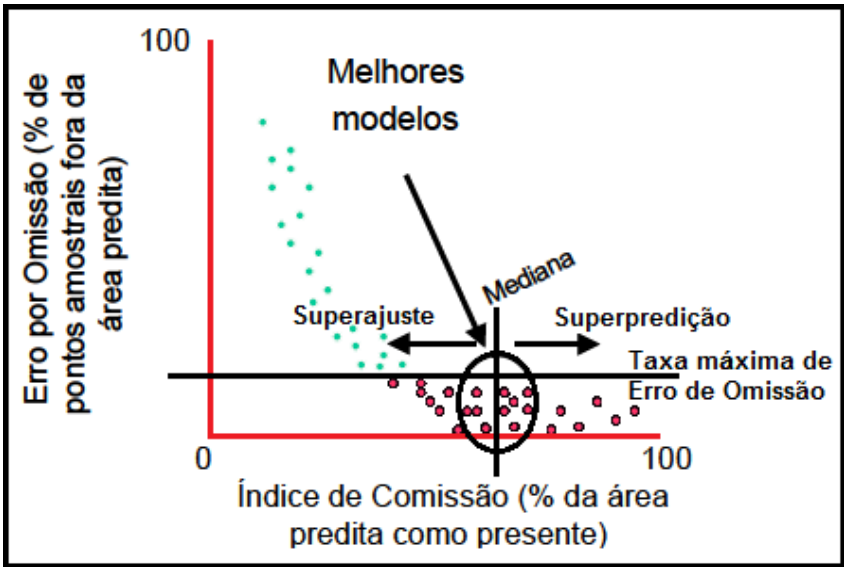


Figura 3.2 - Modelo *Best-subset*;
Fonte: Adaptado de Meyer (2005)

Cada ponto da figura 3.2 representa o resultado de uma das "n" execuções do sistema, cada qual com seu valor de Erro por Omissão e Índice de Comissão. Os "m" modelos escolhidos são aqueles circulados que estão abaixo do limiar máximo de Erro de Omissão e os mais próximos da mediana dos erros de Comissão (Meyer, 2005).

3.2.1 Implantação do método *best-subset* no sistema SAHGA SDM

Após o estudo do método e correção dos erros identificados no sistema, iniciou-se a codificação do método *best-subset*. Para aplicar o *best-subset* foram desenvolvidas duas abordagens.

A primeira, chamada T1, consiste em gerar um único conjunto de pontos de pseudoausência, pelo algoritmo BIOCLIM, e utilizá-lo para ajustar "n" modelos necessários ao método *best-subset*.

A segunda, chamada T2, gera um conjunto de pontos de pseudoausência para cada um dos "n" modelos ajustados pelo sistema SAHGA SDM.

A figura 3.3 mostra os diagramas de atividades das duas abordagens, permitindo visualizar os passos realizados em cada uma delas, bem como suas diferenças. A área selecionada corresponde ao laço de repetição onde são gerados os "n" modelos para aplicação do método *best-subset*.

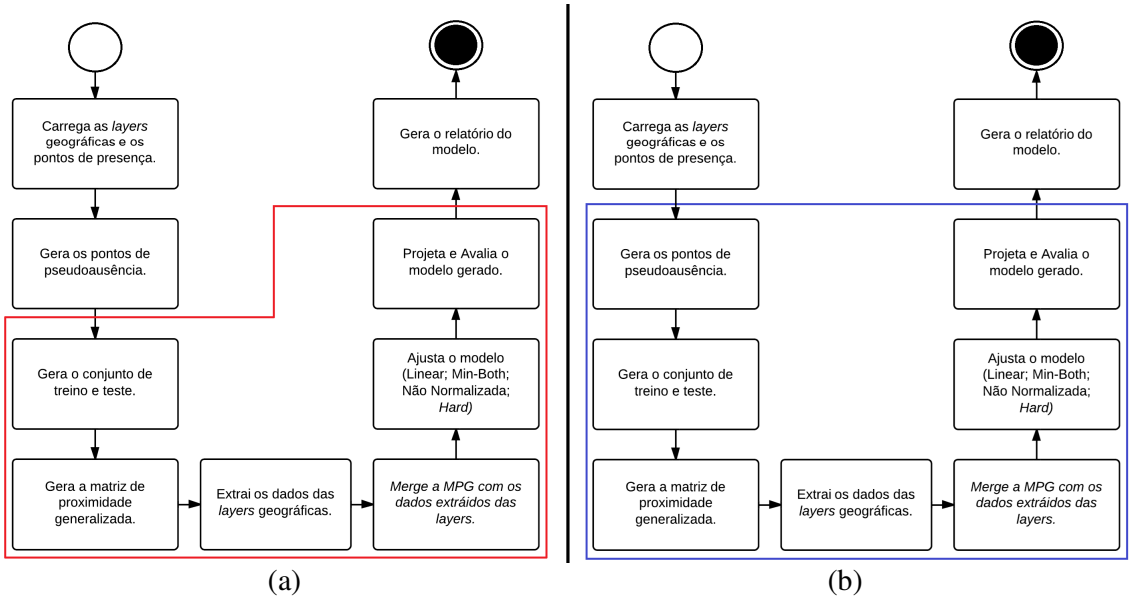


Figura 3.3 – Diagramas de atividades das abordagens desenvolvidas para aplicação do método *best-subset*. (a) abordagem T1; (b) abordagem T2

Para cada um dos "n" modelos ajustados pelo sistema SAHGA SDM são salvos a matriz de confusão, resultante da avaliação do modelo, e a *layer*, em *grayscale*, correspondente à projeção do modelo para a área em estudo, ou seja, o mapa de predição da distribuição da espécie. Ao fim das "n" execuções aplica-se o *best-subset*, conforme descrito no item 3.2.

Como resultado final constrói-se um mapa médio, a partir dos "m" mapas projetados, correspondentes aos modelos selecionados pelo método *best-subset*.

Para colorir o mapa médio de saída, com a finalidade de gerar um mapa com interpretação mais fácil, é calculada uma pseudocor baseada no valor do tom de cinza de cada pixel, que varia entre 0 e 255. O código para pseudocoloração do mapa médio e mostrado na Figura 3.4.

```
if(S->m->MapaMedio->M->M[i][j] == NoData){//se não possui dados, pinta de branco.
    data2[R] = 255;
    data2[G] = 255;
    data2[B] = 255;
}else{
    if(S->m->MapaMedio->M->M[i][j]>127){//se possui dados, e for maior que 127 é presença.
        data2[R] = 127-((S->m->MapaMedio->M->M[i][j])-128);
        data2[G] = 0;
        data2[B] = 127+((S->m->MapaMedio->M->M[i][j])-128);
    }else{//se possui dados, e for menor/igual que 127 é ausência.
        data2[R] = 127-((S->m->MapaMedio->M->M[i][j])-127);
        data2[G] = 0;
        data2[B] =  127+((S->m->MapaMedio->M->M[i][j])-127);
    }
}
```

Figura 3.4 – Código utilizado para fazer a pseudocoloração de um ponto do mapa médio

Quando uma região no mapa não foi georreferenciada, ou seja, não possui valores em sua posição de latitude e longitude, ela é pintada com branco, por exemplo, as regiões oceânicas.

Quanto mais próximo de 0 for o tom de cinza do pixel, mais azul será a pseudocor, correspondendo a uma região de ausência; quanto mais próximo de 255 for o tom de cinza do pixel, mais vermelha será a pseudocor, correspondendo a uma região de presença.

A Figura 3.5. a seguir mostra um exemplo de saída desse código.

