MINISTÉRIO DA CIÊNCIA E TECNOLOGIA
INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS

INPE-15174-TDI/1291

SENSIBILIDADE DE MODELOS DE DISTRIBUIÇÃO DE ESPÉCIES A ERROS DE POSICIONAMENTO DE DADOS DE COLETA

Fábio Iwashita

Dissertação de Mestrado do Curso de Pós-Graduação em .Sensoriamento Remoto, orientada pelos Drs. Silvana Amaral Kampel e Antonio Miguel Vieira Monteiro, aprovada em 30 de março de 2007.

INPE
São José dos Campos
2008

Publicado por: **esta página é responsabilidade do SID**

Instituto Nacional de Pesquisas Espaciais (INPE)
Gabinete do Diretor – (GB)
Serviço de Informação e Documentação (SID)
Caixa Postal 515 – CEP 12.245-970
São José dos Campos – SP – Brasil
Tel.: (012) 3945-6911
Fax: (012) 3945-6919
E-mail: pubtc@sid.inpe.br

Solicita-se intercâmbio
We ask for exchange

MINISTÉRIO DA CIÊNCIA E TECNOLOGIA
INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS

INPE-15174-TDI/1291

SENSIBILIDADE DE MODELOS DE DISTRIBUIÇÃO DE ESPÉCIES A ERROS DE POSICIONAMENTO DE DADOS DE COLETA

Fábio Iwashita

Dissertação de Mestrado do Curso de Pós-Graduação em .Sensoriamento Remoto, orientada pelos Drs. Silvana Amaral Kampel e Antonio Miguel Vieira Monteiro, aprovada em 30 de março de 2007.

INPE
São José dos Campos
2008

528.711.7

Iwashita, F.
Sensibilidade de modelos de distribuição de espécies a erros de posicionamento de dados de coleta / Fábio Iwashita. - São José dos Campos: INPE, 2007.
100 p. ; (INPE-15174-TDI/1291)

1. Modelos de distribuição de espécies. 2. GARP. 3. MAXENT. 4. Erros de posicionamento. 5. Simulação de nicho potencial. I. Título.

Aprovado (a) pela Banca Examinadora em cumprimento ao requisito exigido para obtenção do Título de Mestre em

Sensoriamento Remoto

Dr. Dalton de Morisson Valeriano

Presidente / INPE / SJCampos - SP

Dr. Antonio Miguel Vieira Monteiro

Orientador(a) / INPE / SJCampos - SP

Dra. Silvana Amaral Kampel

Orientador(a) / INPE / SJCampos - SP

Dr. Gilberto Câmara

Membro da Banca / INPE / SJCampos - SP

Dra. Marinez Ferreira de Siqueira

Convidado(a) / CRIA / Barão Geraldo - SP

Aluno (a): Fábio Iwashita

São José dos Campos, 30 de Março de 2007

What the hammer? What the chain?
In what furnace was thy brain?
What the anvil? What dread grasp
Dared its deadly terrors clasp?

William Blake

A minha família dedico

AGRADECIMENTOS

Agradeço

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), pelo auxílio financeiro durante dois anos.

Aos meus orientadores, a Dra. Silvana Amaral Kampel e o Dr. Antônio Miguel Vieira Monteiro pela orientação, dedicação e paciência comigo.

Aos professores Dr. Gilberto Câmara, Dr. Dalton de Morisson Valeriano e Dra. Marinez Ferreira de Siqueira pela contribuição neste trabalho.

Ao Dr. Douglas Francisco Marcolino Gherardi pela atenção, paciência e ajuda durante o curso.

À Etel Rennó e Vera Gabriel da Silva Fontes pela amizade e apoio nos tropeços destes dois anos.

À Helen Borges e Luciana Moreira pela atenção, prestatividade e apoio técnico.

Aos meus pais pelo carinho e apoio durante todos os anos da minha vida, principalmente nas horas difíceis.

À minha querida Leila pelo seu companheirismo, compreensão, carinho e amor.

Aos amigos que fiz em São José dos Campos, Giselle, Polyana, Thais, “os *Forgottens*” e companheiros da Senzala II, Adair, Missae, Olga e Sérgio.

RESUMO

Os chamados modelos de distribuição de espécies utilizam dados de ocorrência de campo e variáveis ambientais para indicar locais adequados para a ocorrência de uma espécie. Apesar dos inúmeros trabalhos que avaliam os mais diversos aspectos dos modelos de distribuição de espécies, os erros posicionamento ainda não foram avaliados. Este trabalho avaliou a sensibilidade dos modelos de distribuição de espécies a erros de posicionamento de dados de coleta. Para que a avaliação dos modelos possa ser efetuada sobre um desenho experimental onde existe um número menor de fatores que podem influenciar o resultado, é preciso ter controle sobre a amostragem. No caso da avaliação da influência dos erros de posicionamento, também é necessário ter um controle dos diferentes tipos de erros de posicionamento. Para cumprir estes propósitos, os erros de posicionamento foram avaliados através de dados artificiais. Foram simulados o nicho fundamental e os pontos de ocorrência de uma espécie vegetal hipotética. Dois métodos de introdução de erros foram desenvolvidos e utilizados, a projeção das coordenadas das amostras para centróides de células e erros com distribuição normal com parâmetros em coordenadas polares. Os erros de posicionamento foram avaliados para os modelos BIOCLIM, GARP *Best Subsets* e MAXENT. Todos os modelos analisados apresentaram sensibilidade aos erros de posicionamento. O BIOCLIM apresentou a maior queda de desempenho. O GARP *Best Subsets* tem baixa sensibilidade a erros de posicionamento, mas prevê uma extensa área de ocorrência. O modelo máxima entropia apresentou a menor sensibilidade a erros. Estes resultados demonstram que a influência dos erros de posicionamento tem que ser considerada no processo de modelagem. São necessários cuidados específicos, como a escolha do método, suas premissas e o conhecimento sobre a precisão dos pontos de ocorrência.

EFFECTS OF SAMPLE PLACEMENT ERRORS ON ACCURACY OF SPECIES DISTRIBUTIONS MODELS

ABSTRACT

The study of environmental and its relationships with species spatial distribution is an old concern in biogeography. Mathematics allied with computers tools make possible a forecast of species distribution. The models know as Species Distribution Models uses occurrence field data and environmental variables to point out suitable places for species. Despite of many works that evaluate species distributions models performance, the placement accuracy influence over habitat suitability models remain unevaluated. We assess the models sensibility for sample placement errors. To keep an experimental design with few unknown factors, a control of sampling conditions are needed. Further, for placement errors analysis, a control over different error is necessary. To fulfill this purpose, placement errors were evaluate through artificial data. We simulated a fundamental niche for virtual plant specie and used a couple of error insertion methods; sample coordinates projection towards cellular center point and errors with normal distribution in polar coordinates parameters. We evaluated BIOCLIM, GARP Best Subsets and Maximum entropy. All models present placement errors sensibility. BIOCLIM exhibited the highest performance decrease. GARP Best Subsets had low sensibility to placement errors, nevertheless predicted a wide range occurrence. The maximum entropy presented the best performance despite of errors placement. These results show the importance to take account the effects of placement errors in modelling process. We need specific cautions, like method choice, its premises and sample placement accuracy.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS E ABREVIATURAS

AVHRR – *Advanced Very High Resolution Radiometer*

CRIA – Centro de Referência em Informação Ambiental

INPE – Instituto Nacional de Pesquisas Espaciais

GDM – *Generalized Dissimilarity Model*

GEOMA – Rede Temática de Pesquisa em Modelagem da Amazônia

GLM – *Generalized Linear Model*

GAM – *Generalized Additive Model*

GPS – Sistema de Posicionamento Global

MDE – Modelo Digital de Elevação

NNETW - *Neural Networks*

NOAA – *National Oceanic and Atmospheric Administration*

SDM – *Species Model Distribution*

SIG – Sistema de Informação Geográfica

SRTM – *Shuttle Radar Topographic Mission*

1 INTRODUÇÃO

As diferentes biotas das grandes massas continentais foram um dos primeiros aspectos da distribuição espacial de espécies estudadas pelos biogeógrafos do século XIX (Brown e Gibson, 1983). Exceto por poucas espécies cosmopolitas que podem ser encontradas em uma ampla variedade de ambientes, espécies vegetais costumam ocupar apenas uma fração dos possíveis habitats aos quais estão adaptadas para crescer e se reproduzir.

O fracasso na dispersão de sementes para ambientes favoráveis, a competição interespecífica, fatores históricos e barreiras biogeográficas são alguns dos determinantes na distribuição das espécies. Analisar e quantificar as relações destes fatores com a ocorrência de espécies representa um grande desafio para a biogeografia.

A modelagem matemática tem sido uma das principais ferramentas empregadas para investigar quantitativamente os fatores que influenciam os padrões de distribuição de espécies (Phillips *et al*., 2006; Rushton *et al*., 2004; Guisan e Zimmermann, 2000).

Entender, modelar e prever a ocorrência de espécies é essencial para os estudos de perda de biodiversidade (Polasky e Solow, 2001), na avaliação de risco ambiental ou sobre impactos de mudanças climáticas na biogeografia de espécies (Pearson et al., 2006). Existem modelos baseados em métodos quantitativos que relacionam as espécies e suas respostas ao meio ambiente, gerando previsões de locais adequados para a ocorrência de uma espécie alvo, determinando assim regiões em potencial para a conservação de espécies raras ou ameaçadas (Engler *et al*., 2004, Araújo e Williams, 2000) além de determinar os melhores locais para reintrodução de espécies (Hirzel *et al*., 2002).

A modelagem matemática aliada às ferramentas computacionais gera a possibilidade da previsão de ocorrência de espécies através da geração de superfícies temáticas, indicando presença ou ausência, com os chamados modelos de distribuição de espécies (*Species Distribution Model* – SDM). Tais modelos são empíricos, pois relacionam observações de campo com variáveis ambientais explicativas, fundamentadas em premissas estatísticas ou teóricas gerando os modelos de distribuição (Guisan e Thuiller, 2005).

As variáveis ambientais podem exercer efeitos diretos ou indiretos sobre as espécies, e devem ser escolhidas de modo a representar os principais fatores que influenciam as espécies (Austin, 2002). Por exemplo, a temperatura pode não influenciar diretamente a fisiologia de uma espécie de planta, mas tem correlação com a luminosidade, esta sim determinante na viabilidade deste indivíduo.

No nível de comunidade, a estrutura e composição das diferentes formações florestais são geralmente influenciadas pelos fatores físico-químicos ocorrentes (Daubenmirre, 1968), por exemplo, os nutrientes essenciais disponíveis presentes no solo. Entende-se o termo nutrientes essenciais para plantas como os elementos químicos sem os quais os indivíduos não conseguiriam completar seu ciclo de vida (gerar sementes viáveis) ou ainda, se estes nutrientes fazem parte de alguma molécula ou constituinte da planta que por si mesmo é essencial (Raven et al., 2001).

Existem fatores limitantes (ou reguladores), que são definidos como controladores da eco-fisiologia (temperatura, água, composição do solo, por exemplo), distúrbios que são qualquer perturbação que afete os sistemas ambientais, e recursos, que são os componentes assimilados pelos organismos (Guisan e Zimmermann, 2000). Incluir esses três fatores, limitações, distúrbios e recursos ao processo de modelagem ainda é um grande desafio a ser superado. Atualmente é quase nulo o emprego de variáveis ambientais com efeito direto sobre as espécies em estudo

(Araújo e Guisan, 2006), e esse é ainda um ponto crítico do processo de modelagem.

A coleta de dados ambientais para modelagem de distribuição de espécies sofreu uma grande mudança na década de 90 quando imagens de sensoriamento remoto se tornaram amplamente acessíveis. Adicionalmente, o crescimento do uso dos Sistemas de Informação Geográfica (SIG), para armazenar e trabalhar dados espaciais levaram a uma expansão no uso de SDMs. O sensoriamento remoto também tornou possível o estudo de áreas mais extensas e de locais de difícil acesso (Guisan e Thuiller, 2005; Rushton, *et al*., 2004), sendo uma fonte alternativa de dados para países de dimensão territorial como o Brasil.

O estudo e a caracterização de habitats através de imagens orbitais têm sido realizados apenas nos últimos anos e as variáveis potencialmente explicativas mais utilizadas até o momento foram climáticas e meteorológicas, topográficas e de uso e ocupação do solo (Hirzel *et al*., 2002; Zaniewski *et al*., 2002; Guisan *et al*., 1999).

Os dados de ocorrência e ausência de espécies utilizados para alimentar, calibrar e avaliar os SDMs raramente são coletados com esse objetivo, se configurando como um primeiro e talvez o maior empecilho da modelagem (Rushton *et al*., 2004). Embora existam grandes bancos de dados de coleções biológicas armazenados, eles geralmente foram adquiridos sem uma estratégia de coleta pré-definida (Stockwell e Peterson, 2002). Dados em museus e herbários apresentam frequentemente tendências (Reddy e Dávalos, 2003; Austin, 2002) ou imprecisão na localização do ponto de coleta e amostragem, muitas vezes indicando apenas proximidade a um ponto de referência, como uma vila ou um rio em uma escala de quilômetros ou mais (Engler *et al*., 2004).

Apenas recentemente o Sistema de Posicionamento Global (*Global Position System* – GPS) passou a ser empregado para coleta de dados biológicos, sendo o

método com maior acurácia para o posicionamento do ponto de coleta. Ainda assim, em florestas de dossel fechado, às vezes é necessário deslocar-se muitos metros do ponto de coleta para obter sinal do GPS.

Iniciativas recentes tornaram diversas bases de dados com ocorrência de espécies disponíveis na rede mundial de computadores. Apesar das coleções não estarem completamente informatizadas, alguns herbários e museus tem sua base de dados digital já disponível como o *Missouri botanical garden* (MBOT) e o *New York botanical garden* (NYBG) nos Estados Unidos e a *Comissíon nacional para el conocimiento y uso de la biodiversidad* (CONABIO) no México.

No Brasil, a Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) promoveu o projeto *speciesLink* coordenado pelo Centro de Referência em Informação Ambiental (CRIA), um sistema distribuído de informação que integra em tempo real coleções de herbários e museus. Contudo em sua maior parte, referem-se a dados primários de museus e herbários, que podem apresentar erros de posicionamento.

Neste trabalho entende-se por erro de posicionamento a diferença entre o ponto real de ocorrências de um indivíduo de uma determinada espécie e a posição registrada e disponível nos bancos de dados das coleções dos museus, herbários, etc.

Estão disponíveis na literatura trabalhos que avaliam o desempenho e a sensibilidade dos modelos de distribuição de espécies em função de diversos tipos de problemas, como tamanho de amostra (Stockwell e Peterson, 2002), estratégia de amostragem (Hirzel e Guisan, 2002) autocorrelação das variáveis ambientais (Segurado et al., 2006) e tendências (Reddy e Dávalos, 2003).

Chapman et al. (2005) discorreu sobre a qualidade das variáveis ambientais potencialmente preditivas, como estas são adquiridas, processadas e

empregadas. Dentre os problemas indicados destaca-se o relacionado à localização dos pontos de ocorrência, os autores empregaram SDM para prever a distribuição de *Rauvolfia nítida* (Apocynaceae) na ilhas do Caribe. Os erros de posicionamento fizeram alguns pontos de ocorrência cair fora da ilha, influenciando no resultado final.

Entretanto, a influência que erros de posicionamento dos dados têm sobre o desempenho dos SDM não foi avaliada. Assim, quão sensíveis são os modelos de distribuição de espécies aos erros de posicionamento?

Avaliar a influência dos erros sobre os modelos não é uma tarefa trivial. Por se tratarem de experimentos que dependem de coletas de campo, dois problemas comuns, mas graves, para o processo de modelagem podem ocorrer: vícios de coleta e número baixo de amostras (Hirzel e Guisan, 2002; Stockwell e Peterson, 2002). Este problema é abordado por Hirzel et al. (2001) e Austin et al. (2006) através de espécies virtuais. A posse do total conhecimento sobre a distribuição espacial da “espécie”, e sobre as variáveis que a determinam, torna o experimento mais controlável, pois há domínio sobre um maior número de variáveis.

O objetivo deste trabalho é avaliar as respostas dos modelos de distribuição de espécies a erros de posicionamento dos dados de coleta de campo, trabalhando com espécies simuladas. Alguns métodos de modelagem são comparados a fim de identificar a sensibilidade de cada método de modelagem aos erros de localização dos pontos de ocorrência da espécie.

Este trabalho está inserido em dois projetos institucionais: a Rede Temática de Pesquisa em Modelagem da Amazônia (GEOMA) e o *openModeller*/CRIA/INPE dos quais a Coordenação Geral de observação da Terra (OBT-INPE) é parte integrante.

O projeto GEOMA tem por objetivo desenvolver modelos computacionais capazes de predizer a dinâmica dos sistemas ecológicos e sócio-econômicos em diferentes escalas geográficas. O GEOMA pretende auxiliar a tomada de decisão nos níveis local, regional e nacional, ao fornecer ferramentas de simulação e modelagem, contribuindo na formação de recursos humanos.

O *openModeller* é um ambiente computacional multi-plataforma desenhado para modelagem de distribuição espacial de nicho fundamental em desenvolvimento coordenado pela equipe do CRIA. Junto com o CRIA e a Escola Politécnica da USP (POLI), o INPE participa do projeto *openModeller*, atualmente financiado pela Fapesp, desenvolvendo ferramentas para o arcabouço computacional, integração com SIG e na avaliação dos modelos para diferentes áreas de estudo.

Nos capítulos a seguir apresentam-se: fundamentação teórica, metodologia, avaliação dos modelos de distribuição de espécies e conclusões.

A fundamentação trata de alguns conceitos de biogeografia utilizados neste trabalho e como estão relacionados com a modelagem de distribuição de espécies. Para o processo de modelagem é realizada uma revisão mais minuciosa incluindo conceitos, exemplos, métodos de avaliação e problemas comumente encontrados no processo de modelagem da distribuição de espécies.

A metodologia se apóia na fundamentação teórica para a escolha dos modelos a serem analisados e nos métodos aplicados para a avaliação. Neste capítulo também é detalhada a simulação da espécie e dos erros de posicionamento.

O capítulo 4 contém os resultados dos modelos gerados a partir das espécies simuladas bem como as respectivas avaliações de sensibilidade destes modelos a erros de posicionamento. O último capítulo apresenta as conclusões acerca dos modelos e traz algumas recomendações com base na apreciação dos resultados deste trabalho.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 A Distribuição espacial de espécies

Com exceção de algumas poucas espécies cosmopolitas, a maioria dos seres vivos possui um padrão de distribuição espacial limitado. Entretanto, esses padrões não são imutáveis, em uma escala geológica, as distribuições podem sofrer grandes mudanças em respostas a variações ambientais (Vivo e Carmignotto, 2004), como mudanças climáticas globais, por exemplo.

Todos os organismos têm alguma capacidade de se mover do seu local de nascimento para novos locais. Plantas superiores e alguns animais aquáticos são sésseis quando adultos, mas nos seus primeiros estágios de desenvolvimento são geralmente capazes de viajar pequenas, mas significantes distâncias a partir da sua fonte (Brown e Gibson, 1983). Esse deslocamento favorece as espécies porque geralmente locais diferentes são mais favoráveis ao desenvolvimento do indivíduo, em parte devido à competição intra-específica, e em parte porque a qualidade do ambiente natal está sempre mudando (Daubenmirre, 1968).

Embora dispersões ocorram continuamente, a maioria não resulta em modificação em suas respectivas distribuições geográficas. A distribuição espacial de espécies está, na maioria dos casos, condicionada a fatores ambientais limitantes que permanecem relativamente constantes no tempo em uma escala ecológica (Bazzaz, 1998).

A biogeografia é a ciência que estuda a distribuição geográfica dos seres vivos, procurando entender os padrões de organização e dispersão espacial e os processos que produzem tais padrões. Os biogeógrafos geralmente distinguem dois tipos de eventos de dispersão. Em alguns casos a espécie cruza com sucesso uma barreira, como um oceano ou uma cadeia de montanhas, e estabelece uma população do outro lado da barreira, em outras situações as

espécies podem simplesmente expandir sua distribuição geográfica (Brown e Gibson, 1983). Assim, através da dispersão, a espécie “procura” locais com condições ambientais favoráveis ao seu desenvolvimento, i.e. por um habitat.

2.1.1 Habitat e nicho ecológico

Habitat é definido como o espaço físico em que uma determinada população de uma espécie vive e se desenvolve. Este conceito é aplicável desde o nível de espécies até o de população, pois duas populações de uma mesma espécie podem viver em habitats com condições ambientais distintas devido a sua plasticidade fenotípica (Raven et al., 2001).

O primeiro conceito de nicho é similar ao do habitat e foi cunhado por Grinnell (1917) que o definiu simplesmente como locais onde os requisitos para uma determinada espécie viver e se reproduzir são preenchidos. Este conceito justifica estudos sobre nichos embasados na distribuição espacial de nutrientes, pois segundo a Lei do “Fator Mínimo” de Liebig, a substância mineral em menor concentração relativa determina o limite para o crescimento e rendimento das plantas (Larcher, 2000).

O termo nicho ecológico se popularizou através de Hutchinson (1957), que o modelou como um hipervolume n-dimensional onde cada dimensão representa o intervalo de condições ambientais ou de recursos necessários para a espécie. Por exemplo, o nicho de uma espécie vegetal poderia incluir o intervalo de temperatura tolerado, a intensidade luminosa necessária para fotossíntese, regimes de umidade e quantidade mínima de nutrientes essenciais presentes no solo.

Para fins de modelagem da distribuição de espécies ressalta-se a diferença entre nicho fundamental e o nicho realizado. O nicho fundamental de espécies inclui os intervalos das condições ambientais necessárias para a existência da espécie sem

considerar a influência de competição interespecífica ou de predação por outras espécies. O nicho realizado descreve a parte do nicho fundamental onde realmente a espécie ocorre. Desse modo, á área definida pelo nicho fundamental é sempre maior que o nicho realizado.

Diversos trabalhos enfatizam a importância em esclarecer o objeto da modelagem, habitat, nicho fundamental, nicho realizado ou apenas distribuição espacial potencial (Araújo e Guisan, 2006; Austin et al., 2006; Rushton et al., 2004; Austin, 2002; Okasanen e Minchin, 2002). Habitat é um dos termos mais escolhidos na definição dessa premissa da modelagem, provavelmente pela amplitude de sua definição.

Segundo Bazzaz (1998) as teorias de seleção de habitats e sua relação com o forrageio, composição e coexistência de espécies são em sua maioria baseadas em animais que se locomovem, e não são diretamente aplicáveis às plantas. Assim, para todas as espécies vegetais, existem alguns fatores comuns na "seleção" de habitats:

1. Disponibilidade de recursos (luz, água, nutrientes, etc.) em quantidade suficiente e em quantidades balanceadas para o crescimento e reprodução.

2. Presença de polinizadores, dispersores e simbiontes.

3. Ausência relativa de herbivoria, predadores e patógenos, com exceção daqueles que atacam os competidores.

Os conceitos de habitat e nicho são fundamentais para estudos de padrões de endemismo, provincialismo, e disjunção de distribuições geográficas. Estes padrões indicam que a dispersão atual de muitos grupos reflete a historia do local de origem, dispersão e extinção local destas espécies (Brown e Gibson, 1983).

Para entender os padrões atuais de distribuição da espécie alvo é importante estudar sua paleo-ecologia.

O termo endemismo significa que a espécie não ocorre em nenhum outro lugar. Organismos podem ser endêmicos em diversas escalas espaciais e em diferentes níveis taxonômicos. Endemismo autóctone é aquele em que a espécie se diferenciou *in situ,* no mesmo local onde é encontrada nos dias atuais, enquanto espécies de endemismo alóctone desenvolveram suas características em outros locais e apenas sobrevivem em uma área restrita (Brown e Gibson, 1983).

A distribuição espacial de uma espécie ou gênero têm uma relação com a linha evolutiva e distribuição dos táxons mais altos como famílias e ordens. Assim, táxons mais baixos tendem a ter uma distribuição espacial mais restrita que os táxons mais altos (Bazzaz, 1998).

Provincialismo é outro termo muito utilizado por biogeógrafos e pode ser definido como a ocorrência de determinadas espécies restritas a uma província biogeográfica (Brown e Gibson, 1983). Reinos, regiões, sub-regiões, províncias e distritos biogeográficos, dentro dessa hierarquia, províncias biogeográficas são áreas onde as características ambientais refletem uma relativa uniformidade na composição de espécies.

Em contrapartida, podem ocorrer também distribuições onde dois ou mais táxons relativamente similares ocupam áreas muito distantes entre si, este padrão de distribuição biogeográfica é denominado disjunto (Figura 2.1). Biogeógrafos dão uma atenção especial a esse tipo de distribuição na tentativa de reconstruir grandes e pequenos eventos na história da Terra, como grandes desastres naturais ou mudanças climáticas em períodos geológicos passados.

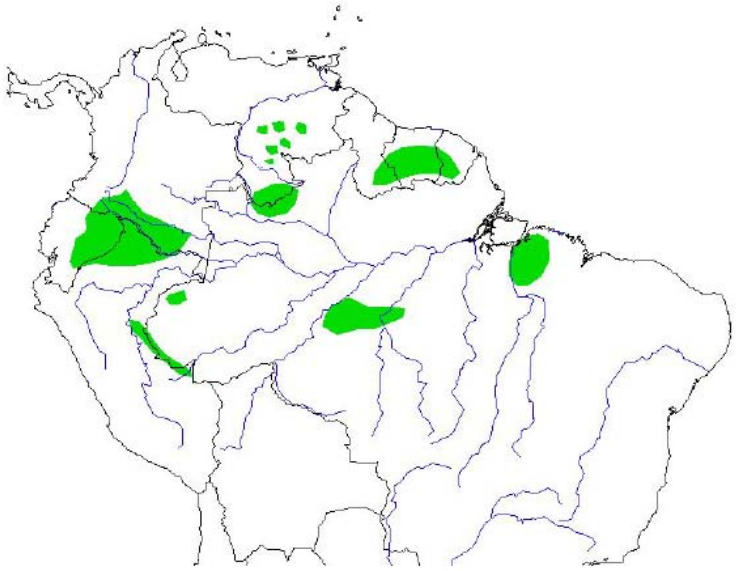

FIGURA 2.1 – Padrão de distribuição disjunto.
Fonte: Bonacorso et al. (2006).

Neste trabalho, como premissa para a modelagem, adota-se que os resultados dos modelos correspondem a uma previsão do nicho fundamental. Para tanto propõe-se estudar uma espécie vegetal virtual de distribuição contínua (não disjunta) típica da Amazônia oriental.

2.2 Modelos de distribuição de espécies

Segundo Guisan e Zimmermann (2000), existem três pilares no estudo de modelos matemáticos aplicados à ecologia: generalidade, realidade e precisão. Desses são derivados três grupos de modelos, onde em cada grupo dois desses aspectos devem ser enfocados em detrimento do terceiro (Figura 2.2).

O primeiro grupo foca a generalidade e a precisão, esses modelos são chamados de analíticos, como as equações de crescimento populacional logístico e as de Lotka-Volterra. O segundo grupo é desenvolvido visando ser realista e generalista, são chamados de mecanicistas, fisiológico, casuais ou modelos de processos, suas previsões são baseadas nas relações reais de causa e efeito. O terceiro

grupo sacrifica a generalidade pela precisão e realidade, são os chamados modelos empíricos, estatísticos ou fenomenológicos.

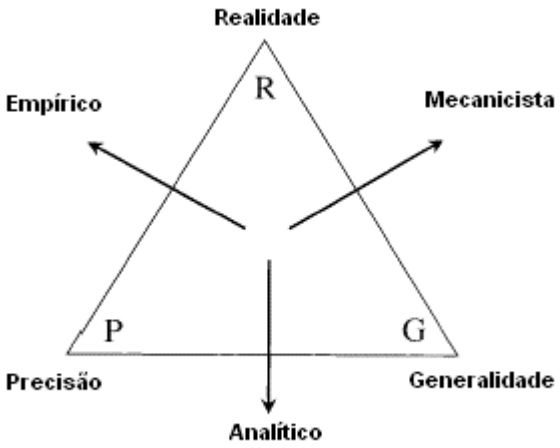

FIGURA 2.2 – Três tipos de modelos: Generalista, mecanicista e empírico. Os SDM geralmente são modelos empíricos.

Os modelos de distribuição de espécies (SDMs) são geralmente empíricos, pois são baseados em amostras de campo (realidade) e são aplicados especificamente para modelar a ocorrência de uma espécie em uma determinada área de estudo através de métodos estatísticos e/ou computacionais.

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).

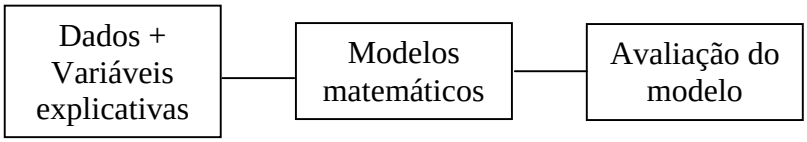

FIGURA 2.3 – Elementos essenciais na modelagem de distribuição de espécies.

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). 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 o habitat seja adequado; c) o habitat é realmente inadequado para a espécie (Phillips *et al*., 2006). Esse tipo de dado é particularmente precioso, porém escasso. Alguns autores vêm contornando esse problema utilizando dados de pseudo-ausência simulados para a modelagem (Engler *et al*., 2004).

2.2.1 Tipos de modelos de distribuição de espécies

Há uma grande variedade de técnicas de modelagem para explorar a relação entre a resposta (ocorrência de espécies) e as variáveis ambientais preditivas.

Elith *et al*. (2006) classificam os SDMs 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.

A Tabela 2.1 apresenta os diversos modelos e *softwares* disponíveis e ilustra o grande número de métodos de modelagem já disponíveis, entretanto foram selecionados apenas três modelos de diferentes categorias, para serem empregados neste trabalho, o Bioclim, o Maxent e o GARP. O texto segue descrevendo genericamente estas categorias de modelos detalhando os utilizados neste estudo.

Métodos de envelopes bioclimáticos – Segundo Guisan e Zimmermann (2000) o envelope bioclimático é um dos métodos mais antigos para modelagem de distribuição, e até recentemente muitos modelos de distribuição de vegetação foram baseados em técnica de envelopes ambientais. O Bioclim é um modelo que utiliza apenas a presença de espécies.

Os envelopes bioclimáticos predizem locais com condições climáticas favoráveis a uma espécie baseados no cálculo de um envelope retilíneo mínimo no espaço climático multidimensional. Um envelope retilíneo pode ser definido como uma árvore de classificação, que consiste em uma partição recursiva do espaço multidimensional definido por variáveis explicativas dentro de grupos que são os mais homogêneos possíveis em termos de sensibilidade (Carpenter et al., 1993).

Para cada variável ambiental o algoritmo calcula a média e o desvio padrão associados aos pontos de ocorrência. O envelope de cada variável é definido por um intervalo de confiança calculado sobre a média, o desvio padrão e o limite de corte estabelecido durante a escolha do limite de corte do desvio padrão. Por exemplo, se o intervalo de confiança é de 95%, o limite de corte deveria ser 1,96.

TABELA 2.1 – Modelos de distribuição de espécies divididos em três grupos: envelopes bioclimáticos, de presença e ausência e modelos de comunidade.

Tipo de modelo	Método	*Software*	Necessita ausência	Página de acesso
ENV	DOMAIN	DIVA-GIS	Não	www.cifor.cgiar.org/docs/_ref/research_tools/domain/index.htm
ENV	BIOCLIM	BIOCLIM/ OpenModeller	Não	www.arcscripts.esri.com http://openmodeller.sourceforge.net/
ENV	LIVES	-	Não	-
PRES	ENFA	BIOMAPPER	Não	www.unil.ch/biomapper
PRES	GLM	S-Plus, R	Sim*	http://www.insightful.com/products/splus/ http://www.r-project.org/
PRES	Máxima entropia	Maxent	Sim*	www.cs.princeton.edu/~schapire/maxent/
PRES	Boosted decision tree	R, GBM	Sim	-
PRES	AG	DK-GARP/ OM-GARP	Sim*	http://nhm.ku.edu/desktopgarp/index.html http://openmodeller.sourceforge.net/
PRES	GAM	S-Plus, R	Sim	http://www.insightful.com/products/splus/ http://www.r-project.org/
PRES	GDM-SS	GDM	Sim	-
COM	GDM	S-Plus	Sim	-
COM	MARS-COMM	R, MDA	Sim	-

* podem ser utilizados dados de pseudo-ausência. ENV, envelopes climáticos; PRES, SDM que utilizam dados de presença e variáveis explicativas; COM, modelos de comunidade. Fonte: Adaptado de Elith *et al*. (2006).

Métodos que utilizam dados de presença e supõem ausência

Esta categoria de modelos utiliza além dos dados de presença, as variáveis ambientais explicativas e os dados de ausência. Caso não exista disponibilidade de dados de ausência alguns modelos simulam ausência. Neste grupo se encaixam a maioria dos modelos empregados atualmente, *Boosted decision tree* (BRT), algoritmos genéticos (GARP), modelo linear aditivo (GAM), modelo linear generalizado (GLM), modelo de dissimilaridade generalizado para apenas uma espécie (GDS-SS), redes neurais (NNETW), *Ecological Niche Factor Analysis* (ENFA).

Dentre estes, o modelo linear generalizado (Guisan et al., 2002; Miller e Franklin, 2002; Guisan et al., 1999) é um dos métodos mais utilizados para modelagem. O GLM é uma extensão da regressão linear múltipla clássica que permite variáveis sem normalidade serem modeladas (Guisan et al., 2002). O GLM geralmente constitui uma escolha preferencial porque pode lidar com muitos tipos de variáveis explicativas (contínuo, binário, qualitativo, ordinal), mas por outro lado precisa também de dados de presença e ausência (Guisan et al., 1999). Para utilizá-lo sem disponibilidade de dados de ausência é possível gerar dados de pseudo-ausências para alimentar os modelos (Engler *et al*., 2004; Segurado e Araújo, 2004).

MAXENT – O modelo Maxent também é um modelo que utiliza somente dados de presença. De acordo com Phillips *et al*. (2006) a máxima entropia (Maxent) é um método para realizar previsões ou inferências a partir de informações incompletas. É aplicado em diversas áreas como astronomia, reconstrução de imagens, física estatística e processamento de sinal. A idéia da aplicação do Maxent para SDM é estimar a probabilidade de ocorrência da espécie encontrando a distribuição de probabilidade da máxima entropia, sujeita a um conjunto de restrições que representam a informação incompleta sobre a distribuição do alvo.

O princípio da máxima entropia é elaborar uma aproximação onde sejam respeitadas todas as restrições conhecidas acerca da distribuição da espécie. Esse princípio pode ser expresso da seguinte forma: denota-se uma distribuição desconhecida por π, sobre um conjunto finito X (que representa o conjunto de pixels da área de estudo). Os elementos de X são tomados como pontos x. Assim a distribuição π define uma distribuição de probabilidade não negativa π(x) para cada x, onde $\sum_{i=1}^{n} \pi(x) = 1$, sendo *n* o número de elementos em X.

Para modelagem é necessária uma aproximação para π, que é denotada como $\hat{\pi}$. Desse modo, a entropia de $\hat{\pi}$ é dada pela equação 2.1.

$$H(\hat{\pi}) = -\sum_{x \in X} \hat{\pi}(x) \ln \hat{\pi} \qquad (2.1)$$

A entropia é um conceito fundamental da teoria da informação originalmente proposto por *Shannon*, cujo índice também é utilizado para mensurar biodiversidade. A informação disponível sobre a distribuição da espécie constitui um conjunto de valores tomados como verdades de campo, chamados “feições”, e suas restrições são os valores esperados de cada feição que devem corresponder com as médias empíricas (valor médio para um conjunto de pontos tomados da distribuição do alvo).

GARP – O GARP (*Genetic Algorithm for Rule Set Production*) é um modelo que define o nicho fundamental através de um conjunto de regras que são selecionadas através de um algoritmo genético. O GARP opera sobre o conjunto de regras, realizando uma “seleção natural”, excluindo regras menos eficientes e criando novos conjuntos de regras a partir de “indivíduos” sobreviventes (Stockwell e Peters, 1999). Cada realização do GARP é uma possibilidade com componentes aleatórias onde cada realização é distinta, o resultado é um somatório dos resultados de várias realizações do mesmo modelo.

A essência do sistema é a capacidade de filtrar e lidar com diversos tipos de erros (Stockwell e Peters, 1999). O GARP é capaz de atingir 90% de acerto sobre um conjunto de 10 amostras (Stockwell e Peterson, 2002). Com 100 amostras o desempenho chega a superar o GLM, tido como mais robusto (Stockwell et al., 2005).

Há uma versão disponível com interface gráfica, o DK-GARP (*Desktop Garp*) e uma versão para o OpenModeller, o OM-GARP (*OpenModeller GARP*).

O GARP possui uma série de parâmetros que devem ser estabelecidos antes de rodar o modelo. Estes parâmetros são importantes porque regem as regras do algoritmo genético. Segue uma rápida descrição segundo Stockwell (2006).