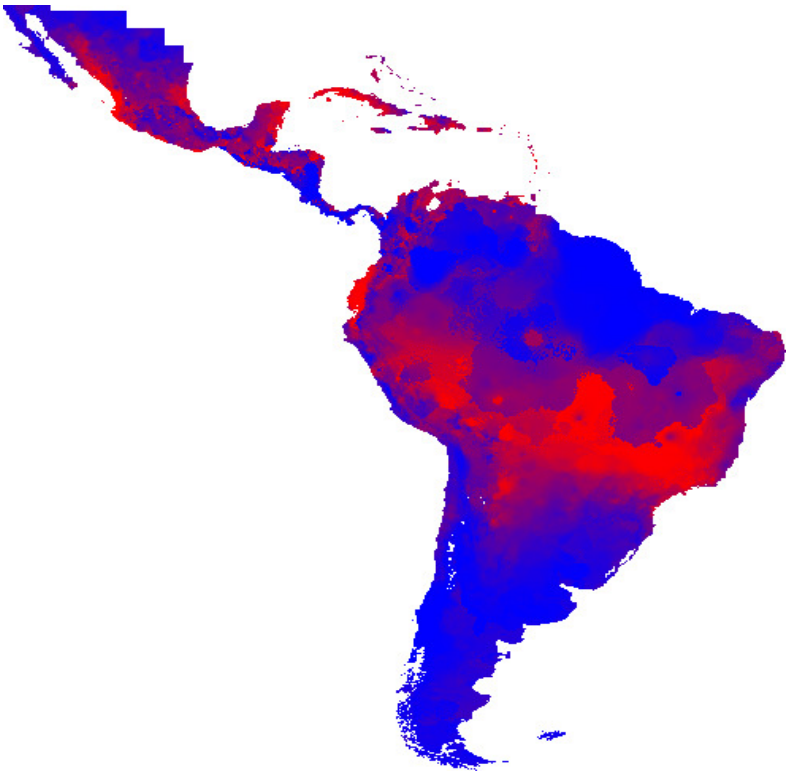


Figura 3.5 – SDM gerado a partir do SAHGA SDM – *best-subset;*

Este processo de pseudocoloração cria um efeito de transição entre as cores azul e vermelha, ou seja, uma transição entre as regiões consideradas como ausência e as regiões consideradas como presença, gerando um gradiente de cor (degradê) na representação do mapa médio.

Capítulo 4

Estudo de Caso

Os dados e parâmetros algorítmicos empregados na avaliação das duas abordagens desenvolvidas (T1 e T2) foram baseados nos estudos de caso desenvolvidos por Finamore (2010). Os dados são da espécie *Thalurania furcata boliviana*, com 65 pontos de presença, e 8 *layers* geográficas correspondentes a variáveis climáticas. Para geração de pontos de pseudoausência utilizou-se o algoritmo BIOCLIM, implementado no sistema; foram gerados 65 pontos de pseudoausência, criando um conjunto amostral balanceado.

Para geração da MPG considerou-se um raio de 60 km; ou seja, se dois pontos, presença ou ausência, estiverem distantes entre si em até metade do raio (até 30 km), eles estão espacialmente relacionados com peso 1; se os dois pontos distarem entre 30 km e 60 km o peso desse relacionamento é 0.5; para distâncias maiores que 60 km o peso do relacionamento é nulo.

Em cada experimento dividiu-se, aleatoriamente, o conjunto de pontos de presença e pseudoausência da espécie em dois subconjuntos, o subconjunto de treino e o subconjunto de teste, cada um deles com 50% do total dos pontos, sendo metade desses pontos presença e a outra metade ausência.

Os parâmetros genéticos utilizados para fazer o ajuste dos modelos foi o *Hard,* por ser um conjunto de parâmetros genéticos que visa assegurar ao algoritmo qualidade da resposta com tempo de convergência aceitável (Santa Catarina, 2009). Esses parâmetros são mostrados na Tabela 4.1.

Optou-se por ajustar modelos do tipo linear, por terem se mostrado suficientes para relacionar a ocorrência da espécie com as variáveis ambientais, para a espécie escolhida. Como função objetivo, o tipo escolhido foi aquele que maximiza o acerto global do modelo, diminuindo os erros de omissão e comissão, e ao mesmo tempo maximiza a qualidade do ajuste do modelo, minimizando a soma dos desvios quadrados. O sistema SAHGA SDM implementa o tipo Min-Both, que implementa a maximização do ajuste do modelo e também

o acerto global do modelo sendo, portanto, o escolhido para a realização dos experimentos (Santa Catarina, 2009; Finamore 2010).

Tabela 4.1 – Conjunto de Parâmetros do SAHGA

Parâmetros	***Hard***
Tamanho da População	100
Número de ciclos	20
Temp. mínima	0,001
Temp. máxima	3
Constante de resfriamento	0,9
Número de repetições	5
Tamanho Elite	1
Taxa cruzamento	80%
Taxa mutação	1%

Fonte: Santa Catarina, 2009

No sistema SAHGA SDM não faz sentido normalizar a variável dependente, pois ela só faz sentido quando o algoritmo SAHGA está processando outros tipos de variáveis dependentes, ou seja, aquelas cuja faixa de variação não está no intervalo entre 0 e 1 (Finamore 2010).

Para avaliar os SDM gerados, ou seja, os mapas médios resultantes do *best-subset* com as abordagens T1 e T2, foram criados dois métodos (M1 e M2). A Figura 4.1 ilustra as atividades realizadas em cada um deles, destacando suas diferenças.

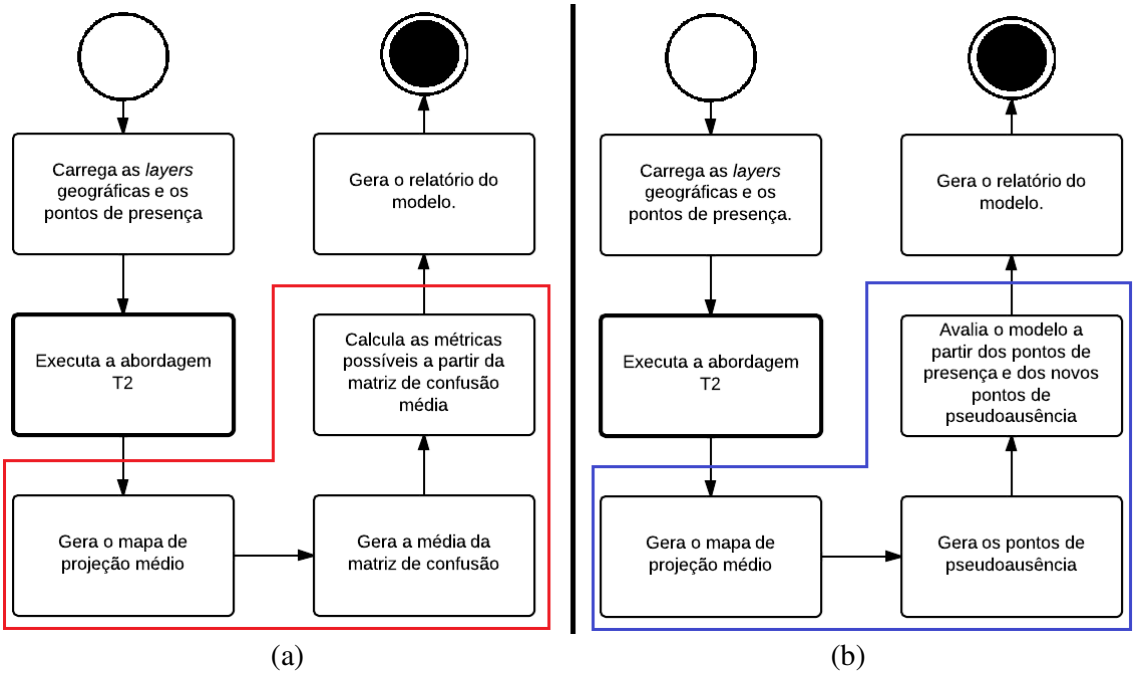


Figura 4.1 – Diagramas de atividades dos métodos desenvolvidos para avaliação dos modelos gerados. (a) método M1; (b) método M2

O primeiro, chamado M1, faz a avaliação do SDM através da média da matriz de confusão dos modelos selecionados pelo *best-subset*. A partir da matriz de confusão média são calculados a acurácia e o CCM; entretanto, não é possível gerar a média de uma curva ROC e calcular o AUC. O segundo, chamado de M2, avalia a capacidade preditiva do SDM (o mapa médio em *grayscale*), utilizando um conjunto teste formado pelos pontos de presença da espécie acrescidos de um novo conjunto de pontos de pseudoausência. Esta avaliação gera uma nova matriz de confusão permitindo calcular a acurácia e o CCM, bem como traçar a curva ROC e estimar o índice AUC. O método M2 permite avaliar o SDM através de um novo conjunto de pontos de teste.

4.1 Resultados

Um dos objetivos para criação das duas abordagens citadas anteriormente, a T1 e a T2, foi analisar se os SDM gerados com diferentes dados de pseudoausência (T2) são mais robustos que aqueles gerados com um único conjunto de pseudoausência (T1), quando aplicado o método *best-subet*.

Para analisar os SDM gerados com as abordagens T1 e T2 foram empregados os métodos M1 e M2. A utilização destes dois métodos não visa descobrir qual o SDM mais preciso, mas sim aquele que possui menor variância nas medidas de avaliação. Baixa variância nas medidas de avaliação é um indicativo da maior confiabilidade do SDM gerado.

4.1.1 Resultado produzido com a abordagem T1

Em cada teste realizado com o sistema SAHGA SDM *best-subset*, obteve-se como saída um SDM (mapa) em tons de vermelho (presença) e azul (ausência). Os resultados encontrados foram avaliados comparando as medidas calculadas através dos métodos M1 e M2. O SDM1 gerado para a espécie *Thalurania furcata boliviana*, utilizando a abordagem T1, pode ser visto na figura 4.2.

Observando-se as medidas de avaliação desse teste, apresentadas na Tabela 4.2, percebe-se que os métodos de avaliação M1 e M2 apresentam resultados similares; devido à pequena variação na taxa de erros de comissão, o método M2 apresenta um CCM ligeiramente menor (0,803 < 0,815). Estes resultados permitem concluir que o SDM1, utilizando a abordagem T1, tem boa capacidade preditiva, com acurácia superior a 89%, independente do método de avaliação utilizado.

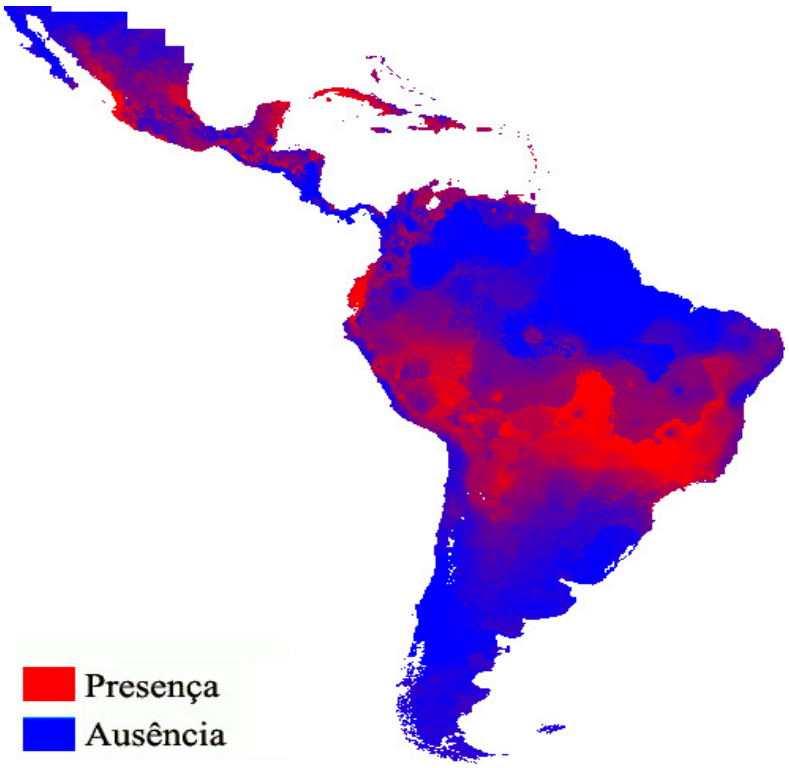


Figura 4.2 – SDM1 ajustado pelo SAHGA SDM *best-subset* utilizando a abordagem T1

O método M2 possibilita construir a curva ROC (Figura 4.3) e calcular o índice AUC. Neste caso, o alto valor para o índice AUC (0,955) permite concluir que o SDM1 ajustado possui ótima capacidade discriminatória, ou seja, alta capacidade para predizer as regiões de presença e ausência.

Tabela 4.2 – Medidas de avaliação do SDM1

Medidas para o SDM1	**Método**	
	M1	**M2**
Acurácia	89,9%	89,2%
Erro de Omissão	0%	0%
Erro de Comissão	20,2%	21,6%
CCM	0,815	0,803
AUC		0,955

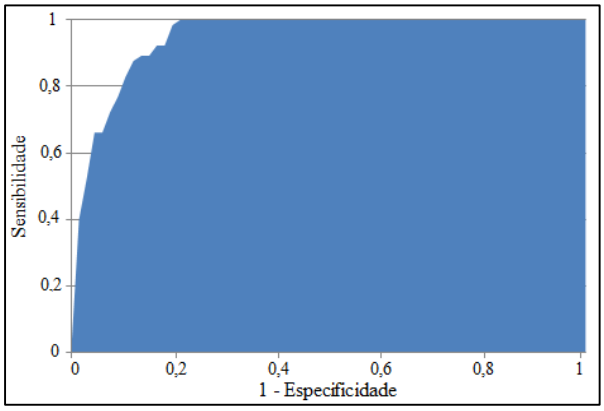


Figura 4.3 – Curva ROC para o SDM1

4.1.2 Resultado produzido com a abordagem T2

Para avaliar a abordagem T2 foram ajustados SDM com os mesmos parâmetros algorítmicos utilizados anteriormente. Na avaliação destes SDM também foram utilizados os métodos M1 e M2. A Figura 4.4 apresenta o SDM2, gerado pelo sistema SAHGA SDM, com

o método *best-subset*, empregando a abordagem T2.

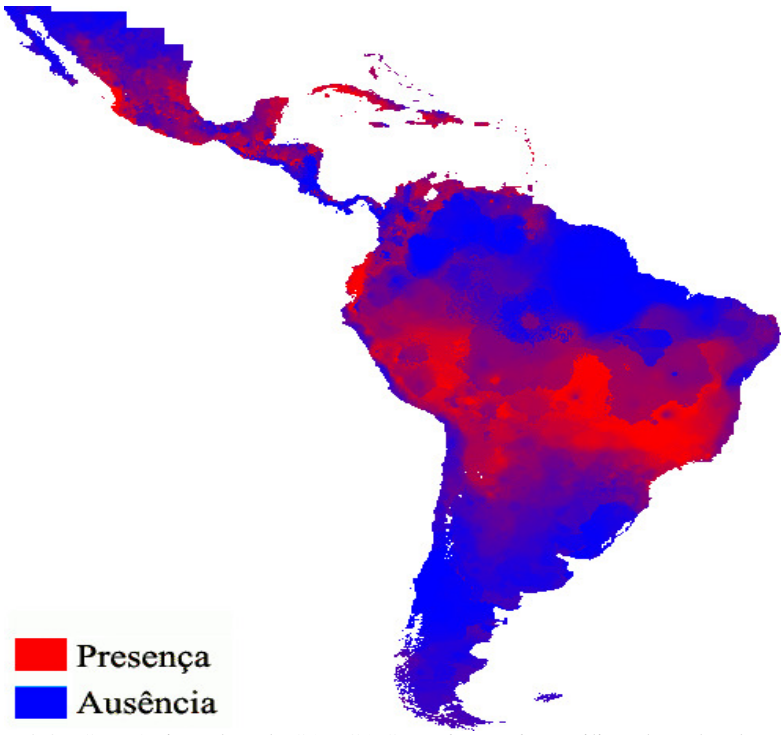


Figura 4.4 – SDM2 ajustado pelo SAHGA SDM *best-subset* utilizando a abordagem T2

Observando-se as medidas de avaliação para o SDM2, apresentados na Tabela 4.3, percebe-se que os resultados calculados através dos métodos M1 e M2 são similares. Devido à pequena variação na taxa de erros de comissão, o método M2 apresenta um CCM ligeiramente maior (0,791 > 0,781). Estes resultados permitem concluir que o SDM2, utilizando a abordagem T2, também teve boa capacidade preditiva, com acurácia superior a 87%, independente do método de avaliação utilizado. A curva ROC (Figura 4.5) e seu índice AUC (0,965) permitem concluir que o SDM2 tem elevada capacidade preditiva.