- O limite de convergência (*convergence limit*) estabelece a condição para cessarem as interações dentro do algoritmo genético. Seus valores estão geralmente entre 0,01 e 0,1 e caso seja zerado, o algoritmo só vai parar quando o número de interações estabelecidas atingir o seu máximo;

- Número máximo de gerações (*max generations*) é o número máximo de interações e estabelece outra condição para a parada do algoritmo. Este parâmetro faz o algoritmo parar mesmo se o limite de convergência ainda não for atingido. Um número maior de interações produz resultados mais estáveis;

- Proporção de treinamento (*training proportion*) é a quantidade do total de amostras que será utilizada para treino e para teste;

- O número de modelos (*total runs*) gerados é utilizado para compor o conjunto *Best Subsets;*

- Máximo número de processos (*maximum number of threads*) é o número máximo de processos executados simultaneamente;

- Limite de omissão "*hard*" (*hard omission threshold*) é o número de modelos abaixo no limite de omissão que serão considerados no resultado (Figura 2.4);

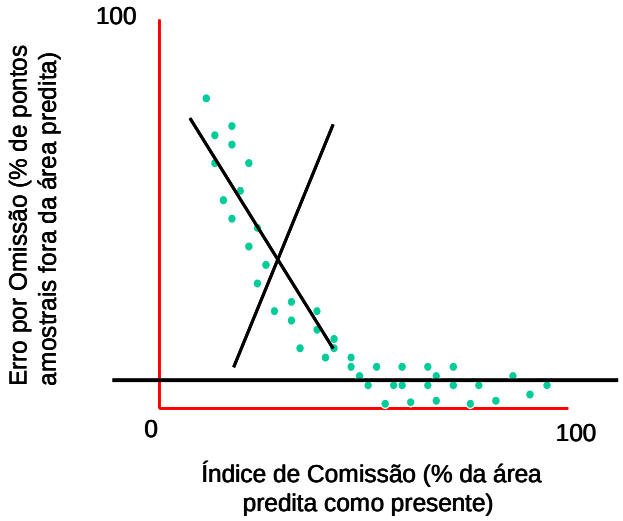

FIGURA 2.4 – Critério de limite de omissão *hard* para a seleção de melhores modelos no GARP-BS.

Fonte: Kansas applied remote sensing program (2005).

- Tamanho da população (*population size*) – é o número máximo de regras mantidas na solução.

- Tamanho da amostra de comissão (*commission sample size*) é o número de amostras utilizadas para calcular os erros de comissão;

- O limite de comissão (c*ommission threshold*) é outro critério para a seleção dos melhores modelos (Figura 2.5) onde, neste caso, os melhores estão próximos à mediana;

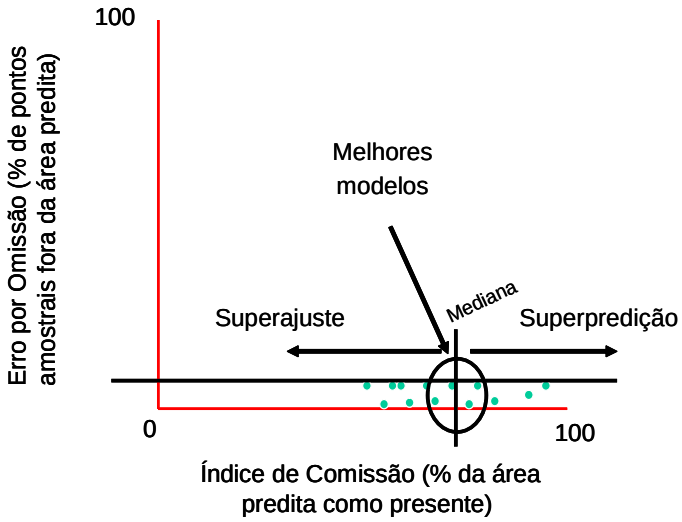

FIGURA 2.5 – Critério de limite de comissão para a seleção de melhores modelos no GARP-BS. Nesse caso o limite é a mediana.

Fonte: Kansas applied remote sensing program (2005).

Modelos baseados na comunidade

Apesar de apresentarem um grande potencial para incluir processos ecológicos como predação e competição à modelagem de distribuição de espécies, os modelos de comunidade não foram empregados neste trabalho e ainda são pouco utilizados.

Os modelos de comunidade também são chamados de modelos generalizados de dissimilaridade (GDM), e modelam o volume espacial na composição da comunidade entre pares de locais como uma função das diferenças ambientais entre estes locais. Este método combina elementos da matriz de regressão e dos GLM, permitindo respostas não lineares do ambiente que captura relações realistas ecologicamente entre a dissimilaridade e a distância ecológica (Elith *et al*., 2006). Pertencem a este grupo o próprio GDM e o *Multivariate Adaptative Regression Splines for Communities* (MARS-COMM).

2.2.2 Escala de estudo

Um problema presente na modelagem de distribuição está na identificação da escala apropriada para a amostragem. A escolha da escala de estudo envolve duas questões, a extensão da área de estudo e a resolução espacial dos planos de informação (Rushton *et al*., 2004).

Ao definir a resolução espacial deve-se levar em consideração que um *pixel* com resolução espacial baixa resulta em dados mais simples de tratar, mas em contrapartida, caso exista autocorrelação (Segurado et al., 2006), os dados não podem ser agregados em uma célula maior, pois não são independentes. Em contraste uma resolução espacial mais fina pode representar melhor alguns dos processos ecológicos (Engler *et al*., 2004, Collingham et al., 2000). Por isso a ocorrência de organismos sésseis, como plantas podem ser melhor inferidas com resoluções espaciais finas (Guisan e Thuiller, 2005; Collingham et al., 2000).

A escolha do tamanho da área de estudo também é crucial, pois padrões observados em uma escala podem não aparecer em outra, esta restrição pode levar a interpretações incorretas se apenas parte de um importante gradiente ambiental for amostrado (Guisan e Thuiller, 2005). Assim, para determinar o tamanho da área de estudo é preciso ter um conhecimento *a priori* dos gradientes que influenciam a distribuição da espécie em toda a área de estudo para que possam ser evitados problemas de autocorrelação (Segurado et al., 2006), caso o modelo pressuponha independência, por exemplo (Austin, 2002). Caso exista autocorrelação, é possível determinar o intervalo de amostragem com o estudo do variograma da variável com base no alcance do semivariograma experimental (Isaaks e Srivastava, 1989).

Dados ambientais com resolução espacial variável também representam uma dificuldade para a modelagem de distribuição de espécies. Em geral, os trabalhos com SDM adotam uma das seguintes estratégias: agregar os dados em grades

regulares com resolução espacial que seja compatível com a pior precisão encontrada nos dados coletados; interpolar os dados ou; descartar os dados com baixa resolução espacial.

Outro fator que deve influenciar na escolha da escala é a densidade populacional de indivíduos. A competição intra-específica pode diminuir a quantidade de ocorrências por área, e o número de amostras influencia no desempenho do modelo (Stockwell e Peterson, 2002). Por isso um balanço entre tamanho amostral e tamanho da área de estudo deve ser obtido. A Figura 2.6 ilustra que o número de ocorrências da espécie *Eryngium alpinumem* cai quando a resolução espacial aumenta (Engler *et al*., 2004).

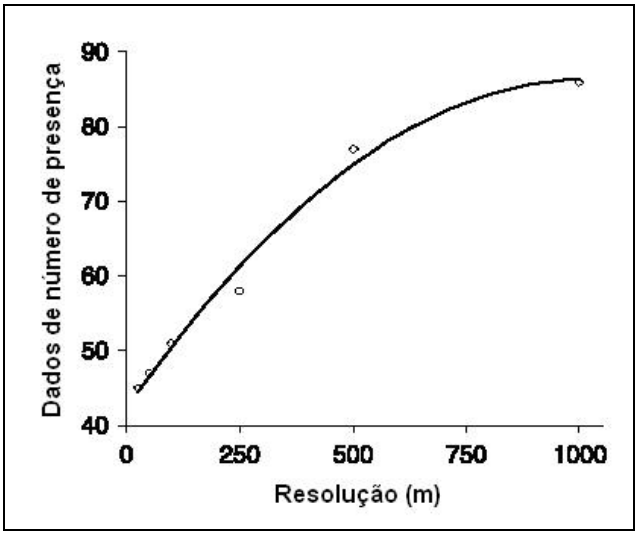

FIGURA 2.6 – Número de ocorrências de *Eryngium alpinumem* em diferentes resoluções.
Fonte: Engler *et al*. (2004).

O cenário se agrava quando ocorrem simultaneamente problemas na resolução espacial das variáveis e na extensão da área de estudo. Chapman et al. (2005) geraram um modelo de distribuição para *Rauvolfia nítida* (*Apocynaceae*) nas ilhas do Caribe, em uma situação onde há variáveis climáticas com resolução espacial baixa e uma área de estudo pequena. Desse modo, mesmo que exista um bom número de pontos de ocorrência, boa parte deles cairá na mesma célula.

Este problema também é constatado por Tobler et al. (2007) que verificou este problema de agregação de pontos de ocorrência em dados de herbários para espécies das famílias *Moraceae* e *Myristicaceae* na Amazônia peruana. A área de estudo foi coberta por 252 células de 0,5º, onde metade das ocorrências das 46 espécies da família *Myristicaceae* caíram em apenas 9 células, e metade dos pontos de ocorrência das 134 espécies da família *Moraceae* caíram em 6 células. Este trabalho demonstra a importância de coletas botânicas sistemáticas no tempo e nos espaço, com um desenho amostral bem definido para que seja possível a construção de bancos de dados robustos.

Como a modelagem é baseada em dados coletados em campo, e devido aos custos de coleta e a logística, muitos conjuntos de dados são pequenos ou são coletados de forma concentrada em pequenas áreas ou regiões de fácil acesso, como rios ou locais próximos a estradas (Barry e Elith, 2006). Em conseqüência disso, muitos dados disponíveis, mesmo que coletados com localização precisa, podem apresentar alguma tendência (Reddy e Dávalos, 2003).

Nestes casos, os dados inevitavelmente não são representativos espacialmente, além disso, é possível que os dados apresentem autocorrelação (Segurado et al., 2006) ou outra forma de não independência (Rushton *et al*., 2004), sendo necessária a realização de um estudo detalhado sobre esta questão antes da realização dos experimentos.

Alguns trabalhos chegaram a utilizar dados com um metro de resolução espacial (Lassueur et al., 2006). Trabalhos assim são exceções, são raros os estudos com resolução espacial abaixo de 1 km. Zaniewski et al. (2002) e Vargas et al. (2004) utilizaram dados ambientais com 1 km^2 de resolução para prever a ocorrência de espécies vegetais. Luoto et al. (2005) utilizaram uma grade de 10 x 10 km para avaliar incertezas em envelopes climáticos. Phillips et al. (2006), empregaram modelos com resolução espacial relativamente alta (0,05º) em uma área de estudo

continental para comparar o desempenho do Maxent com o GARP. Esses trabalhos evidenciam a preocupação com a influência da escala espacial nos resultados dos modelos. Porém a disponibilidade de dados de ocorrência, de ausência e mesmo dados ambientais é muitas vezes o fator determinante na escolha da escala de trabalho.

2.2.3 Escolha das variáveis

A seleção das variáveis é particularmente importante para os modelos GLM e GAM (Guisan et al., 2002). Apesar dos atuais satélites oferecerem variáveis potencialmente preditivas, muitas variáveis ambientais são coletadas em campo, e acabam sendo interpoladas para gerar um plano de informação com resolução compatível com outras variáveis. Este processo introduz uma série de incertezas que terminam se propagando junto com as incertezas inerentes aos SDMs.

Em contrapartida, existem casos onde há excesso de variáveis, que pode parecer vantajoso para um não estatístico, contudo, estas variáveis podem estar correlacionadas ou não aumentarem significativamente a explicação sobre a variação dos dados (Guisan et al., 2002). Para contornar esse problema e eliminar variáveis, uma alternativa é utilizar o método *stepwise forward* ou *backward* (Netter *et al*., 1996). Outras abordagens para a modelagem como NNETW e algoritmos genéticos possuem seus próprios critérios para a seleção das variáveis (Guisan e Thuiller, 2005).

Os programas analisados neste trabalho não precisam de uma prévia seleção de variáveis ambientais. O Bioclim encontra o menor envelope multidimensional que engloba todas as variáveis, no GARP a seleção de variáveis é parte da rotina, e o Maxent maximiza a informação contida em todas as variáveis ambientais.

2.2.4 Avaliação do modelo

O método tradicionalmente utilizado para avaliação dos modelos lineares é o teste de hipóteses, que verifica se os coeficientes de regressão das variáveis preditivas são significativamente diferentes de zero (Rushton *et al*., 2004; Guisan et al., 2002).

O método de avaliação mais comum em quase todos os modelos de distribuição de espécies é a matriz de confusão de acertos e erros associados à previsão dos modelos (Tabela 2.2). Os itens “a” e “d” são os verdadeiros positivos e verdadeiros negativos, i.e. as presenças e ausências preditas corretamente. Os possíveis erros dos modelos são os falsos positivos e falsos negativos, itens “b” e “c” respectivamente.

TABELA 2.2 – Matriz de confusão

Predito	Real	
	+	-
+	a	b
-	c	d

Analisar a matriz de confusão é essencial para evitar uma superestimativa (Figura 2.7a), um super-ajuste (Figura 2.7b) ou uma omissão alta (Figura 2.7c) do modelo.

Erros de comissão, sobreprevisão ou superestimativa (Figura 2.8) não são considerados verdadeiras falhas de modelos. A região de ocorrência prevista pode ser adequada para a ocorrência da espécie, mas por motivos históricos ou ecológicos a espécie está ausente. O maior equívoco ocorre nos erros de omissão onde o modelo falha em prever a ocorrência da espécie. Um super-ajuste também prejudica a utilidade do modelo, visto que muitos trabalhos visam projetar modelos para outras áreas ou condições climáticas.

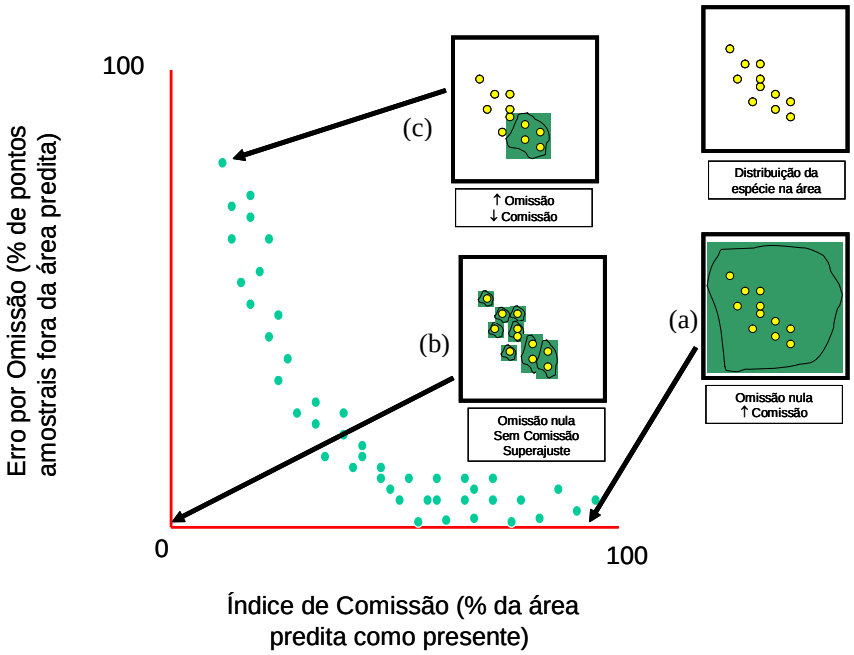

FIGURA 2.7 – Problemas de alta omissão, super-ajuste e superestimativa dos modelos. Fonte: Kansas applied remote sensing program (2005).

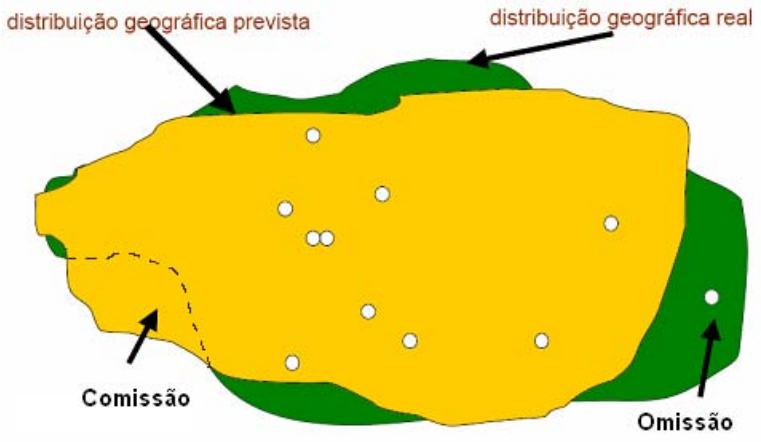

FIGURA 2.8 – Representação dos erros de omissão e comissão.

Fonte: Modificado de Siqueira (2005).

Com os valores dos diferentes tipos de erros (Tabela 2.2) é possível obter uma série de medidas (Tabela 2.3) para a avaliação de desempenho de SDM (Fielding e Bell, 1997). Os cálculos da Tabela 2.3 consideram os itens da Tabela 2.2 como valores absolutos e não proporcionais.

TABELA 2.3 – Medidas derivadas da matriz de confusão de resultados dos SDMs.

Medida	Cálculo
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)
Kappa	$\frac{(a+d) - \frac{((a+c)(a+b)+(b+d)(c+d))}{N}}{N - \frac{((a+c)(a+b)+(b+d)(c+d))}{N}}$

Fonte: Fielding e Bell (1997).

Dentre estas métricas, a prevalência, a sensibilidade e a especificidade são as mais usadas. Prevalência é o total (%) da área de estudo em que a espécie realmente ocorre. Sensibilidade é uma medida que descreve a probabilidade de um *pixel* x ser corretamente classificado como ocorrência. Especificidade é a probabilidade de um *pixel* ser corretamente classificado como ausência (Segurado e Araújo, 2004; Guisan e Zimmermann, 2000; Fielding e Bell, 1997).

É possível avaliar os modelos através de sua área mínima estimada (Engler *et al*., 2004), ou prevalência, que é a mínima superfície obtida considerando todos os *pixels* com previsões acima de um limiar de probabilidade (e.g. 0,7). Ao avaliar um

mapa de distribuição potencial com dados de apenas presença, um mapa indicando ocorrência por todas as partes possuiria a melhor avaliação (baixo erro de omissão), contudo um mapa tão otimista seria de pouca utilidade. Assim, a idéia por trás da área mínima estimada é baseada na premissa de que um bom mapa de distribuição de espécie obtido a partir de dados com apenas presença deveria prever áreas com potencial de ocorrência as menores possíveis, enquanto inclui um número máximo de ocorrência de espécies ao mesmo tempo.

O índice Kappa também é obtido a partir da matriz de confusão. Este índice é considerado uma boa medida de desempenho de modelos de distribuição de espécies porque faz uso de todas as informações contidas na matriz de confusão (Fielding e Bell, 1997). O índice Kappa varia de 0 a 1 e pode ser classificado de acordo com a TABELA 2.4.

TABELA 2.4 – Qualidade do índice Kappa

Índice Kappa	Qualidade
0,01 a 0,20	Ruim
0,21 a 0,40	Razoável
0,41 a 0,60	Boa
0,61 a 0,80	Muito Boa
0,81 a 1,00	Excelente

Fonte: Landis e Koch (1977).

Ao medir o desempenho de um classificador, Pontius Jr. (2000) alerta para o fato que o índice kappa pode conduzir a conclusões equivocadas, pois a tabela de contingência trabalha com proporções, não considerando a localização dos rótulos atribuídos. A matriz de confusão deste trabalho foi construída comparando os planos de informação *pixel* a *pixel*, evitando os erros do índice kappa quanto à localização.