Tabela 4.3 – Medidas de avaliação do SDM2

Medidas para o SDM2	**Método**	
	M1	**M2**
Acurácia	87,9%	88,5%
Erro de Omissão	0%	0%
Erro de Comissão	24,2%	23,1%
CCM	0,781	0,791
AUC		0,965

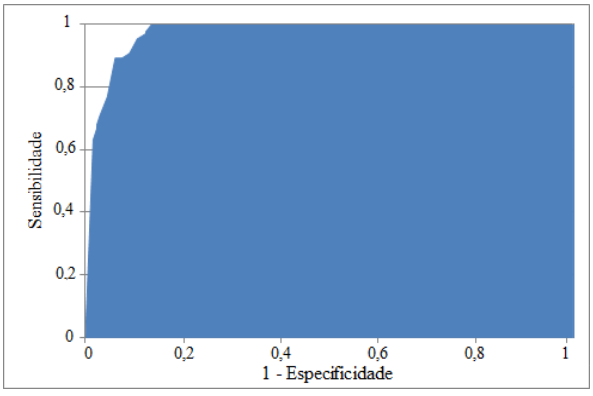


Figura 4.5 – Curva ROC para o SDM2

4.1.3 Comparação entre as abordagens e métodos de avaliação

Para comparar os SDM gerados pelo SAHGA SDM com os SDM gerados por outros sistemas, fez-se uma análise sobre qual abordagem e qual método se mostraram mais confiáveis, para isso foram gerados novos testes com os mesmos parâmetros algorítmicos apresentados no início do capítulo.

Foram feitas comparações entre os resultados de dois SDM (SDM3 e SDM4) ajustados com a abordagem T1, os SDM gerados são mostrados na Figura 4.6 a seguir.

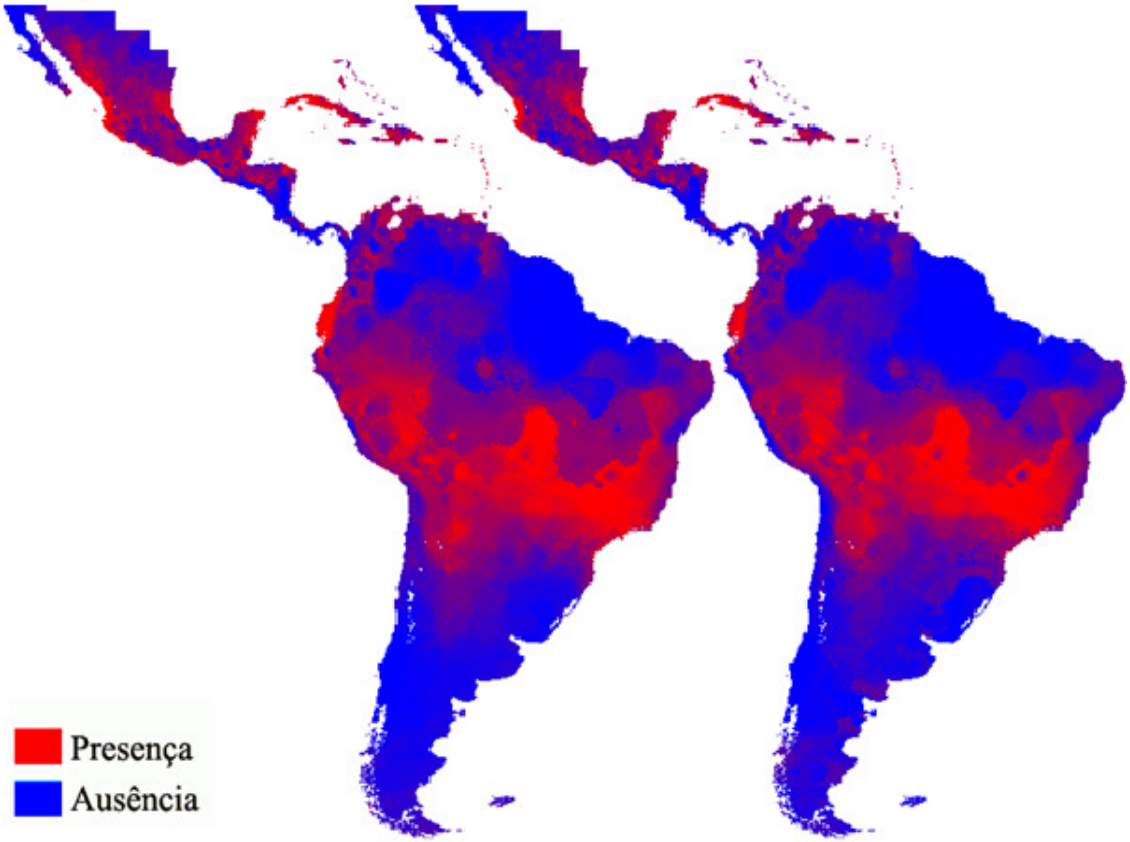


Figura 4.6 – SDM3 e SDM4, gerados com a abordagem T1

Os SDM gerados, mostrados na Figura 4.6, apresentam uma suave diferença, indicando a robustez do sistema SAHGA SDM com *best-subset*, pois o modelo foi gerado através da média dos melhores modelos entre os 100 modelos ajustados inicialmente

Na sequência foram comparados os resultados de outros dois SDM (SDM5 e SDM6) ajustados com a abordagem T2. A Figura 4.7 permite observar que os SDM5 e SDM6 também são similares, apresentado pequenas diferenças em algumas áreas de transição. Isso mostra que a abordagem T2 também apresenta robustez na geração dos SDM.

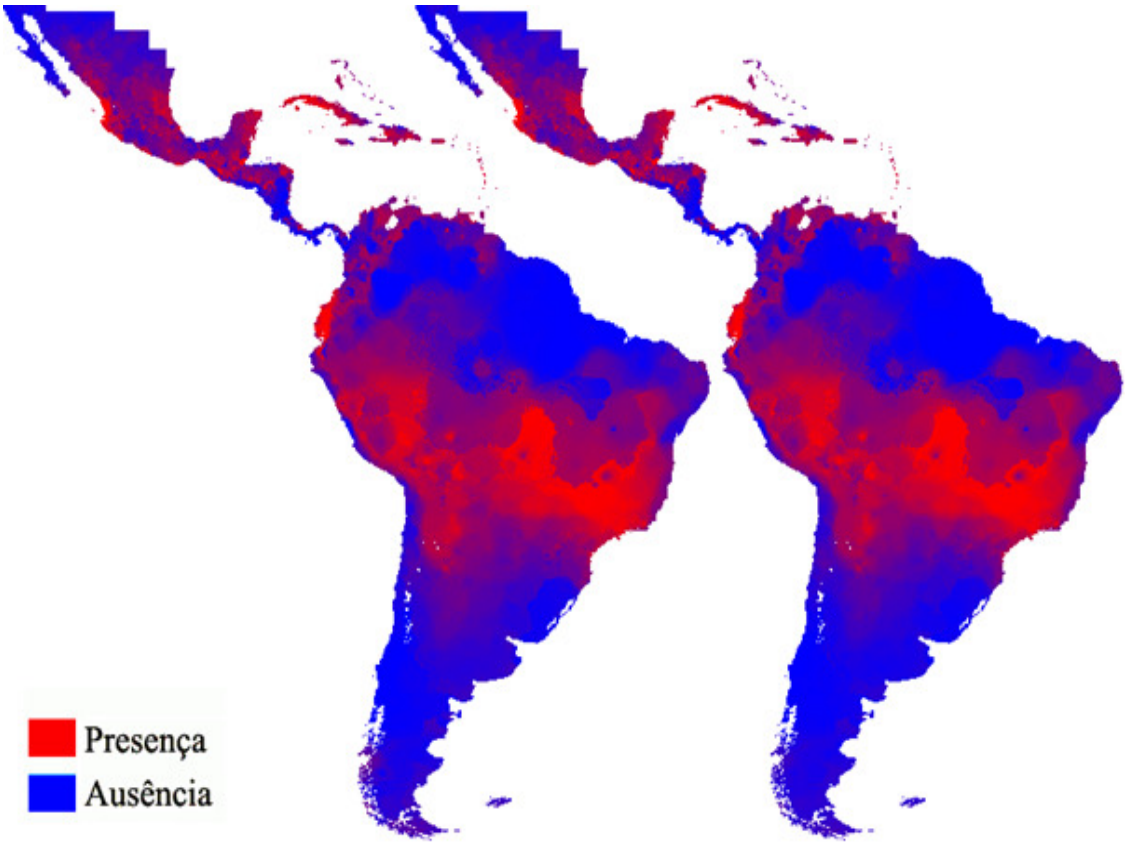


Figura 4.7 – SDM5 e SDM6, gerados com a abordagem T2

Analisando as Figuras 4.6 e a 4.7 observaram-se modelos projetados (mapas) similares, pois ambas calculam um modelo médio a partir dos melhores entre 100 modelos ajustados. Para avaliar qual a melhor abordagem para criar os SDM compararam-se as medidas de avaliação calculadas para cada modelo.

Os resultados apresentados na Tabela 4.4, referentes aos testes da abordagem T1, mostram que, apesar da aparente similaridade entre os mapas na Figura 4.6, as suaves diferenças visuais configuram-se em medidas distintas. Na avaliação, utilizando tanto o método M1 como o método M2, a diferença entre os testes está na taxa de erros de comissão; elas foram maiores no SDM3 utilizando o método M1, porém maiores no SDM4 quando utilizado o método M2, indicando alta variação de resultados para um sistema que calcula a média dos melhores modelos, mostrando que a abordagem T1 não foi capaz de gerar modelos com medidas de avaliação similares.

Tabela 4.4 – Comparação entre dois SDM gerados com a abordagem T1

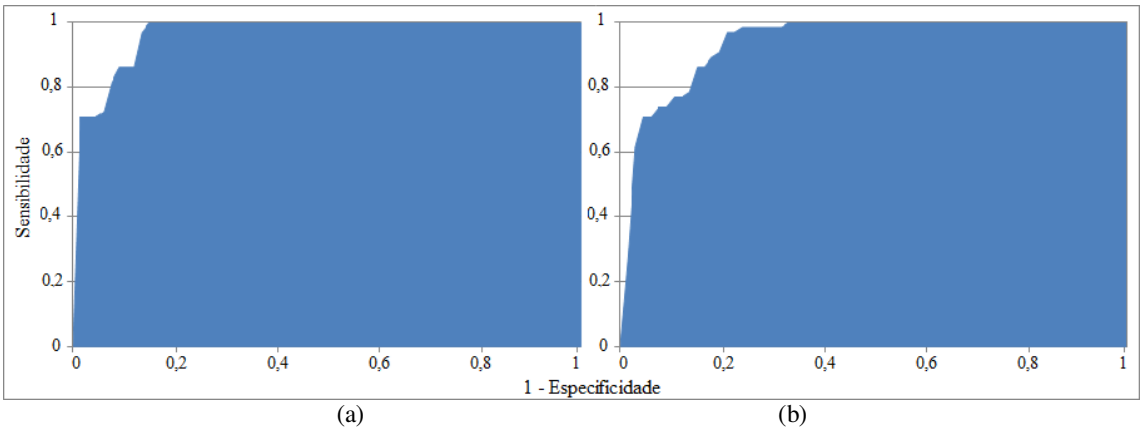
Medidas	SDM3		SDM4	
	M1	M2	M1	M2
Acurácia	87,4%	93,1%	92,1%	80,8%
Erro de Omissão	0%	0%	0%	0%
Erro de Comissão	25,2%	13,8%	15,8%	38,5%
CCM	0,773	0,870	0,853	0,667
AUC		0,973		0,946

Em relação às medidas de avaliação, apresentadas na Tabela 4.5, referentes aos testes da abordagem T2, mostram que há pouca variação na acurácia dos SDM5 e SDM6 quando avaliados pelo método M1 (88% e 88,4%). Quando os SDM são avaliados pelo método M2, a variação na acurácia mostra-se acentuada (84,6% e 93,1%), assim como a variação no índice CCM (0,728 e 0,870).

Tabela 4.5 – Comparação entre dois SDM gerados com a abordagem T2

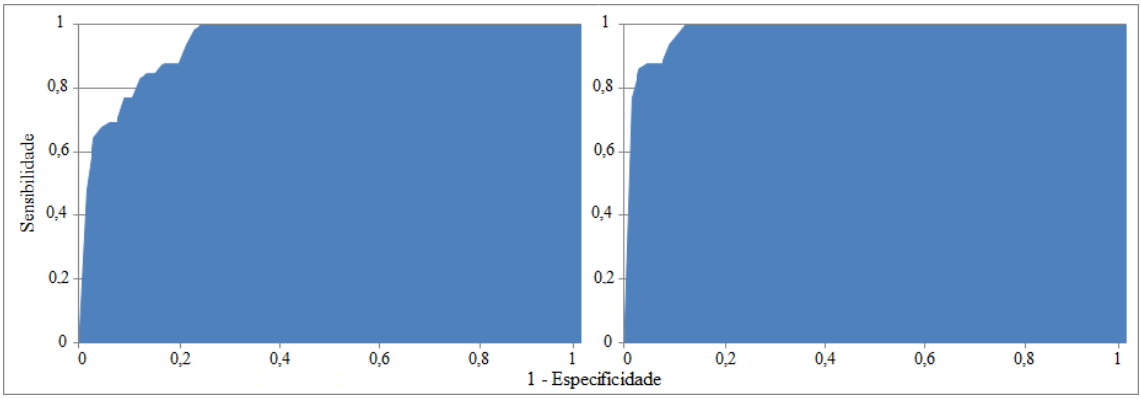
Medidas	SDM5		SDM6	
	M1	M2	M1	M2
Acurácia	88,0%	84,6%	88,4%	93,1%
Erro de Omissão	0%	0%	0%	0%
Erro de Comissão	24,1%	30,8%	23,2%	13,8%
CCM	0,782	0,728	0,790	0,870
AUC		0,951		0,987

Observando esses resultados pode-se verificar a alta variação nas medidas quando os modelos são avaliados pelo método M2, independente da abordagem utilizada. Isto acontece porque este método de avaliação emprega um novo conjunto de pontos de pseudoausência o que contribui para as elevadas taxas de erros de comissão calculadas para o SDM4 e SDM5. Na figura 4.8 e 4.9 podemos observar essas diferenças na forma da curva ROC.



(a) (b)

Figura 4.8 – Curvas ROC dos SDM3 (a) e SDM4 (b) ajustados com a abordagem T1



(a) (b)

Figura 4.9 – Curvas ROC dos SDM5 (a) e SDM6 (b) ajustados com a abordagem T2

Com os testes realizados pode-se observar que a abordagem T2, com o método de avaliação M1, apresentou os resultados mais estáveis. Para verificar esta afirmativa foram gerados outros 4 modelos (SDM5 a SDM8) utilizando novamente os mesmos parâmetros algorítmicos apresentados no inicio do capítulo, cujas medidas de avaliação são apresentadas na Tabela 4.6.

Tabela 4.6 – Comparação entre SDM gerados com a abordagem T2 e método M1

Medidas	**M1**				
	SDM2	**SDM5**	**SDM6**	**SDM7**	**SDM8**
Acurácia	87,9%	88,0%	88,4%	89,1%	88,3%
Erro de Omissão	0,0%	0,0%	0,0%	0,0%	0,0%
Erro de Comissão	24,2%	24,1%	23,2%	21,8%	23,3%
CCM	0,781	0,782	0,790	0,801	0,788

Com isso pode-se afirmar que, a partir dos estudos realizados, ajustar SDM utilizando a abordagem T2 e o método de avaliação M1 resultam em SDM com medidas mais estáveis. A acurácia variou entre 87,9% e 89,1%, porém isso não quer dizer que estes modelos são os mais precisos, mas que são SDM mais úteis e confiáveis, pois a aleatorização de diversos conjuntos de pontos de pseudoausência conferem diversidade aos dados de entrada, contribuindo para a representatividade e independência dos dados amostrais.