A fim de garantir medidas apropriadas de desempenho, Pontius e Schneider (2001) recomendam o uso do gráfico *Receiver Operating Characteristic* (ROC-plot). O chamado ROC-plot é uma opção para avaliar os modelos, onde é representado em um gráfico a sensibilidade contra a especificidade (Figura 2.9). O ROC-plot é uma medida independente da prevalência (Manel et al., 2001) e correlacionada com o Kappa (Anderson et al., 2003) onde a área sob a curva (*Area Under the Curve* – AUC) é a medida de desempenho. Quanto mais próximo de um for a área, melhor o desempenho (Phillips *et al*., 2005; Rushton *et al*., 2004). Este método é bastante utilizado porque é uma medida global de desempenho independente de limites de corte (Fielding e Bell, 1997), geralmente empregados na construção da matriz de confusão.

Guisan e Zimmermann (2000) propõem mais duas abordagens para a avaliação do modelo: calibrar o modelo e realizar a validação cruzada, *Jack-knife (leave-one-out)* ou *bootstrap*; ou utilizar dois conjuntos independentes de dados, um para calibrar e outro para validar, como também é feito no processo de regressão linear múltipla (Netter *et al*., 1996).

O emprego da validação cruzada, do *Jack-knife* ou *bootstrap* é mais adequado quando o conjunto de dados é demasiado pequeno para separá-lo em dado de calibração e dado de avaliação (Fielding e Bell, 1997). O método *bootstrap* é uma técnica de re-amostragem que permite investigar a tendência de uma estimativa através da realização de múltiplas re-amostragens (com reposição) dentro do conjunto de dados de calibração, que então o remove para obter uma estimativa não tendenciosa (Guisan e Zimmermann, 2000).

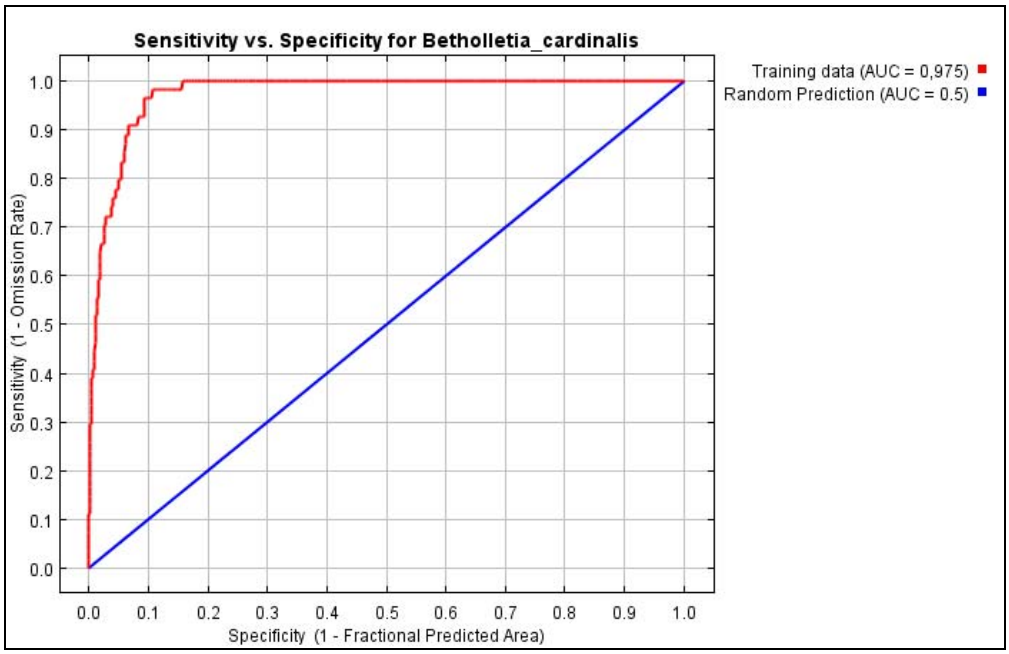

FIGURA 2.9 – Exemplo de um ROC-plot. A linha azul representa o resultado de uma previsão completamente aleatória. A linha vermelha é a análise do modelo, quanto maior a diferença entre o resultado do modelo e a aleatoriedade, melhor seu desempenho.

Após comparar seis métodos para modelar a distribuição espacial de 44 espécies de anfíbios e répteis, Segurado e Araújo (2004) discutem que dificilmente será encontrado o “melhor” modelo, pois cada método possui pontos fortes, bem como fraquezas. A escolha do método apropriado depende dos dados, das premissas e dos objetivos. Segundo estes autores, isso deixa duas alternativas para os pesquisadores: utilizar um sistema especialista (GARP, por exemplo) que compara métodos automaticamente e escolhe o melhor método para cada espécie ou um método que é mais robusto genericamente (como o GLM) ou; a segunda opção é escolher um método que é robusto particularmente para o tipo de dados e o objetivo do trabalho.

3 METODOLOGIA

A metodologia deste trabalho pode ser divida em três grandes blocos de procedimentos. No primeiro são selecionados os modelos a serem avaliados, as variáveis ambientais empregadas, a área de estudo e a escala da modelagem (Figura 3.1a). O segundo bloco consiste na simulação do nicho fundamental e dos pontos de ocorrência da espécie. A partir destes pontos são simulados erros de posicionamento com dois métodos diferentes, através da projeção em centróides de células e introdução de erros com parâmetros de coordenadas polares (Figura 3.1b). O último bloco é a avaliação da sensibilidade dos modelos a erros de posicionamento através de quatro métricas de avaliação (Figura 3.1c).

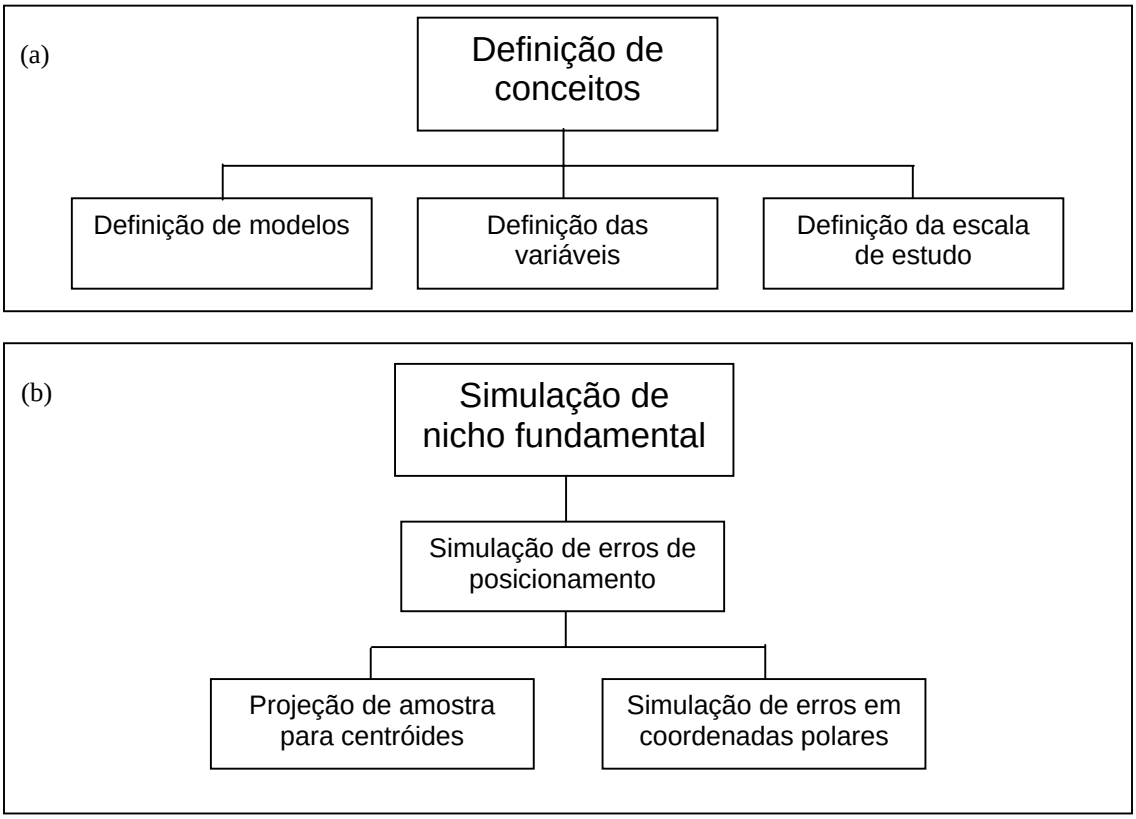

FIGURA 3.1 – As três principais etapas para a avaliação dos modelos. (a) Definição de conceitos e premissas, (b) simulação do nicho fundamental e dos erros de posicionamento e (c) avaliação da sensibilidade dos modelos a erros de posicionamento.

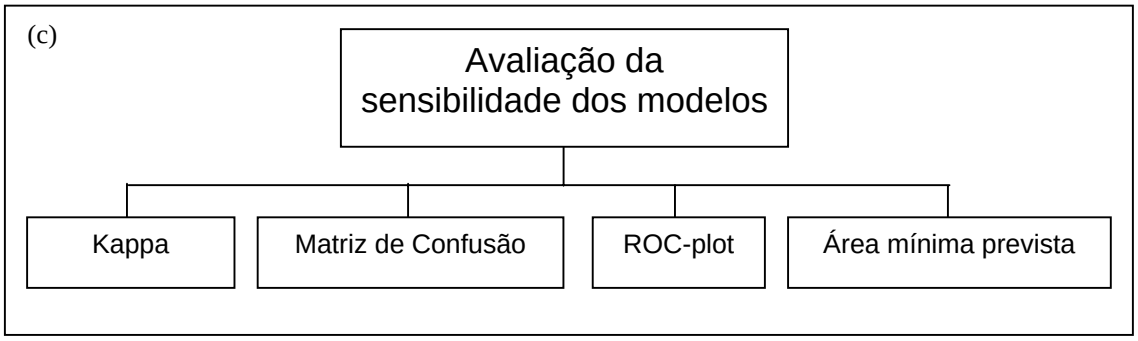

FIGURA 3.1 cont. – As três principais etapas para a avaliação dos modelos. (a) Definição de conceitos e premissas, (b) simulação do nicho fundamental e dos erros de posicionamento e (c) avaliação da sensibilidade dos modelos a erros de posicionamento.

3.1 Definição dos conceitos da modelagem

3.1.1 Seleção dos modelos

Foram analisados o BIOCLIM, o OM-GARP *Best subsets* e o Maxent. A escolha destes métodos para a execução dos modelos se deve inicialmente ao fato de que, de acordo com Elith *et al*. (2006), estes se encaixam em diferentes categorias de desempenho (Figura 3.2), e o BIOCLIM e o GARP estão implementados no *openModeller*.

O BIOCLIM representa um dos modelos mais empregados para realizar prognósticos sobre os mais diversos cenários de mudanças climáticas globais (Beaumont et al., 2005). O GARP, também bastante utilizado, possui uma capacidade inerente de lidar com erros através da seleção de modelos ótimos (Stockwell e Peters, 1999). O Maxent caracteriza-se pela qualidade de realizar previsões sobre informações incompletas (Phillips et al., 2006).

Como a maioria dos dados reais provém de herbários ou museus onde raramente são registradas as ausências, os modelos devem ser do tipo que necessita somente de dados de presença.

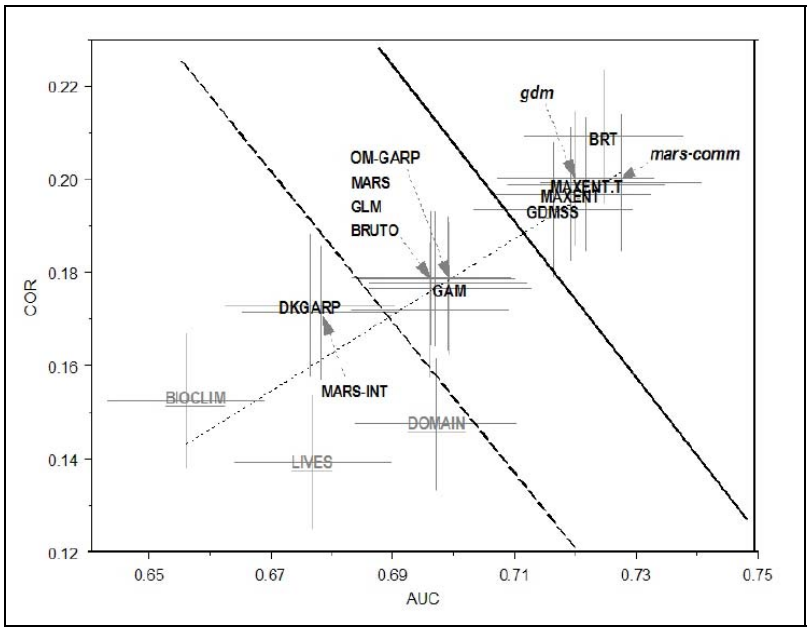

FIGURA 3.2 – Área sob a curva (AUC) média vs correlação (COR) média para os métodos de modelagem. A AUC é uma medida de desempenho do modelo independente da matriz de confusão, e a COR é outra medida que testa amostras contra área predita. Assim, os modelos que estão no canto superior direito têm melhor desempenho e os modelos do canto inferior esquerdo, menor desempenho. Fonte: Elith *et al*. (2006).

3.1.2 Variáveis ambientais preditivas

Adotou-se a premissa de que apenas as características físicas do ambiente são determinantes do nicho fundamental da espécie de estudo. Para esta caracterização foram selecionadas informações climáticas, de relevo e de solo disponíveis. Assim, como variáveis ambientais preditivas foram escolhidas a precipitação, temperatura, relevo e umidade do solo (Tabela 3.1) em um total de 41 planos de informação.

As Figuras 3.3, ilustra a variabilidade espacial do relevo e das médias anuais das variáveis climáticas. As Figuras 3.4, 3.5 e 3.6 mostram como as médias da umidade do solo, da temperatura e da precipitação variam ao longo do ano.

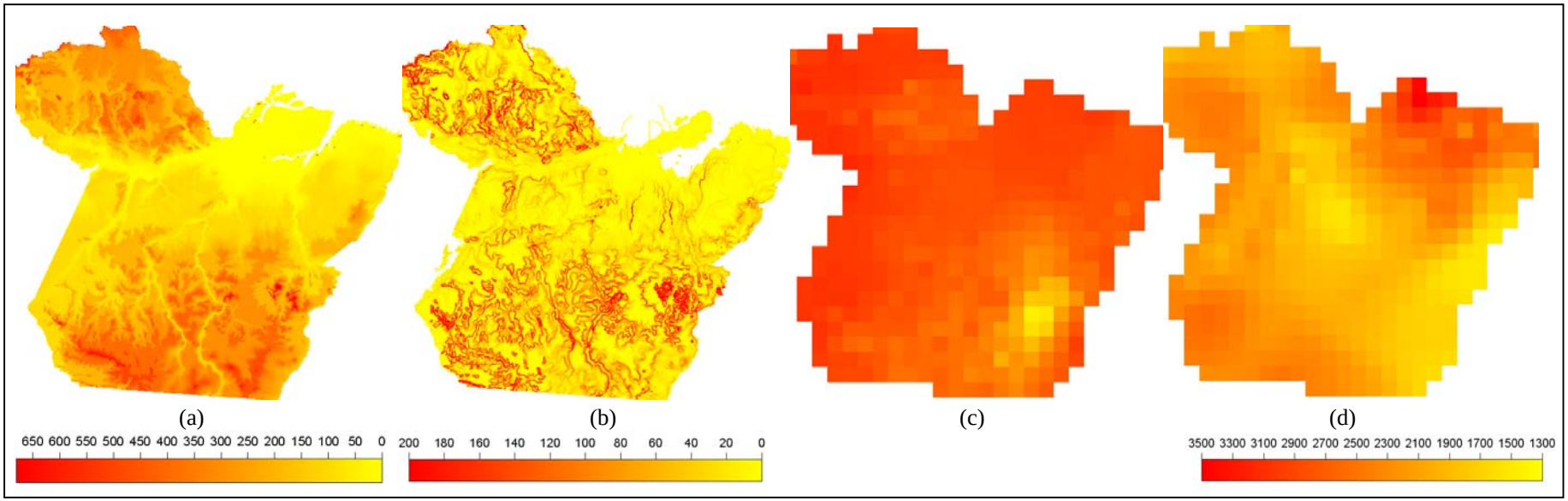

FIGURA 3.3 – Dados de relevo. (a) elevação e (b) declividade e valores médios de (c) Temperatura e (d) Precipitação

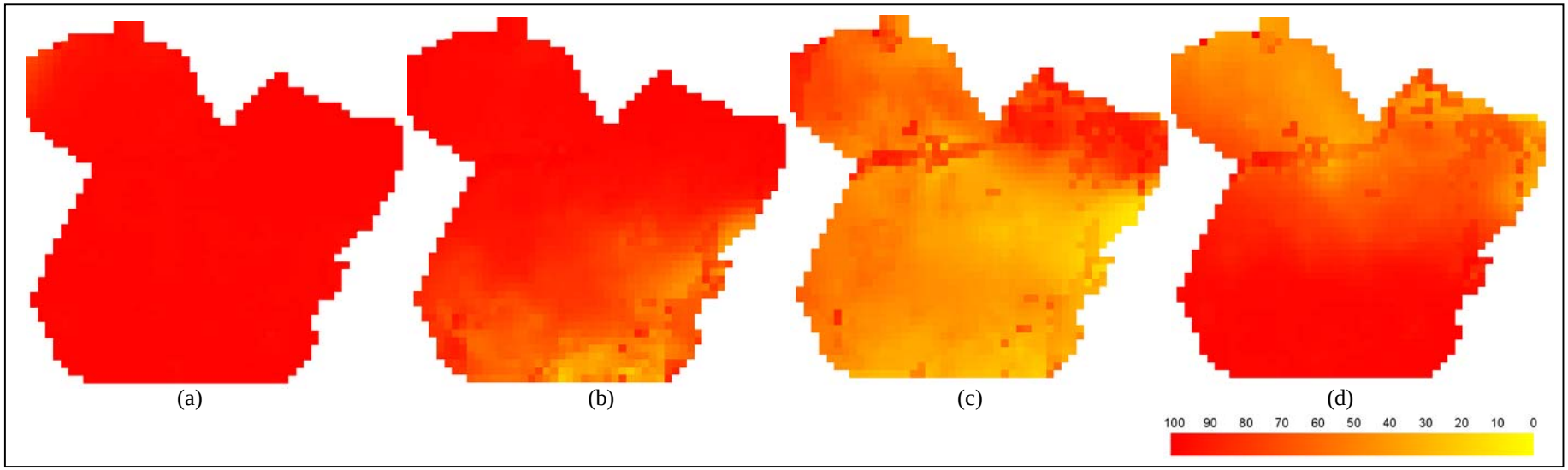

FIGURA 3.4 – Umidade relativa (%) presente no solo, média para os meses de (a) Março, (b)Junho, (c) Setembro e (d) Dezembro