4.2 Comparações com outros sistemas

De acordo com resultados apresentados anteriormente, os SDM mais confiáveis foram aqueles que utilizaram a abordagem T2 para geração dos modelos e o método de avaliação M1. Estes SDM foram comparados com os SDM ajustados pelo sistema SAHGA SDM, sem o método *best-subset*, e com SDM ajustados pelo algoritmo GARP *best-subset*.

O sistema GARP - *Genetic Algorithm for Rule-set Prediction*, que é um algoritmo utilizado na geração de modelos de distribuição de espécies, mais precisamente na predição da distribuição potencial de espécies. Este algoritmo não considera os relacionamentos espaciais e, consequentemente, a dependência espacial; ele ajusta modelos e realiza predições observando apenas os valores pontuais das amostras (Santa Catarina, 2009). O algoritmo também está disponível no software openModeller Desktop v1.0.6 (CRIA et al., 2008).

4.2.1 Comparação com o SAHGA SDM sem o método *best-subset*

Para efetuar a comparação do método foram gerados dois SDM (SDM9 e SDM10) com o sistema SAHGA, sem o método *best-subset*, utilizando o conjunto de parâmetros algorítmicos apresentados na Tabela 4.1. Estes SDM são visualizados na Figura 4.10.

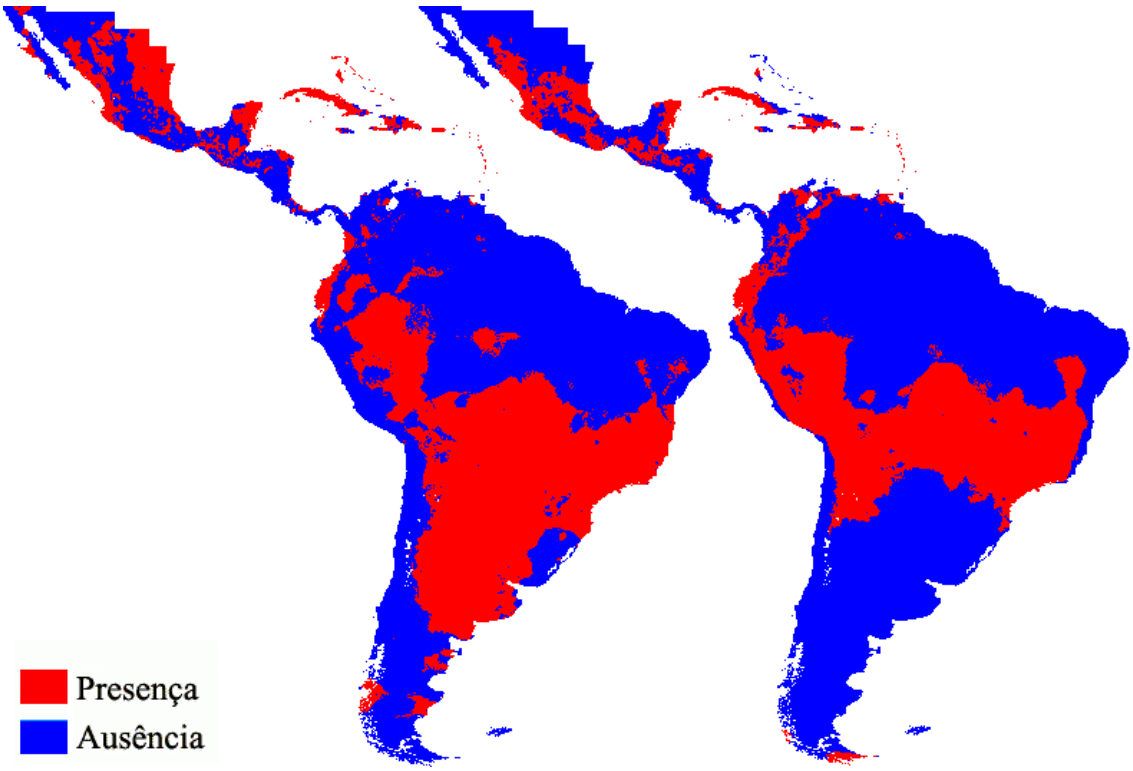


Figura 4.10 – SDM9 e SDM10 gerados pelo sistema SAHGA SDM, sem aplicação do método *best-subset;*

Observando-se esta figura pode-se concluir que os SDM9 e SDM10, ajustados com os mesmos parâmetros e diferentes conjuntos de pontos de pseudoausência, geraram resultados bem distintos.

Esta variação nos SDM reflete-se nas medidas de avaliação, apresentadas na Tabela 4.7. Os modelos SDM9 e SDM10 apresentaram alta variação na acurácia, com índices iguais a 74,2% e 98,5%, respectivamente. Analisando os outros índices da tabela, observa-se também a

variação de todas as taxas entre uma execução e outra do sistema SAHGA SDM sem o *best-subset*.

Tabela 4.7 – Medidas de avaliação para os SDM gerados pelo sistema SAHGA SDM, sem o método *best-subset*

Medidas	SDM9	SDM10
Acurácia	74,2%	98,5%
Erro de Omissão	6,1%	0%
Erro de Comissão	45,5%	3,0%
CCM	0,528	0,970
AUC	0,863	0,995

A seguir temos a Figura 4.11, onde é possível fazer a comparação do SDM gerado usando o SAHGA SDM com o *best-subset* com o SAHGA SDM sem o *best-subset*, mostrando que a representatividade do SDM com o *best-subset* ficou melhor por possuir a área de transação de onde é presença (vermelho) para área onde é ausência(azul).

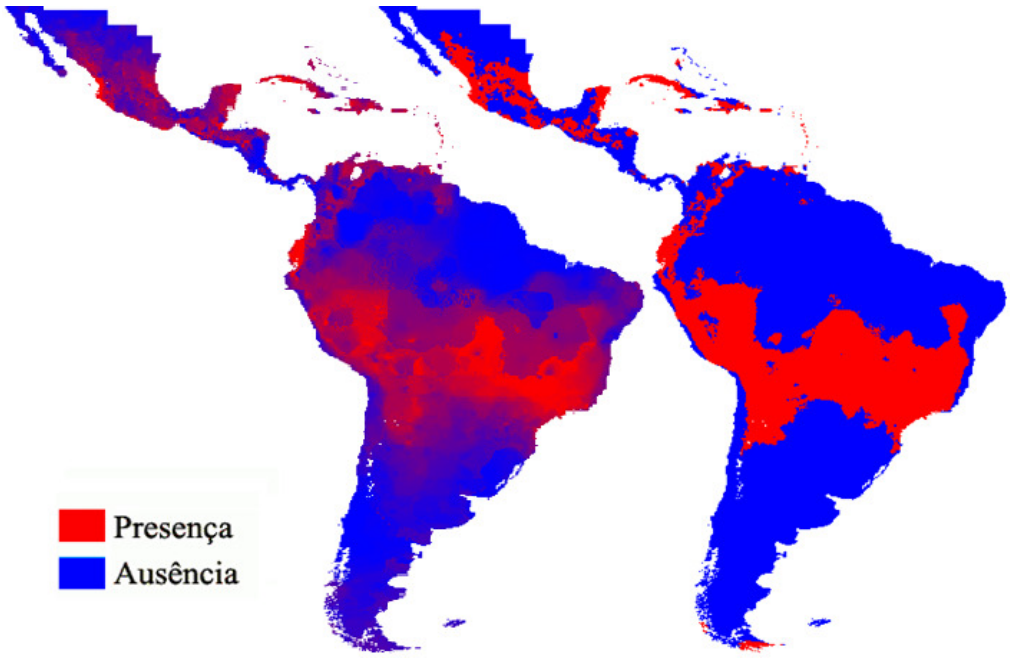


Figura 4.11 – Comparação entre os SDM gerados pelo SAHGA SDM com e sem o *best-subset*.

4.2.2 Comparação com o GARP *best-subset*

Um dos motivos para comparar os resultados do SAHGA SDM *best-subset* com o GARP *best-subset* está no fato de que ambos utilizam AG em seus núcleos de otimização. Os parâmetros utilizados para executar o GARP *best-subset* foram os apresentados na tabela 3.1.

Na figura 4.12 a seguir mostra dois SDM gerados usando o algoritmo do GARP – *best-subset*.

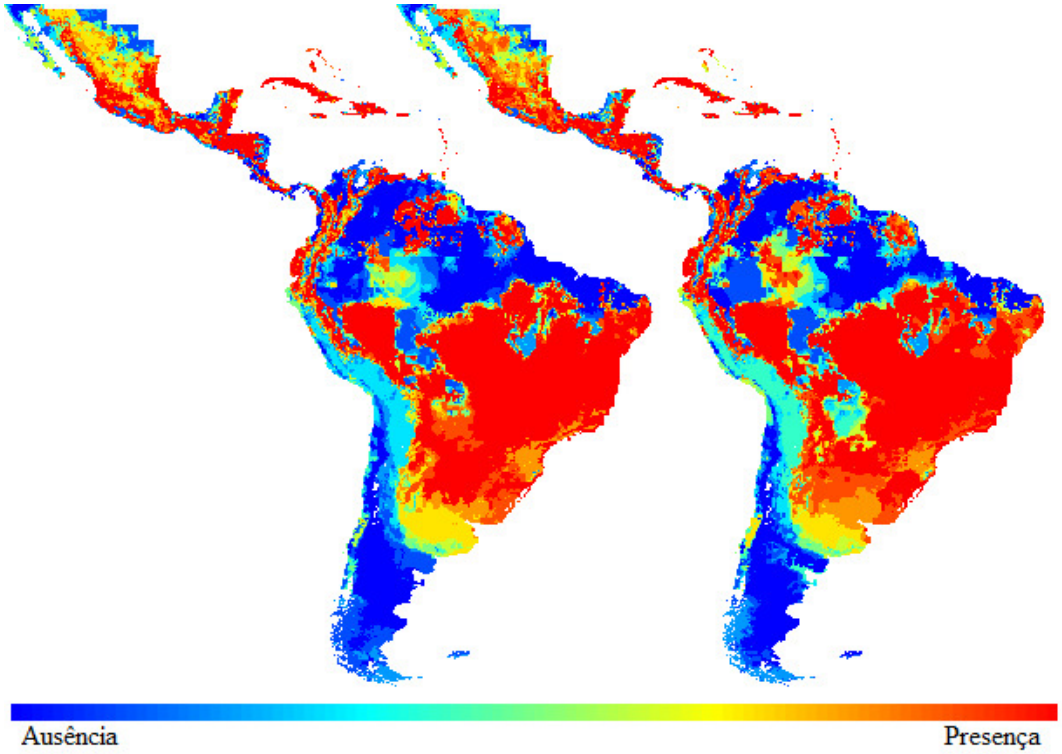


Figura 4.12 – SDM11 e SDM12 ajustados com o algoritmo GARP – *best-subset;*

Analisando a Figura 4.12 observa-se os modelos SDM11 e SDM12 apresentam algumas diferenças, principalmente nas áreas de transição entre áreas de presença e ausência. Porém, analisando as medidas de avaliação (Tabela 4.8) as diferenças são mínimas.

Tabela 4.8 – Medidas de avaliação para os SDM gerados pelo sistema GARP *best-subset*

Medidas	**SDM11**	**SDM12**
Acurácia	86,2%	86,2%
Erro de Omissão	0%	0%
Erro de Comissão	27,7%	26,2%
CCM	0,753	0,746

4.2.3 Discussão dos resultados

A comparação visual dos SDM gerados pelo GARP *best-subset* e pelo SAHGA SDM *best-subset* fica comprometida, pois os mapas de saída utilizam representações nas escalas de cores distintas. Entretanto pode-se compará-los através das medidas de avaliação calculadas para SDM ajustados com os mesmos dados de entrada.

Para facilitar esta comparação construiu-se a Tabela 4.9, com as medidas de avaliação de cinco SDM ajustados pelo SAHGA SDM *best-subset*, utilizando a abordagem T2 e o método de avaliação M1, ao lado das medidas de avaliação de quatro modelos ajustados pelo algoritmo GARP *best-subset*.

Tabela 4.9 – Medidas de avaliação para os SDM gerados pelo SAHGA SDM *best-subset* e GARP *best-subset*

Medidas	**SAHGA SDM *best-subset***					**GARP *best-subset***			
	SDM2	**SDM5**	**SDM6**	**SDM7**	**SDM8**	**SDM11**	**SDM12**	**SDM13**	**SDM14**
Acurácia	87,9%	88,0%	88,4%	89,1%	88,3%	86,2%	86,2%	89,2%	86,9%
Erro de Omissão	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%
Erro de Comissão	24,2%	24,1%	23,2%	21,8%	23,3%	27,7%	26,2%	21,5%	26,2%
CCM	0,781	0,782	0,790	0,801	0,788	0,753	0,746	0,803	0,765

Os dados apresentados na Tabela 4.9 indicam que o sistema SAHGA SDM *best-subset* foi capaz de ajustar SDM tão bons quanto aqueles ajustados pelo algoritmo GARP *best-subset.*

Capítulo 5

Conclusões

Neste trabalho foi implementado o método *best-subset* no sistema SAHGA SDM. Utilizando este sistema foram ajustados diferentes SDM para a espécie *thalurania furcata boliviana*. Estes SDM apresentaram boas medidas de avaliação e são mais confiáveis que os modelos gerados pelo SAHGA SDM sem o *best*-subset.

Para implementar o que foi proposto, foram criadas e comparadas duas abordagens, T1 e T2. Analisando os resultados dessas abordagens percebeu-se que a abordagem T1 é dependente da qualidade representativa dos pontos de pseudoausência gerados; se o algoritmo BIOCLIM gerar pontos ruins, os resultados baseados nesses pontos também serão ruins.

Já a abordagem T2 mostrou-se mais robusta. Isso não quer dizer que os modelos serão mais precisos, mas sim que serão SDM mais confiáveis, pois a aleatorização de diversos conjuntos de pontos de pseudoausência conferem diversidade aos dados de entrada, contribuindo para a representatividade e independência das amostras. Esta foi a abordagem escolhida para gerar os resultados e ser incorporada ao sistema SAHGA SDM *best-subset*.

Para analisar os SDM gerados com as abordagens T1 e T2, foram criados e comparados os dois métodos para o cálculo das medidas de desempenho: os métodos M1 e M2. O método M1 apresenta as medidas de desempenho com menor variação do que o método M2, ou seja, as medidas calculadas pelo método M1 são mais estáveis. O método M2 tem sua valia, pois permite analisar os SDM com outro conjunto de testes, visando uma avaliação independente; porém, os resultados tornaram-se instáveis, o que mostrou que o método M2 não é um método adequado para avaliação dos SDM ajustados. Conclui-se que o método M1 e a abordagem T2 formam o par adequado para gerar e testar os SDM ajustados pelo SAHGA SDM *best-subset*.

Na comparação dos SDM ajustados pelo SAHGA SDM *best-subset* com os SDM ajustados pelo SAHGA SDM sem *best-subset*, conclui-se que o método *best-subset* conferiu robustez aos SDM ajustados pelo primeiro sistema. As medidas de desempenho para os modelos

ajustados apresentaram pequenas diferenças, mas mantiveram acurácia sempre superior a 86%.

Em comparação com o sistema GARP *best-subset* os testes realizados mostraram medidas de desempenho similares. Conclui-se que, para os testes realizados, o sistema SAHGA SDM *best-subset* é um sistema tão bom quando o GARP *best-subset*. Entretanto, recomenda-se testar o sistema desenvolvido com outros conjuntos de dados, gerando SDM para outras espécies.