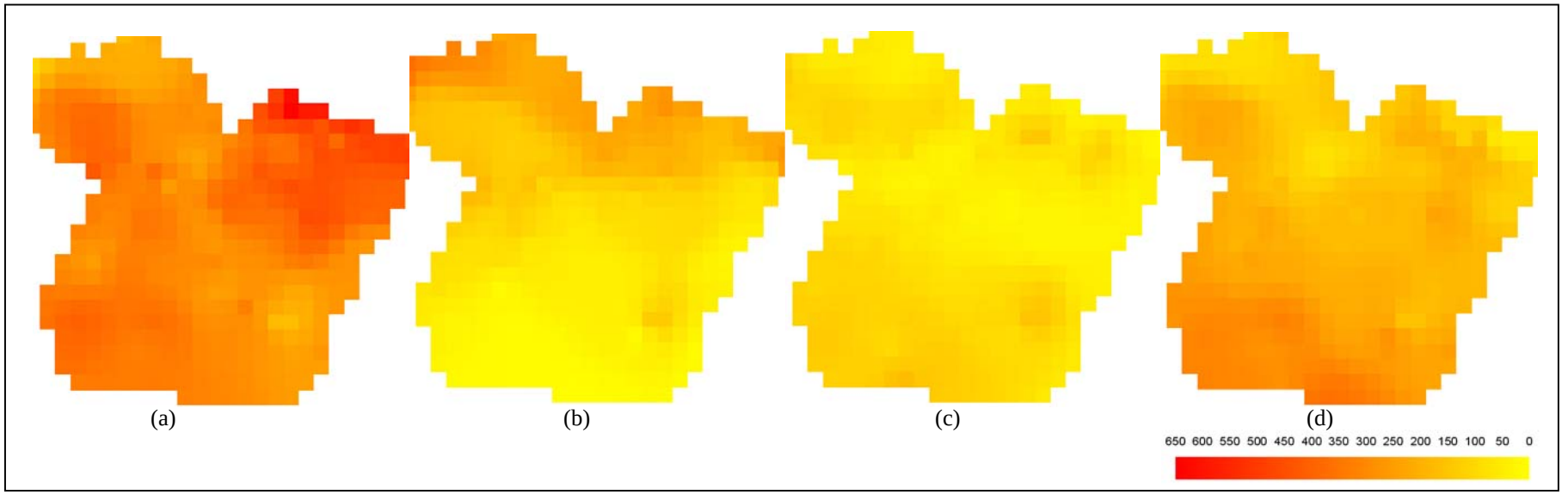

FIGURA 3.5 – Precipitação (mm) média mensal para os meses de (a) Março, (b)Junho, (c) Setembro e (d) Dezembro

57

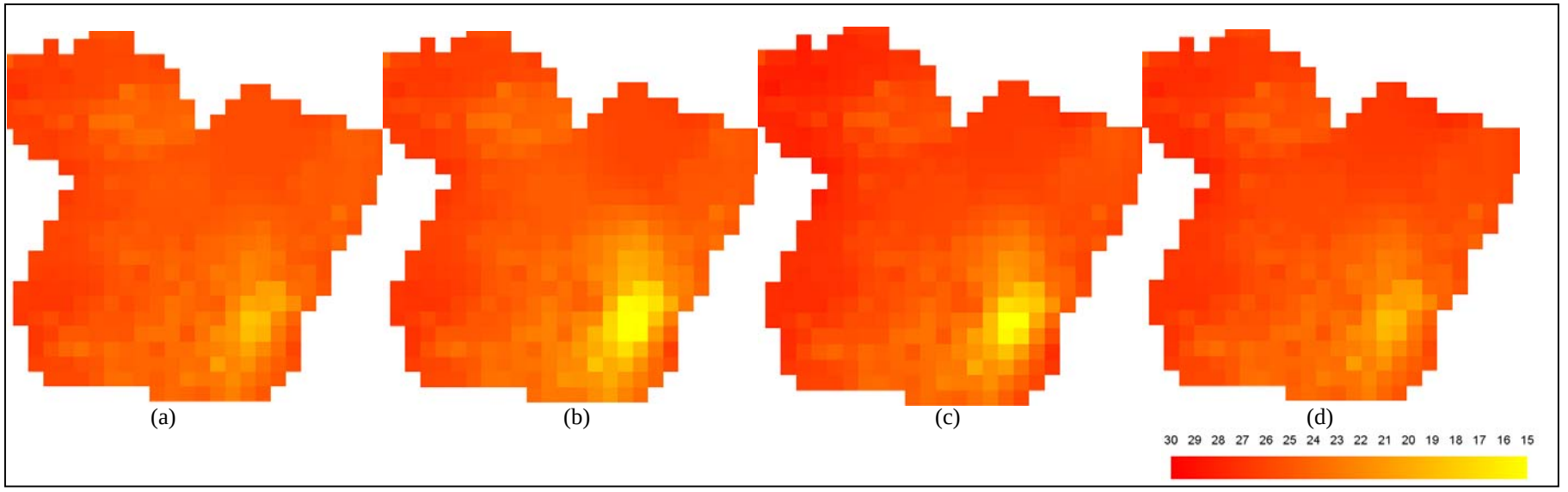

FIGURA 3.6 – Temperatura (ºC) média mensal para os meses de (a) Março, (b)Junho, (c) Setembro e (d) Dezembro

Para o clima, a base de dados do *Worldclim* (Hijmans et al. 2005), comumente utilizada em estudos de modelagem de espécies foi adotada. Os dados de precipitação e temperatura foram obtidos de uma rede mundial de estações de coleta (Hijmans et al., 2005). Os dados são uma média do período de 1960-1990 e foram gerados por interpolação dos dados das estações, possuindo uma resolução espacial de 50 km.

TABELA 3.1 – Resolução espacial das variáveis preditivas.

	Resolução espacial (km)
Elevação	1
Declividade	1
Aspecto	1
Temperatura média anual	50
Temperatura média mensal (Jan a Dez)	50
Precipitação média anual	50
Precipitação média mensal (Jan a Dez)	50
Umidade do solo média mensal (Jan a Dez)	25

O modelo digital de elevação do SRTM (*Shuttle Radar Topographic Mission*) tem uma resolução original de 90 m, contudo a base de dados foi reamostrada para 1 km. A partir do modelo digital de elevação foram calculados o aspecto e a declividade. Os dados foram fornecidos pelo JPL (*Jet Propulsion Laboratory*) da NASA (*National Aeronautics and Space Administration*).

Para a caracterização do solo optou-se por utilizar dados de umidade, fator limitante para muitas espécies vegetais. Foram usados os mapas de umidade relativa produzidos por Rossato et al. (2004) de todo o território brasileiro. Estes dados representam uma média do período de 1971 a 1990 para cada mês a partir de dados do sensor AVHRR (*Advanced Very High Resolution Radiometer*) dos satélites da série NOAA (*National Oceanic and Atmospheric Administration*), de

amostras de solos de campo e de dados meteorológicos de agências governamentais. Para este produto, foram processados dados da Agência Nacional de Energia Elétrica, do Departamento de Águas e Energia do Estado de São Paulo, do Sistema Meteorológico do Paraná, da Superintendência de Desenvolvimento do Nordeste e do Instituto Nacional de Meteorologia.

O período de junho a setembro, devido ao período de estiagem, apresenta os menores valores de umidade no solo. O aumento do percentual ocorre a partir de outubro se estendendo até fevereiro. O estado do Pará apresenta baixos índices de umidade relativa até o mês de novembro (Rossato et al., 2004). Assim, para as operações de geoprocessamento, na simulação do nicho fundamental, foram utilizados os dados dos meses de agosto e janeiro, representando um mês seco e outro chuvoso.

3.1.3 Escala de estudo

Segundo levantamentos do PRODES (2006), o Pará é o estado com a segunda maior taxa de desmatamento no Brasil, atrás apenas do Mato Grosso. O estado também possui o município com a maior área desmatada até 2006, São Félix do Xingu, com 84.249 km^2 desflorestados.

O desmatamento e o processo de ocupação do Pará estão vinculados à expansão da fronteira agrícola, apropriação fundiária e exploração dos recursos naturais (Escada et al., 2005) gerando um cenário de conflitos sócio-econômicos e ambientais.

O projeto GEOMA, no qual este trabalho está inserido, possui o objetivo de desenvolver modelos para avaliar e prever condições de sustentabilidade sob diferentes tipos de atividades humanas e de políticas públicas (GEOMA, 2006). Esse panorama levou a escolha do estado do Pará como área de estudo deste trabalho.

É necessário definir a escala de estudo em dois aspectos, extensão e resolução espacial. A extensão da escala de estudo foi definida pela própria dimensão da área de estudo, o estado do Pará. A resolução espacial de estudo foi determinada pela resolução mais fina entre as variáveis ambientais disponíveis, 1 km (Tabela 3.1).

3.2. Simulação da espécie

Vícios de coleta (Reddy e Dávalos, 2003; Hirzel e Guisan, 2002) ou baixo número de ocorrência (Stockwell e Peterson, 2002) são problemas que amostras coletadas em campo podem apresentar e que influenciam significativamente no desempenho dos modelos. Para que a avaliação dos modelos possa ser efetuada sobre um desenho experimental onde existe um número menor de fatores que podem influenciar o resultado, é preciso ter controle sobre a amostragem. No caso da avaliação da influência dos erros de posicionamento, também é necessário ter controle dos diferentes tipos de erros de posicionamento.

Nesse sentido, propõe-se o uso de uma espécie virtual conforme sugerido por outros autores (Austin et al., 2006; Hirzel et al., 2001). A amostragem e os erros de posicionamento foram simulados sobre um ambiente real, o estado do Pará. O procedimento proposto simula o nicho fundamental (Araújo e Guisan, 2006) de uma espécie vegetal arbórea. A partir deste nicho fundamental os erros de posicionamento foram também simulados e avaliados no processo de modelagem de distribuição da espécie.

Para simular o nicho fundamental tomou-se por base a teoria dos envelopes bioclimáticos. Modelos de envelopes bioclimáticos predizem locais com condições climáticas favoráveis a uma espécie baseados no cálculo de um intervalo de confiança no espaço climático multidimensional (Guisan e Zimmermann, 2000). O processo de envelopes é similar uma operação booleana de geoprocessamento, onde os intervalos ótimos para a ocorrência da espécie são delimitados em cada plano de informação.

Para simular o nicho fundamental de uma espécie arbórea virtual, foram delimitados intervalos ótimos das variáveis de relevo, clima e umidade do solo. Operações de intersecção entre as variáveis, através das ferramentas de geoprocessamento permitiram a definição de mapas base e integrá-los em um mapa final de nicho potencial para a espécie.

A definição das condições favoráveis à espécie virtual foi baseada em Locatelli *et al.* (2003) que descrevem os aspectos ambientais adequados para a castanha do Brasil (*Bertholettia excelsa* HBK). Esta espécie foi escolhida como referência para simular a espécie virtual deste trabalho por sua importância econômica nas comunidades locais e por ser uma espécie protegida por lei e com ocorrência no estado do Pará.

Para a simulação, definiu-se que a espécie virtual é encontrada em ambientes de terra firme, não suportando os solos ácidos e encharcados das várzeas. A intolerância à acidez de solo a torna ausente no cerrado. Apesar de associada a ambientes de terra firme, a espécie não ocorre em altitudes superiores a 1600 m e em declividades acima de 80%. A faixa ótima de temperatura para seu desempenho fisiológico encontra-se entre 22º C a 32º C, acima de 1700 mm anuais e entre 10% a 65 % de umidade do solo.

A Figura 3.7 indica o nicho fundamental simulado por geoprocessamento para a espécie virtual, no estado do Pará. O nicho fundamental simulado serve de base para a próxima etapa que consiste em simular os pontos de amostragens nos locais favoráveis à ocorrência da espécie e os erros de posicionamento.

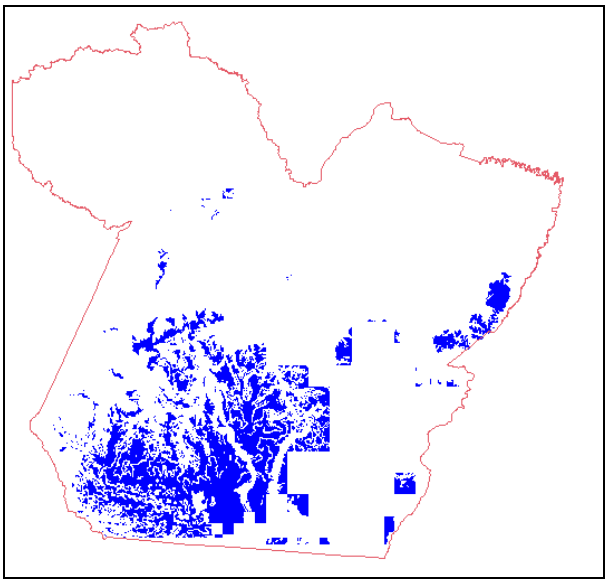

FIGURA 3.7 – Distribuição espacial da castanheira simulada por geoprocessamento para o estado do Pará.

Austin et al. (2006) e Hirzel et al. (2001) empregaram espécies simuladas para avaliar SDM. Em seus trabalhos foram gerados mais de 1000 pontos de ocorrência. Neste trabalho, tomando como base a área do nicho simulado, foram gerados 150 pontos de ocorrência artificiais na tentativa de aproximar a simulação à realidade da disponibilidade de dados reais e da viabilidade de trabalhos de campo.

3.2.1 Simulação dos erros

Neste trabalho é admitido que as ocorrências geradas foram “coletadas” em condições ótimas, i.e., utilizando GPS. A essas amostras são atribuídos erros de posicionamento em diferentes escalas e em dois diferentes métodos; projeção em centróides de células e erros com distribuição normal em coordenadas polares. Os erros assumidos são de até 10 km, até 0,25º, até 0,5º e até 1º. Essas dimensões foram escolhidas baseadas na resolução espacial das variáveis ambientais preditivas, que são de 1 km, 25 km e 50 km. Próximo à linha do equador, 0,25º equivale a

aproximadamente 27,75 km, 0,5º a 55 km e 1º a 111 km. Os erros de até 10 km foram escolhidos por estarem em uma resolução intermediária entre 1 km e 25 km

Projeção em centróides de células

Hirzel e Guisan (2002) testaram diferentes tipos de amostragem e obtiveram melhores desempenhos de modelos para amostras distribuídas espacialmente de forma regular. Edwards Jr. et al. (2006) também obtiveram melhores resultados através de uma amostragem probabilística com uma malha regular. Nesse sentido, buscou-se simular erros projetados em centróides de células de modo a buscar uma distribuição uniforme de amostras (Figura 3.8).

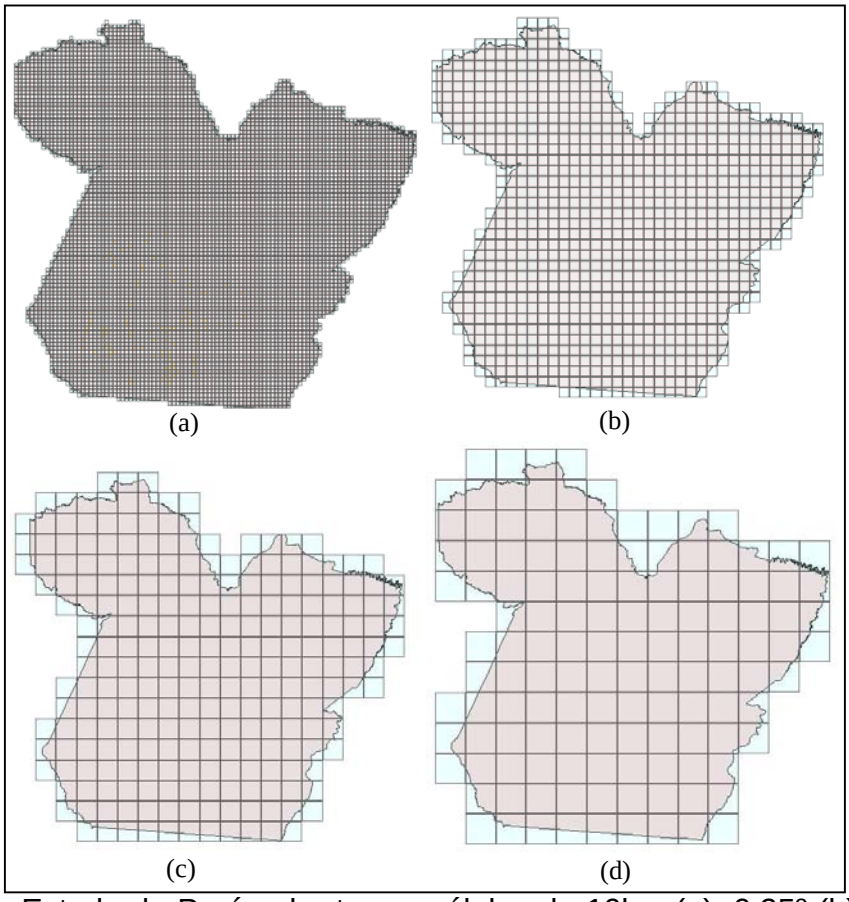

Figura 3.8 – Estado do Pará coberto por células de 10km (a), 0,25º (b), 0,5º (c) e 1º (d).

Esta simulação parte da situação hipotética que existem ocorrências em uma região delimitada por uma célula, mas sua localização exata é desconhecida. Gera-se uma rede de células e as amostras têm sua localização projetada para o centróide da célula que a contém (Figura 3.9).

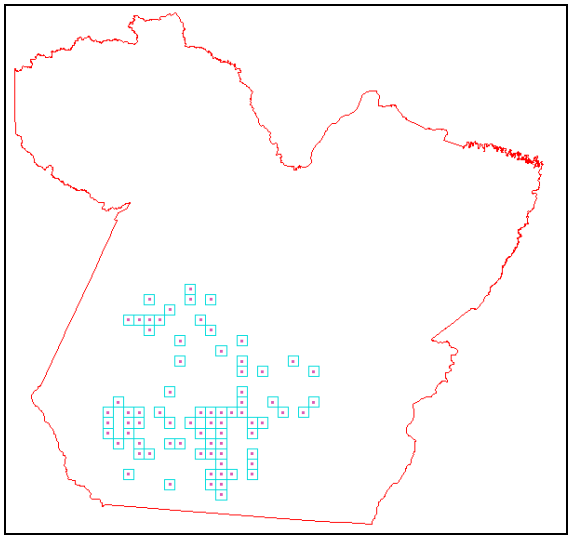

FIGURA 3.9 – Células de 0,25º selecionada por consulta espacial com centróides para a simulação de erros.

No caso das ocorrências simuladas, como sua localização é conhecida, é realizada uma consulta espacial para sua projeção nos centróides das respectivas células. Todas as amostras contidas na área da célula são projetadas para um mesmo ponto (Figura 3.10), pois se pressupõe que ao utilizar dados reais suas verdadeiras localizações não serão conhecidas.