5.1 Trabalhos Futuros

Ao concluir esse trabalho percebeu-se que o sistema SAHGA SDM pode evoluir e ter algumas funcionalidades adicionadas. São desafios para trabalhos futuros:

1. Otimizar o código do SAHGA SDM pois, comparado ao GARP, sua execução é mais lenta; o SAHGA é cerca de 5 vezes mais lento que o GARP, utilizando o mesmo conjunto de dados e parâmetros para o método *best-subset*.

2. Incluir rotinas de pré-análise de dados para eliminar as variáveis com alto índice de correlação.

Referências Bibliográficas

AGUIAR, A. P. D. *et al*. Modeling spatial relations by generalized proximity matrices. In: **Brazilian Symposium on Geoinformatics**, Campos do Jordão – SP, v. 5, 2003. Anais eletrônicos... São José dos Campos: INPE, Novembro 2003. Disponível em: <http://www.geoinfo.info/geoinfo2003/papers/geoinfo2003-11.pdf>. Acesso em: 22/07/2013.

ANDERSON, R. P.; LEW, D.; PETERSON, A. T. Evaluating predictive models of species' distributions: criteria for selecting optimal models. **Ecological Modelling**, v. 162, n. 3, p. 211-232, Abril 2003.

AUSTIN, M. P. Spatial prediction of species distributions: an interface between ecological theory and statistical modelling. **Ecological Modelling**, v. 157, n. 2. p. 101-118, Novembro 2002.

BALDI, P. *et al*. Assessing the accuracy of prediction algorithms for classification: an overview. **Bioinformatics**, v. 16, n. 5, p. 412-424, Maio 2000.

BEAUMONT, L. J.; HUGHES, L.; POULSEN, M. Predicting species distributions: use of climatic parameters in BIOCLIM and its impact on predictions of species' current and future distributions. **Ecological Modelling,** v. 186, n. 2, p. 251-270, Agosto 2005.

BRAGA, A. C. S. **Curvas ROC: aspectos funcionais e aplicações**. Tese (Doutorado em Engenharia de Produção e Sistemas) - Universidade do Minho, Braga, 2000.

BUSBY, J.R. BIOCLIM – A bioclimatic analysis and predictive system. In **Nature Conservation: Cost Effective Biological Surveys and Data Analysis**. Margules, C. R. and M. P. Austin, editors. (eds.). Melbourne, pp. 64–68, CSIRO. 1991.

CASSEMIRO, F. A. S., GOUVEIA, S. F. & DINIZ-FILHO, J. A. F. Distribuição de Rhinella granulosa: integrando envelopes bioclimáticos e respostas ecofisiológicas **Revista da Biologia**, v. 8, p. 38–44. 2012.

CRIA, Centro de Referência em Informação Ambiental; POLI-USP, Escola Politécnica da USP; INPE, Instituto Nacional de Pesquisas Espaciais. **OpenModeller**. 2008. Disponível em: <http://openmodeller.sourceforge.net/>. Acesso em: 22/07/2013.

DAUBENMIRRE, R. Plant Communities: a textbook of plant synecology. **Harper & Row**, New York, 1968. 300 p.

DELEO, J. M. Receiver operating characteristic laboratory (ROCLAB): software for developing decision strategies that account for uncertainty. In: **Proceedings of the Second International symposium on Uncertainty Modelling and Analysis**, Los Alamitos, California: IEEE Computer Society Press, p. 318–325. 1993.

ELITH, J., GRAHAM, C.H., and NCEAS Modeling Group. Novel methods improve prediction of species' distributions from occurrence data. **Ecography**, v.29, n. 2, p. 129-151, Março 2006.

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, Abril 2004.

FIELDING, A. H.; BELL, J. F. A review of methods for the assessment of prediction errors in conservation presence/absence models. **Environmental Conservation**, v. 24, n. 01, p. 38-49, Março 1997.

FINAMORE, P. P. **Avaliação de Três Métodos para Geração de Pontos de Pseudo-Ausência Sobre a Qualidade dos Modelos de Distribuição de Espécies Ajustados pelo Sistema SAHGA SDM**. Monografia (Graduação em Ciência da Computação) – Universidade Estadual do Oeste do Paraná (UNIOESTE), Cascavel, PR, 2010.

GUISAN, A.; THUILLER, W. Predicting species distribution: offering more than simple habitat models. **Ecology Letters**, v. 8, n. 9, p. 993-1009, Setembro 2005.

GUISAN, A.; ZIMMERMANN, N. E. Predictive habitat distribution models in ecology. **Ecological Modelling**, v. 135, n. 2-3, p. 147-186, Dezembro 2000.

IWASHITA, F. **Sensibilidade de modelos de distribuição de espécies a erros de posicionamento de dados de coleta**. Dissertação (Mestrado em Sensoriamento Remoto) - Instituto Nacional de Pesquisas Espaciais (INPE), São José dos Campos, SP, 2007.

LEVINS, R., The strategy of model building in population ecology. **American Science**, v. 54, p. 421–431. 1966.

MATTHEWS, B. W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. **Biochimica et Biophysica Acta – Protein Structure**, v. 405, n. 2, p. 442-451, Outubro 1975.

MEYER, E. M. Ecological niche modelling: inter-model variation, best-subset models selection. In: **Workshop on Biodiversity Data Modelling**, Cidade do México, 2005. Anais eletrônicos... Copenhague: GBIF, Abril 2005. Disponível em: <http://www.gbif.org/prog/ocb/modeling_workshop/bangalore/presentations/ENMIV>.Acesso em: 22/07/2013.

NIX, H. A. A biogeographic analysis of Australian elapid snakes. In: **LONGMORE, R. (Ed.) Atlas of elapid snakes of Australia**. Canberra: Australian government publishing service, v.7, p. 4–15, 1986 (Australian flora and fauna series).

OPENSHAW, S.; ABRAHART, R. J. GeoComputation. In: **CRC Press**, London, 2000. 413 p.

OPENSHAW, S.; OPENSHAW, C. Artificial intelligence in geography. In: **John Wiley & Sons,** West Sussex, 1997. 348 p.

PEARSON, R. G.; RAXWORTHY, C. J.; NAKAMURA, M. & PETERSON, A. T. Predicting species distributions from small numbers of occurrence records: a test case using cryptic geckos in Madagascar. **Journal of Biogeography**, v. 34, p. 102-117, 2007.

PEDROSA, B. M. **Ambiente computacional para modelagem dinâmica**. Tese (Doutorado em Computação Aplicada) - Instituto Nacional de Pesquisas Espaciais, São José dos Campos, SP, 2003.

PHILLIPS, S. J.; ANDERSON, R. P.; SCHAPIRE, R. E. Maximum entropy modeling of species geographic distributions. **Ecological Modelling**, Amsterdam, v. 190, n. 3-4, p. 231-259, Janeiro 2006.

SANTA CATARINA, A. *SAHGA* – **Um algoritmo genético híbrido com representação explícita de relacionamentos espaciais para análise de dados geoespaciais**. Tese (Doutorado em Computação Aplicada) - Instituto Nacional de Pesquisas Espaciais (INPE), São José dos Campos, SP, 2009.

SEGURADO, P.; ARAÚJO, M. B. An evaluation of methods for modelling species distributions. **Journal of Biogeography**, v. 31, n. 10, p. 1555-1568, Setembro 2004.

SIQUEIRA, M. F. **Uso de modelagem de nicho fundamental na avaliação do padrão de distribuição geográfica de espécies vegetais**. Tese (Doutorado em Ciências de Engenharia Ambiental) - Universidade de São Paulo (USP), São Carlos, SP, 2005.

SOLORZANO, A. **Análise Fitogeográfica do cerradão: conexões florísticas, padrões estruturais, relações ecológicas e modelagem de sua distribuição potencial.** Tese (Doutorado em Ecologia), Universidade de Brasília, Brasília, 2011.

VIVO, M. D.; CARMIGNOTTO, A. P. Holocene vegetation change and the mammal faunas of South America and Africa. **Journal of Biogeography**, Oxford, v. 31, n. 6, p. 943-957, 2004.