O maior erro possível é a metade da diagonal da célula (E_3 da Figura 3.10), dessa forma temos as dimensões das células (Tabela 3.2) por trigonometria (Equação 3.1):

$$Lado = \frac{2E_3}{\sqrt{2}} \qquad (3.1)$$

TABELA 3.2 – Dimensões das células

Erro máximo	Lado da célula em graus (º)
10 km	0,127407
0,25º	0,353553
0,5º	0,707107
1º	1,414214

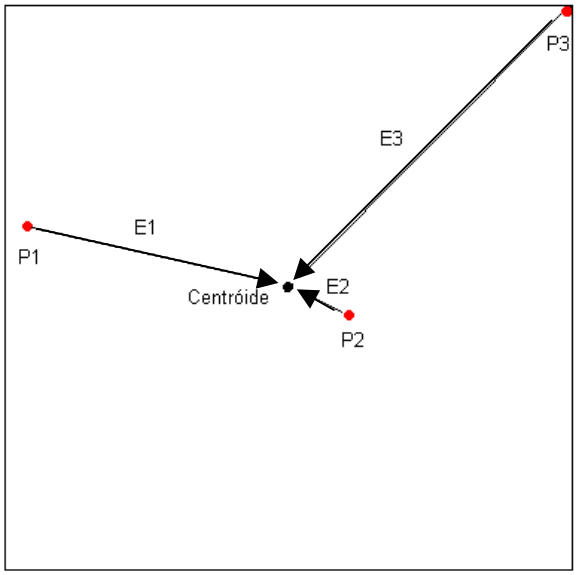

FIGURA 3.10 – Projeção de amostras com localização desconhecidas para o centróide da célula. P1, P2 e P3 são pontos de ocorrência com localização "desconhecida". Todas as ocorrências desta célula são projetadas para o mesmo ponto, o centróide da quadrícula. E1, E2 e E3 são os erros de posicionamento associados a cada amostra.

Erros em coordenadas polares

A projeção de pontos para centróides parte de posições "desconhecidas", os pontos simulados originais, e geram uma posição conhecida, os centróides. Uma forma mais intuitiva para simulação de erros é um método inverso ao de células, onde a partir de pontos conhecidos, as 150 amostras simuladas, os erros são atribuídos gerando posições "desconhecidas" (Figura 3.11). Dessa forma é

possível atribuir erros conhecidos e medidos por distância euclidiana (Equação 3.2).

$$d = \sqrt{x^2 + y^2} \tag{3.2}$$

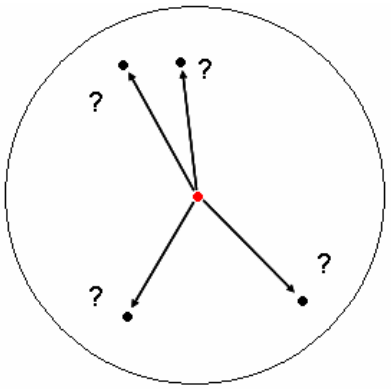

FIGURA 3.11 – A introdução de erros em coordenas polares é intuitiva.

Este método pressupõe que os erros de posicionamento contidos no conjunto de amostras tendem para uma distribuição normal de acordo com o Teorema do Limite Central. Este tipo de erro é uma tentativa de aproximação da realidade, onde existem coordenadas para o ponto, mas o erro de posicionamento associado é desconhecido.

As coordenadas polares são uma alternativa às medidas deste sistema cartesiano que apresenta algumas dificuldades devido ao termo racional na equação. Assim, dado um ponto P do plano, utilizando coordenadas cartesianas (retangulares), sua localização é descrita no plano escrevendo P=(x_e,y_e) onde x_e é a projeção de P no eixo *x* e y_e, a projeção no eixo *y*. Em coordenadas polares (Figura 3.12) a localização de P é descrita a partir da distância de P à origem O do sistema e do ângulo formado pelo eixo *x* e o segmento OP, caso P≠O. Denota-se P=*(r,θ)* onde *r* é a distância de P a O e *θ* o ângulo tomado no sentido anti–horário, da parte positiva do eixo O*x* ao segmento OP, caso P≠O.

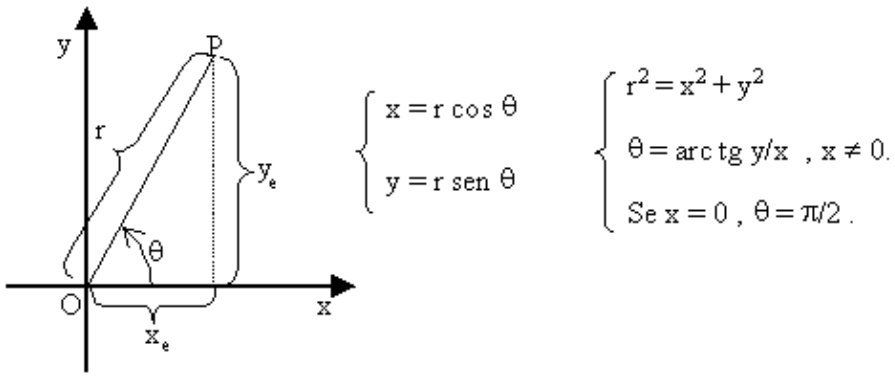

FIGURA 3.12 – Conversões de coordenadas cartesianas para polares.

Análogo à Figura 3.12 temos erros de posicionamento em coordenadas polares (Figura 3.13) para uma amostra em coordenadas geográficas: um erro angular θ, entre 0 e 2π radianos e um erro radial *r*, que varia de 0 até o erro máximo (10km, 0,25º, 0,5º ou 1º).

O ponto de partida para a simulação é a geração de uma série de conjuntos de números aleatórios uniformes. O método de *Box-Müller* (proposto por George Edward Pelham Box e Mervin Edgar Muller em 1958) é utilizado para simular dados de uma distribuição normal padrão (Equação 3.3).

$$Z = \cos(2\pi U_1) * \sqrt{-2\ln(U_2)} \qquad (3.3)$$

Onde U_1 e U_2 são valores de duas distribuições uniformes entre 0 e 1, independentes entre si. Sobre os conjuntos de Z (que possuem valores entre -1 e 1) foram aplicadas operações aritméticas de forma de forma que os valores permaneçam contidos entre 0 e 2π radianos para ângulos e de 0 até os respectivos erros máximos de cada nível de erro (10km, 0,25º, 0,5º ou 1º).

As medidas obtidas de θ e *r* são atribuídas aos pontos originais que em seguida são convertidos de volta para coordenadas cartesianas (Figura 3.12).

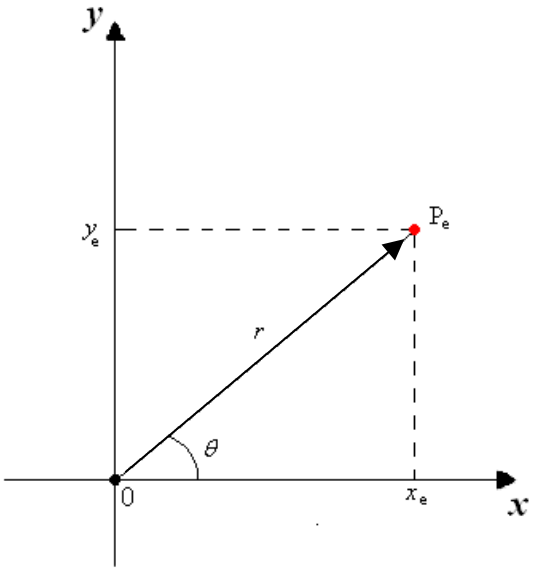

FIGURA 3.13 – Erros de posicionamento em coordenadas polares. O zero é a posição original, atribui-se um erro θ e um erro r para obter a posição P_e. As coordenadas cartesianas de P_e são dadas por x_e e y_e.

3.2.2. Avaliação dos modelos

Não foram encontrados na literatura trabalhos que relatem a sensibilidade dos modelos a erros de posicionamento de dados de entrada. Além disso, os trabalhos que comparam o desempenho de modelos não classificam os dados quanto à precisão do posicionamento (Austin et al., 2006; Elith *et al*., 2006; Phillips *et al*., 2006; Guisan e Thuiller, 2005; Segurado e Araújo, 2004, Anderson et al., 2003; Hirzel et al., 2002; Manel et al., 2002).

Ao analisar o desempenho de modelos na previsão de ocorrência de espécies seria usual comparar os resultados dos modelos com o habitat simulado, nossa verdadeira distribuição assumida (Austin et al., 2006, Hirzel et al., 2001). Mas o objetivo é analisar a influência dos erros de posicionamento sobre diferentes SDMs, assim, a comparação é realizada sobre o melhor resultado que o modelo gera em relação à qualidade do posicionamento dos pontos de coleta. Ou seja, a

saída do modelo onde os dados foram “coletados por GPS” é referência para ser comparada com os resultados dos modelos que foram alimentados com amostras que apresentam erros de posicionamento em diferentes níveis (Figura 3.14).

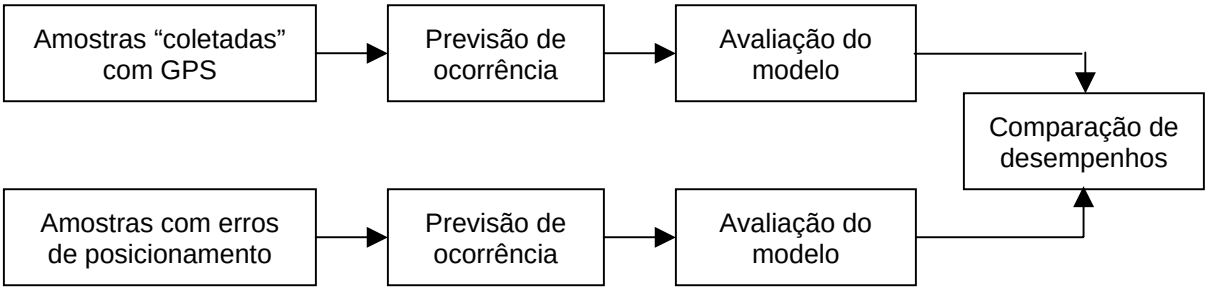

FIGURA 3.14 – Método empregado para a comparação de modelos

Cada modelo (Bioclim, Garp e Maxent) é rodado com as amostras de treino e em cada nível de erro (10 km, 0,25º, 0,5º e 1º) com dois métodos de simulação de erros (associados a centróides e em coordenadas polares), em um total de 10 previsões por modelo. Como todas as ocorrências contidas em uma célula são projetadas para o mesmo centróide, há uma diminuição no número de amostra de acordo com o tamanho da quadrícula (Tabela 3.3).

TABELA 3.3 – Número de amostras simuladas e empregadas por modelo e por escala

Erros	**Modelos**					
	Centróides de células			**Erros polares**		
	BIOCLIM	OM-GARP	MAXENT	BIOCLIM	OM-GARP	MAXENT
Treino	100	100	100	100	100	100
10 km	100	100	100	100	100	100
0,25º	66	66	66	100	100	100
0,5º	39	39	39	100	100	100
1º	24	24	24	100	100	100

Métricas de avaliação

Fielding e Bell (1997) estabelecem uma série de critérios para a avaliação de modelos. Três desses itens são relevantes para este trabalho:

- Caso as previsões estejam restritas a uma área homogênea, deve-se considerar a partição dos dados em treino e teste.
- Considerar o efeito de prevalência. Medidas de acurácia global pode ser um guia pobre para avaliar modelos.
- Se os SDM serão hierarquizados quanto ao seu desempenho, comparações baseadas no ROC-plot são mais robustas já que é uma medida independente da matriz de confusão.

A região do estado do Pará é relativamente homogênea quanto à temperatura e precipitação, fato compensado pela variabilidade temporal (Figura 3.5 e 3.6). Assim as ocorrências simuladas foram separadas em treino e teste. A proporção ideal é de 50% para cada conjunto (Hirzel e Guisan, 2002), mas optou-se por gerar 100 pontos para treino e 50 para testes a fim de simular uma situação que se aproxime da realidade do pesquisador.

Se as amostras de treino forem muito maiores que as de teste, o modelo gerado pelo treino terá um desempenho melhor que o seu subconjunto, o teste (Figura 3.10). Em contrapartida, conjuntos grandes diminuem a variância dos erros. Além disso, avaliar o modelo baseado no resultado gerado a partir de um conjunto menor de dados pode levar a conclusões equivocadas (Fielding e Bell, 1997).

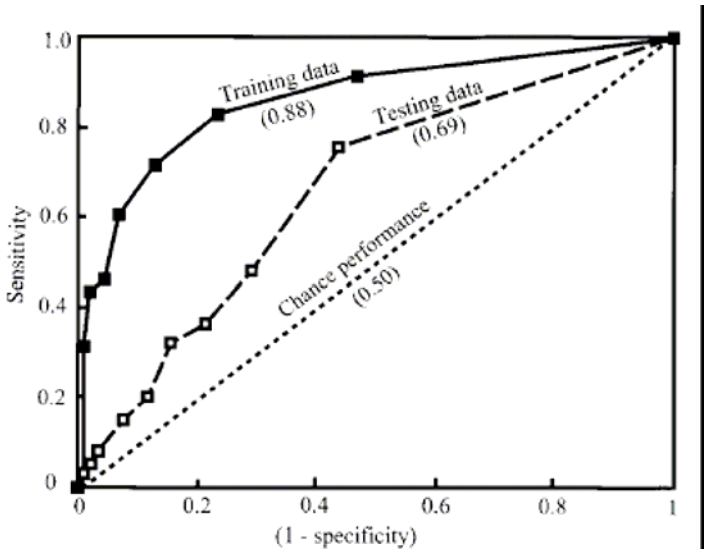

FIGURA 3.15 – Os dados utilizados para teste geralmente possuem um desempenho abaixo dos dados de treinamento.
Fonte: Fielding e Bell (1997)

Outro método de avaliação é a comparação das saídas através de regressão linear (Lassueur et al., 2006). Onde cada saída de um determinado modelo contendo erros é comparada ao melhor modelo gerado, aquele cuja entrada foram as amostras que simulam coletas com GPS. Em um primeiro momento é criada uma malha de pontos sobre a área de estudos e dessa malha de pontos são sorteados aleatoriamente 140 pontos, garantindo uma boa representatividade nos testes de avaliação. Cada modelo utilizado neste trabalho gera um plano de informação com valores associados à probabilidade de ocorrência da espécie. Para cada resultado realiza-se uma intersecção do plano de informação com as amostras sorteadas, desse modo são extraídos os valores correspondentes a estes pontos em cada plano de informação resultante dos modelos. Os valores destes pontos são utilizados para a comparação entre os modelos.

A partir destes pontos e é possível montar uma regressão linear onde a variável independente são os pontos de treino, e assim testar a hipótese que a curva possui uma inclinação de 45º. Se a curva possuir uma inclinação estatisticamente igual a 1, as saídas dos modelos são consideradas significativamente iguais.

Para testar se um modelo de regressão linear (Guisan e Zimmermman, 2000; Fielding e Bell, 1997) utilizados em GLMs, por exemplo, é significativamente diferente de zero é realizado o teste (Equação 3.4):

$$t = \frac{b_1 - \beta_1}{s(b_1)}, \beta_1 = 0 \quad (3.4)$$

Onde β_1 é a inclinação da curva de regressão, b_1 é uma estimativa do seu valor, $s(b_1)$ é o desvio padrão estimado e t é o desvio padrão associado a uma distribuição normal padrão.

O caso da avaliação dos erros de posicionamento é ligeiramente diferente, deseja-se saber se os resultados dos modelos são iguais, onde a inclinação da reta de regressão deveria ser de 45º caso os resultados sejam significativamente iguais, assim testa-se:

$$t = \frac{b_1 - 1}{s(b_1)}, \beta_1 = 1 \quad (3.5)$$

Portanto, para avaliar os modelos foram utilizados os métodos de separação de amostras de ocorrências em treino e teste (matriz de confusão), o índice kappa, o gráfico ROC-plot, e a regressão linear.

4 AVALIAÇÃO DOS MODELOS DE DISTRIBUIÇÃO DE ESPÉCIES

Este capítulo apresenta os resultados das avaliações dos modelos Bioclim, OM-GARP *Best Subsets* e Máxima entropia. Em primeiro lugar, para cada modelo é apresentado o resultado do modelo gerado a partir do conjunto de amostras de treino, que não possui erros. A discussão segue com os modelos gerados a partir dos conjuntos amostrais com erros de posicionamento. As análises dos modelos finalizam com as métricas de avaliação.

4.1 BIOCLIM

Como observado na seção 3.3, o processo que estabelece os limites ótimos do envelope resultante do Bioclim é semelhante à simulação do nicho fundamental através de técnicas de geoprocessamento. Por isso, entre os três modelos (Bioclim, GARP e Maxent) e suas respectivas respostas, visualmente o Bioclim (Figura 4.1) é o que modela melhor o nicho fundamental, quando as amostras não possuem erros de posicionamento. Deste primeiro resultado é extraída a área de ocorrência predita pelo conjunto de treino, que corresponde a 6,91% da área de estudo.

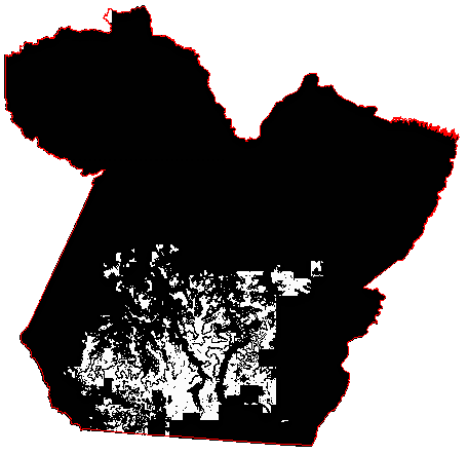

FIGURA 4.1 – Previsão de ocorrência por envelope bioclimático com conjunto de amostras de treino

FIGURA 4.2 – Modelos Bioclim com erros projetados em centróides: (a) Erros de até 10 km (b) Erros de até 0,25º (c) Erros de até 0,5º e (d) Erros de até 1º, e com erros em coordenadas polares: (e) Erros de até 10 km (f) Erros de até 0,25º (g) Erros de até 0,5º e (h) Erros de até 1º.

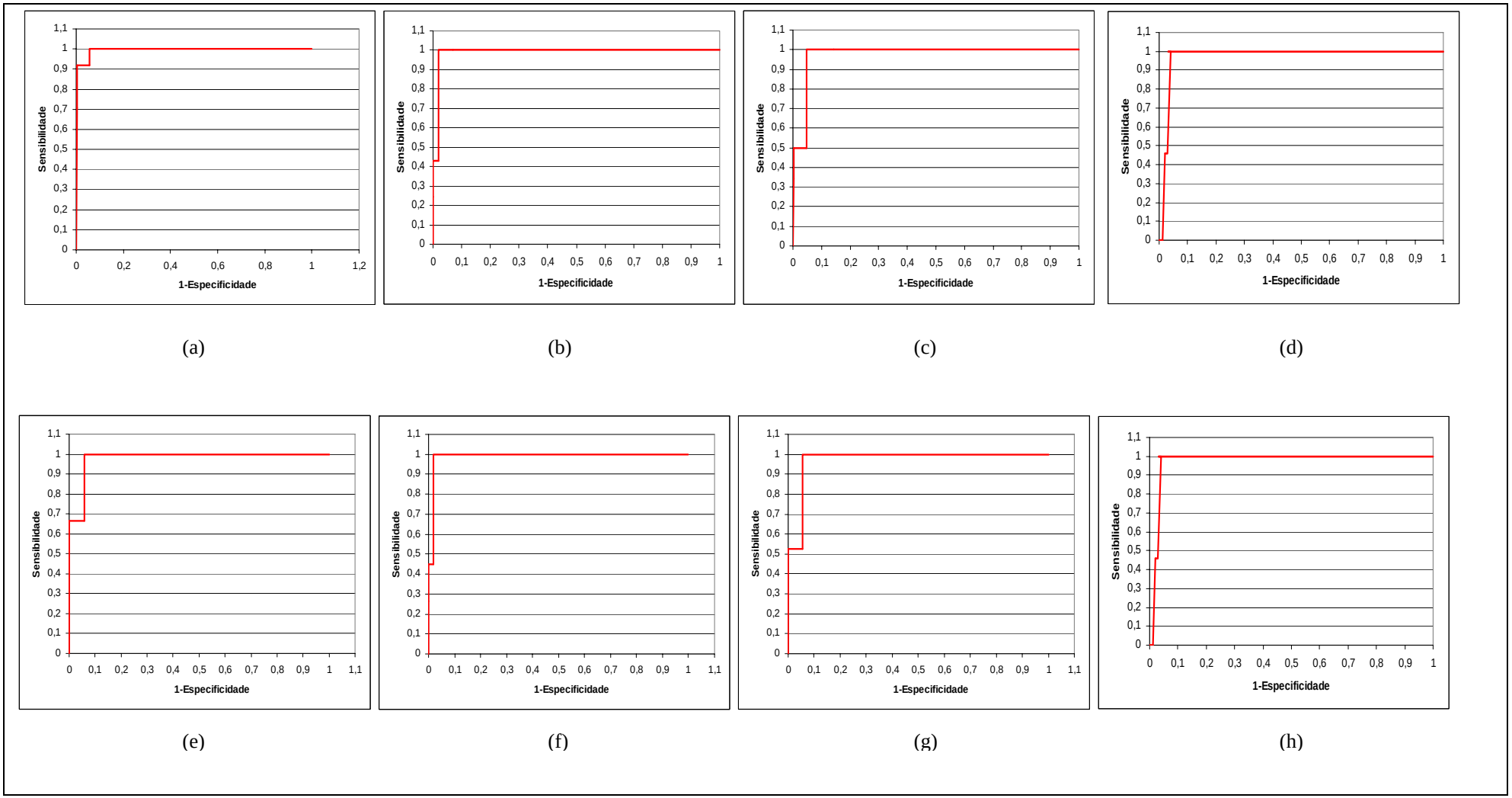

FIGURA 4.3 – ROC-plot do modelo Bioclim com erros projetados em centróides: (a) Erros de até 10 km (b) Erros de até 0,25º (c) Erros de até 0,5º e (d) Erros de até 1º, e com erros em coordenadas polares: (e) Erros de até 10 km (f) Erros de até 0,25º (g) Erros de até 0,5º e (h) Erros de até 1º.

A Figura 4.2 mostra como os erros de posicionamento alteram as saídas do Bioclim de acordo com os níveis de erros de posicionamento projetados em centróides de células e erros com parâmetros de coordenadas polares. Uma análise visual permite notar que erros a partir de 0,25º alteram provocam um aumento na proporção de área de ocorrência prevista, ou seja, há uma piora no desempenho do modelo.

Os ROC-plots da Figura 4.3 não permitem uma análise visual, apenas os valores de AUC contidos na Tabela 4.7 é que explicitam a variação no desempenho do Bioclim em função dos erros de posicionamento.

4.2 – GARP BEST SUBSETS

Uma crítica comum ao modelo GARP é que o algoritmo possui a característica de apresentar resultados diferentes para cada realização, com os mesmo parâmetros e mesmas variáveis (Anderson et al., 2003). Esta falta de convergência de resultados deve-se ao fato do algoritmo genético ser estocástico. Siqueira (2005) constatou que não existe uma falta de convergência significante nos resultados do GARP, e esse aspecto é importante quando são comparadas as medidas de desempenho dos modelos.

O GARP-BS é uma composição de um número de modelos determinados durante a escolha dos parâmetros de otimização (Seção 2.2.1). Dessa forma, os valores que são apresentados no mapa de saída representam o número de modelos que previram como positiva a ocorrência da espécie. Os valores mais altos são representados pela cores claras, i.e., um número maior de modelos indica a presença da espécie.

A área prevista pelo modelo resultante das amostras de treino (Figura 4.4) foi de 20,54% do total da área de estudo. Mesmo o modelo de referência do GARP *Best*

Subsets que foi gerado a partir de amostras sem erros prevê ocorrência sobre uma área muito maior do que a prevista pelo modelo de treino do Bioclim (6,91%).

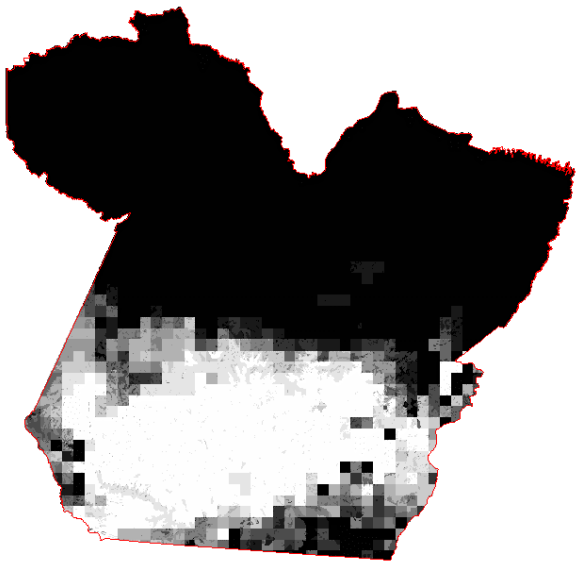

FIGURA 4.4 – Modelo OM-GARP *Best Subsets* com conjunto de amostras de treino

A Figura 4.5 mostra os resultados dos modelos gerados com erros. A avaliação deve ser realizada considerando-se que o número de amostras diminui conforme o erro aumenta (Tabela 3.3) visto que todas as ocorrências contidas em uma célula são projetadas para o mesmo centróide (Figura 3.5). O aumento da área de ocorrência é similar aos resultados de Manel et al. (2001) que obteve superestimativas de ocorrências para espécies com um número pequeno de amostras.

A Figura 4.7 contém os diagramas de dispersão dos resultados do modelo de treinamento versus os resultados com erros de posicionamento. Esses diagramas explicitam os valores para b1, onde fica claro que as saídas dos modelos são diferentes em relação à referência, o treinamento. Os erros em coordenadas polares apresentam um IC mais estreito que o IC dos erros em centróides. Isso pode ser confirmado na Tabela 4.1 e no diagrama de dispersão da Figura 4.7.

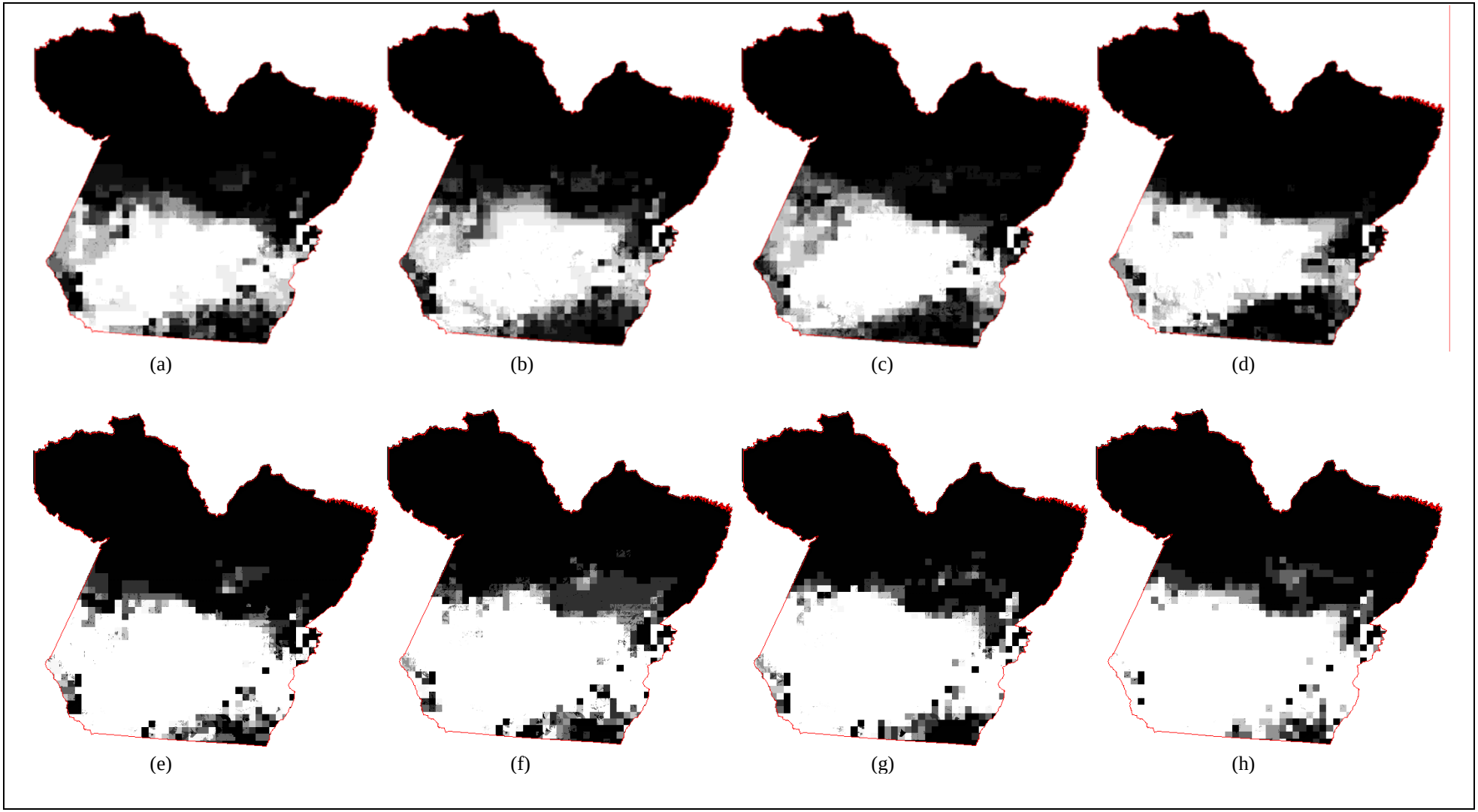

FIGURA 4.5 – Modelos GARP *Best Subsets* com erros projetados em centróides: (a) Erros de até 10 km (b) Erros de até 0,25º (c) Erros de até 0,5º e (d) Erros de até 1º, e com erros em coordenadas polares: (e) Erros de até 10 km (f) Erros de até 0,25º (g) Erros de até 0,5º e (h) Erros de até 1º.

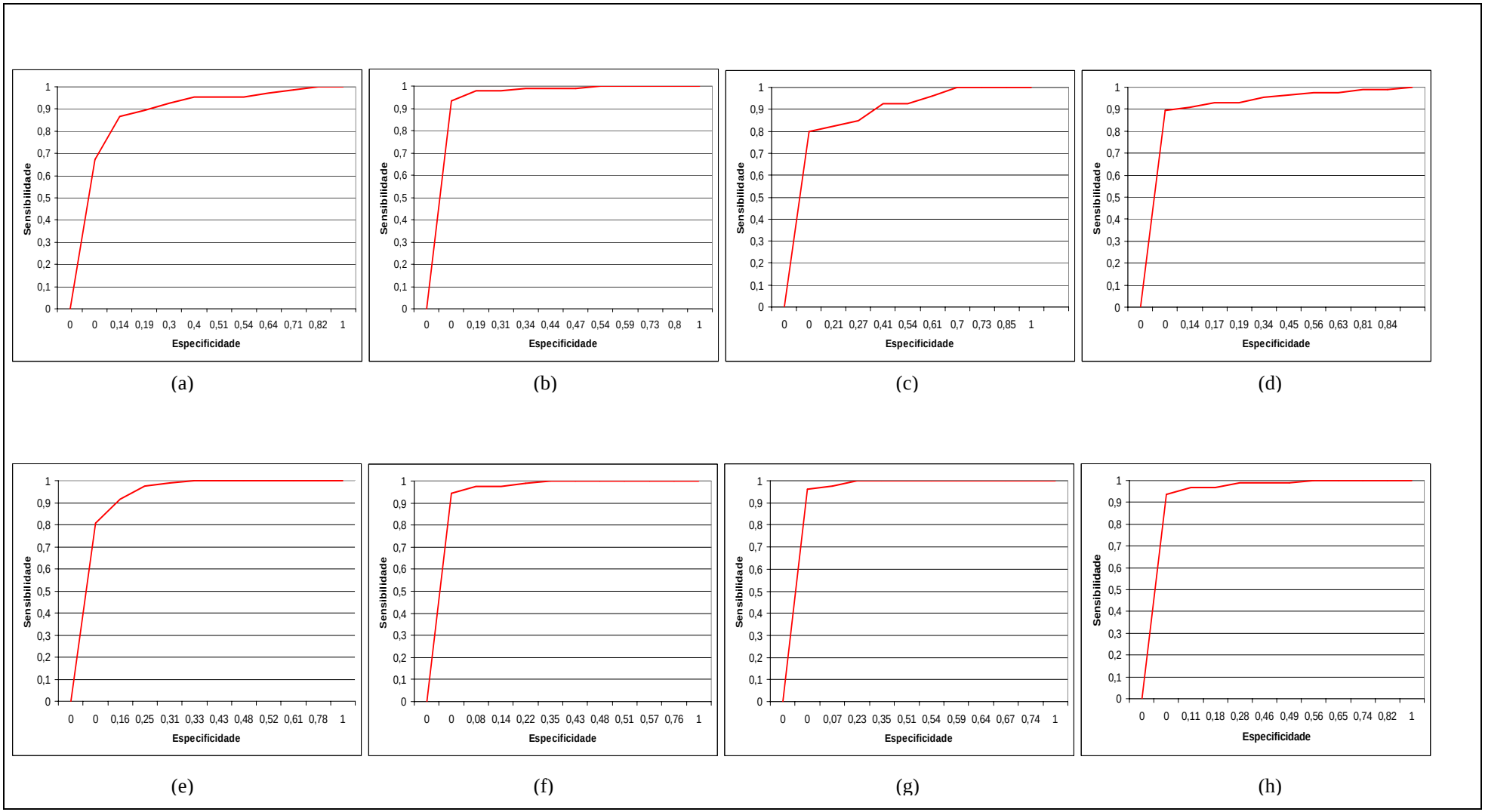

FIGURA 4.6 – ROC-plot do GARP *Best Subsets* com erros projetados em centróides: (a) Erros de até 10 km (b) Erros de até 0,25º (c) Erros de até 0,5º e (d) Erros de até 1º, e com erros em coordenadas polares: (e) Erros de até 10 km (f) Erros de até 0,25º (g) Erros de até 0,5º e (h) Erros de até 1º.

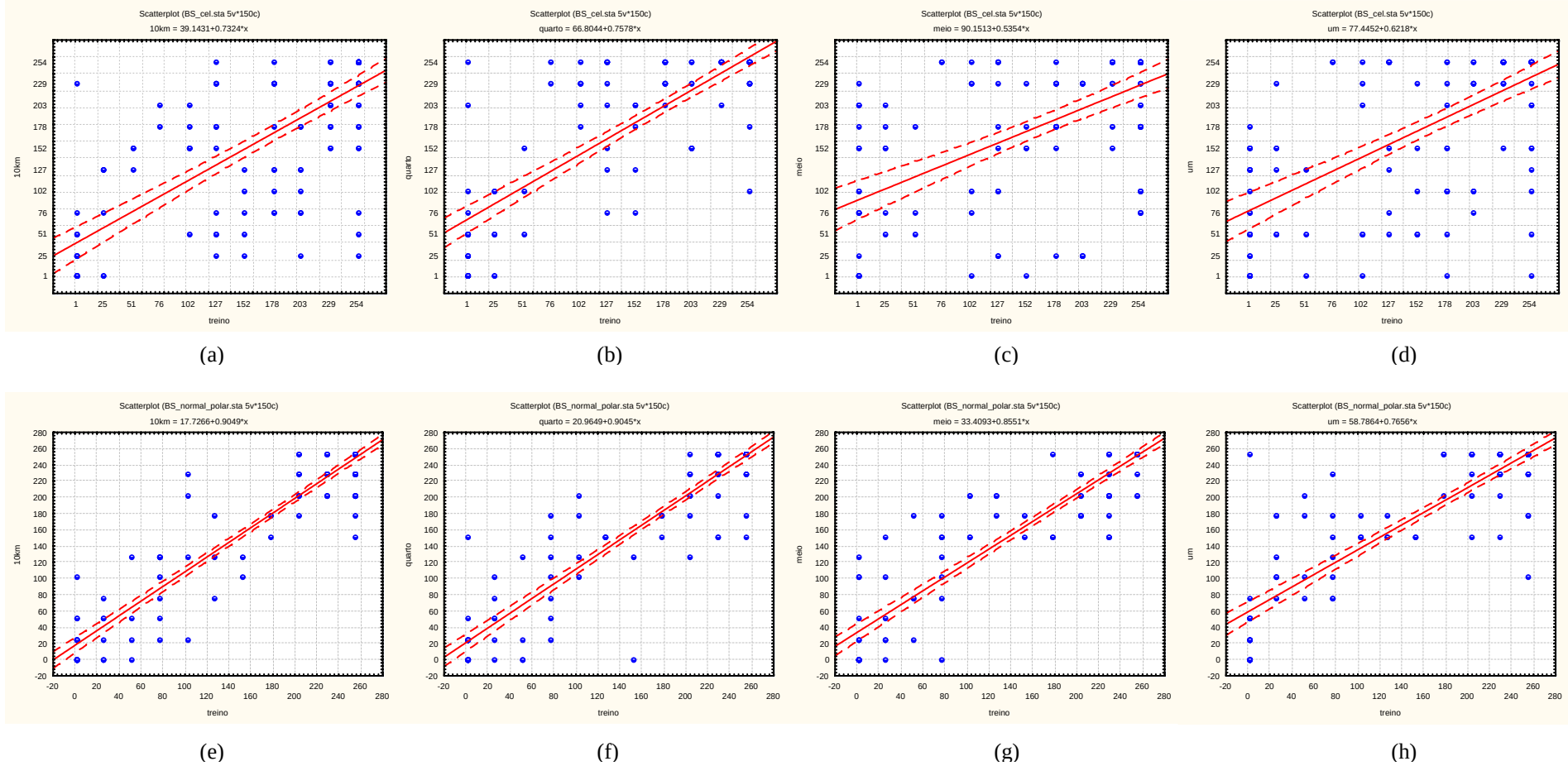

FIGURA 4.7 – Diagrama de dispersão dos erros em centróides de células do GARP. A amostra de treinos é a variável independente e os erros em coordenadas polares as variáveis dependentes (a) Erros de até 10 km (b) Erros de até 0,25º (c) Erros de até 0,5º e (d) Erros de até 1º.

A Tabela 4.1 possui os *p*-valores, a inclinação estimada *b*1 e o intervalo de confiança para a inclinação da reta. O *p*-valor refere-se ao teste cuja hipótese nula é $\beta_1 = 1$. De acordo com esta tabela, nenhum modelo apresentou similaridade de resultados significativa, ou seja, os erros de posicionamento alteraram as saídas de todos os modelos. Todos os *p*-valores são próximos de zero, indicando que β_1 é diferente de 1 a níveis de significância maiores que 99,9%, i.e. os resultados dos modelos não podem ser considerados iguais. O intervalo de confiança da Tabela 4.1 facilita a interpretação, onde a 95% de confiança, a inclinação da curva está contida nesses intervalos, e nenhum dos intervalos inclui um b1 = 1.

TABELA 4.1 – Teste para $\beta_1 = 1$ para o GARP

	Centróides				Cood. Polar			
	p	b1	IC-95%	IC+95%	*p*	b1	IC-95%	IC+95%
10 km	1,86E-06	0,732	0,633	0,831	1,54E-04	0,905	0,858	0,950
0,25º	3,52E-07	0,757	0,675	0,840	5,52E-04	0,905	0,853	0,955
0,5º	2,49E-10	0,535	0,417	0,653	1,27E-06	0,855	0,802	0,907
1º	1,24E-08	0,622	0,510	0,732	5,20E-10	0,766	0,704	0,826

4.3 Máxima entropia

Os resultados do modelo máxima entropia com as amostras de treinamento são apresentados na Figura 4.8. O mapa indica a probabilidade de ocorrência, onde os valores mais altos estão em cores quentes. As áreas com alta probabilidade em vermelho são similares à área predita pelo Bioclim e o ao nicho fundamental simulado.

A Figura 4.9 ilustra os resultados tabelados, expondo o modelo Maxent com erros projetados em centróides de células e com erros em coordenadas polares. As áreas em cores quentes representam uma probabilidade alta de ocorrência. Nesta

figura nota-se o aumento da área predita em vermelho, que é causado pelos erros de posicionamento.

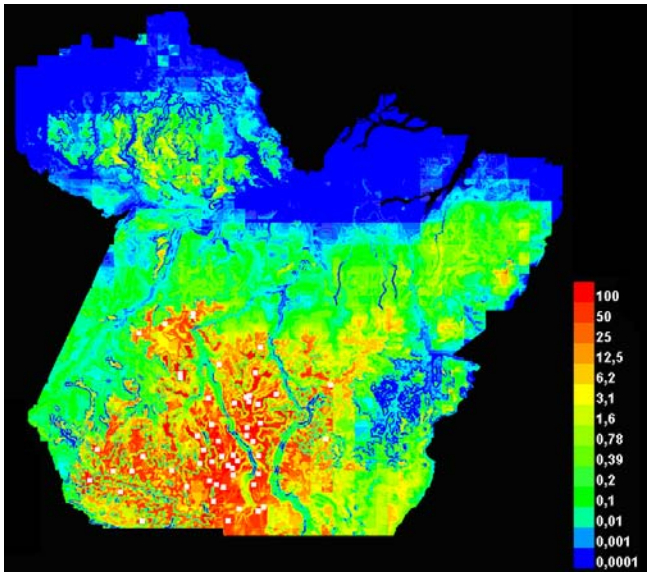

FIGURA 4.8 – Previsão de ocorrência do MAXENT com o conjunto de amostras de treino.

A Figura 4.11 representa o diagrama de dispersão que compara o resultado do modelo de treinamento da máxima entropia com as diversas escalas de erros. Nota-se um grande número de pontos próximos na origem, isso ocorre porque a máxima entropia realiza previsões em escala logarítmica. Não existe um valor zero no plano de informação resultante do modelo, mas a maioria da área de estudo possui probabilidades de ocorrência muito próximas de zero.

A Tabela 4.2 contém os *p*-valores para o teste de $\beta_1 = 1$ e o intervalo de confiança para b_1. Para erros de até 10 km e de até 0,25º a inclinação da reta é significativamente diferente de 1, a 95% de confiança. Em contrapartida as curvas dos erros maiores de 0,5º e 1º possuem uma inclinação significativa de 45º.

TABELA 4.2 – Teste para $\beta_1 = 1$ para o Maxent

	Centróides				Cood. Polar			
	p	b1	IC-95%	IC+95%	p	b1	IC-95%	IC+95%
10 km	0,002	0,832	0,726	0,937	0,008	0,856	0,752	0,959
0,25º	0,019	0,785	0,611	0,960	0,034	0,807	0,633	0,982
0,5º	0,289	0,894	0,700	1,089	0,358	0,908	0,713	1,103
1º	0,819	0,981	0,819	1,142	0,806	0,976	0,791	1,162

A projeção para o centróide das células gera o problema, já citado, de diminuição do número de amostras. A introdução de erros em coordenadas polares não traz esse problema. O GARP é robusto mesmo com um número pequeno de ocorrências (Stockwell e Peterson, 2002), mas ainda não existem trabalhos que avaliem a influência do tamanho da amostra no desempenho do Maxent.

O método de máxima entropia foi apresentado recentemente (Phillips et al., 2006) e ainda não possui muitos estudos sobre seu desempenho e incertezas. Contudo os resultados mostram que essa metodologia é uma boa alternativa quando os dados possuem erros de posicionamento.

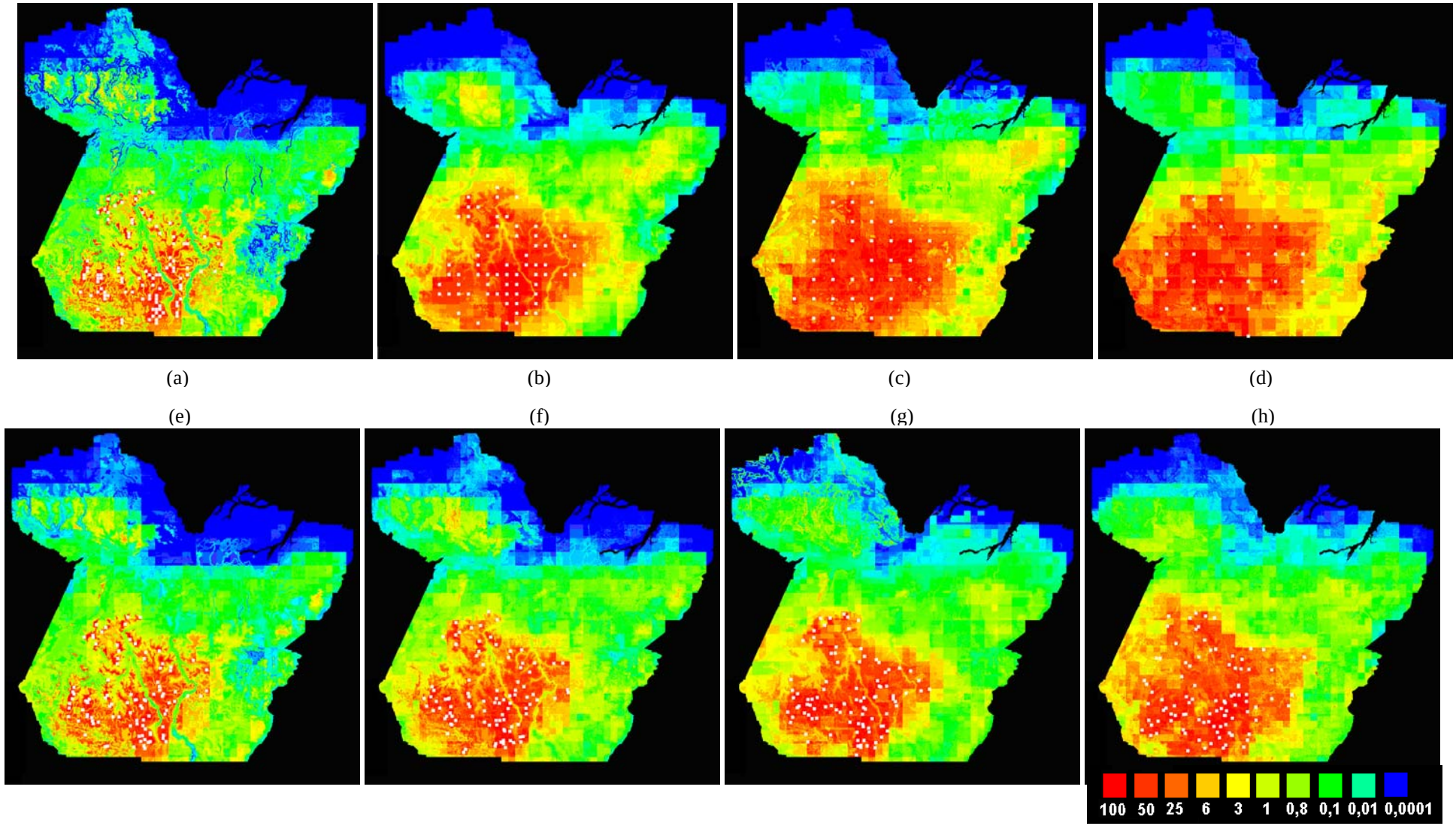

FIGURA 4.9 – Previsão de ocorrência com MAXENT com erros projetados em centróides (a) Erros de até 10 km (b) Erros de até 0,25º (c) Erros de até 0,5º e (d) Erros de até 1º.

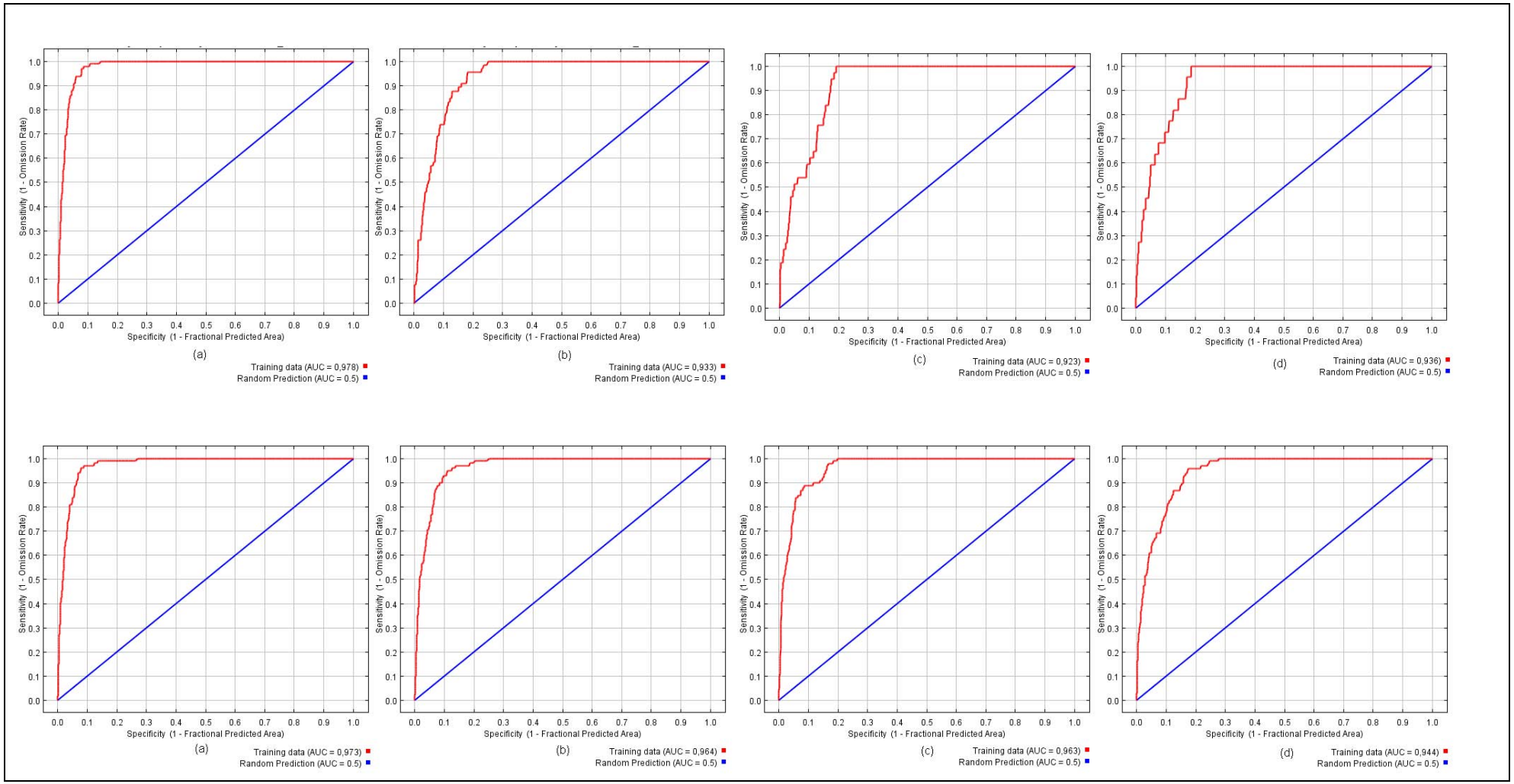

FIGURA 4.10 – ROC-plot e AUC do Maxent com erros projetados em centróides (a) Erros de até 10 km (b) Erros de até 0,25º (c) Erros de até 0,5º e (d) Erros de até 1º.

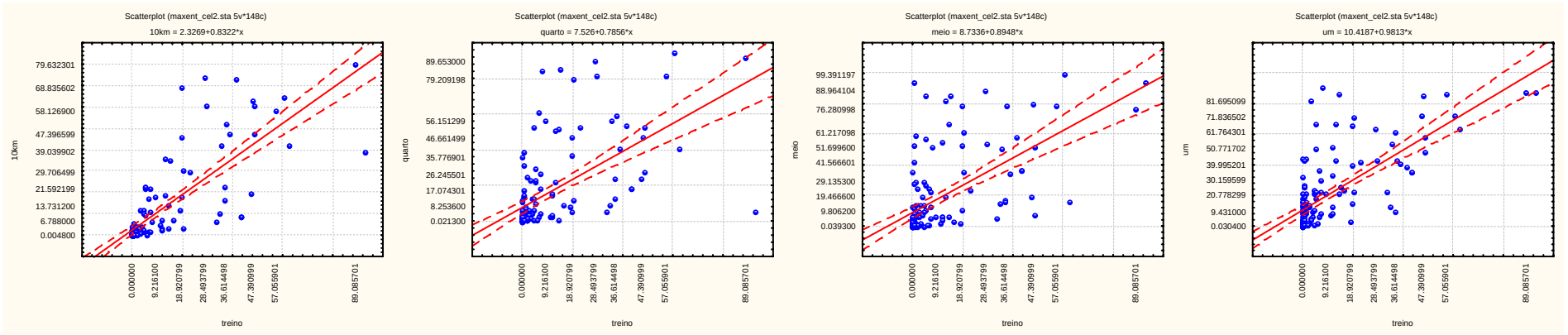

FIGURA 4.11 – Diagrama de dispersão dos erros em centróides de células do Maxent. A amostra de treinos é a variável independente e os erros em coordenadas polares as variáveis dependentes (a) Erros de até 10 km (b) Erros de até 0,25º (c) Erros de até 0,5º e (d) Erros de até 1º.

4.4. Métricas de avaliação

4.4.1. Omissão e comissão

O método mais simples de avaliação do desempenho de modelos de distribuição de espécies é a matriz de confusão (Manel et al., 2001), calculada a partir dos pontos de teste.

Os erros de comissão (Tabela 4.3) aumentam para todos os modelos conforme crescem os erros de posicionamento. Os erros de comissão podem não representar um erro real do modelo, mas está correlacionado com o aumento área de ocorrência, representando para este trabalho uma queda no desempenho dos modelos.

TABELA 4.3 – Erros comissão (%)

Comissão						
	Bioclim		**GARP**		**Maxent**	
	Centróide	Polar	Centróide	Polar	Centróide	Polar
10 km	3,33	10	2,67	4	4,83	4,76
0,25º	24	23,33	18,67	6,67	11,04	10,88
0,5º	25,33	26	10,67	2,67	12,41	12,25
1º	24	32	15,33	10,67	14,48	14,29

Os erros de omissão (Tabela 4.4) crescem apenas para o Bioclim e para o Maxent. Essa verificação ocorre tanto quando as ocorrências são projetadas em centróides de células quanto com erros em coordenadas polares.

Quanto aos erros de omissão do GARP-BS, Anderson et al. (2003) obtiveram resultados similares, detectaram que os erros de omissão são inversamente proporcionais aos erros de comissão, quando os erros de omissão são baixos, os erros de comissão tendem a ser altos.

TABELA 4.4 – Erros de omissão (%)

Omissão						
	Bioclim		**GARP**		**Maxent**	
	Centróide	Polar	Centróide	Polar	Centróide	Polar
10 km	0	0	14,67	10,67	4,14	3,4
0,25º	0,67	0,67	4	3,33	5,52	4,76
0,5º	4	1,33	10,67	2	4,83	4,76
1º	6	2,11	6	4	2,07	2,04

A análise dos erros de omissão e comissão apresenta resultados semelhantes aos de Phillips et al. (2006), que compararam os desempenhos do GARP e do Maxent e obtiveram menores erros de omissão e de áreas preditas para o Maxent.

O Bioclim mostrou um aumento dos erros de comissão e dos erros de omissão, apresentando a maior queda no desempenho. Considerando estas métricas não é possível analisar a sensibilidade do GARP-BS aos erros de posicionamento, pois os erros de omissão, considerados mais importantes, não apresentam tendências de aumento ou queda. Os erros de comissão e omissão no Maxent permanecem menores que os do Bioclim em todas as escalas. Portanto, se forem considerados apenas os erros de omissão e comissão é possível concluir, talvez equivocadamente, que o Maxent possui menor sensibilidade a erros de posicionamento.

4.4.2. Área mínima

Considerando-se a área mínima prevista (Tabela 4.5) dos modelos Bioclim e Maxent, o conjunto de treinamento, i.e., sem erros de posicionamento apresenta o menor valor de área se comparados aos modelos com erros de posicionamento: Bioclim, 6,91% e Maxent 2,94 %. Em contrapartida o modelo de treinamento do GARP-BS apresentou uma previsão de ocorrência de 20,54% do total da área de estudo, e o modelo com erros de 0,5º em centróides de 19,61%, e que de acordo

com Engler et al. (2004), uma menor área prevista significa um melhor desempenho.

Para o Bioclim e o Maxent, modelos com erros de até 10 km geram uma área de ocorrência similar em forma e proporção aos modelos de treinamento. A partir de erros de 0,25º a área prevista pelo Bioclim aumenta de 7,45/8,24% a 10 km para 11,00/12,17% a 0,25º. O Maxent também apresenta essa piora no desempenho com erros acima de 10 km.

Em uma análise descritiva, o GARP-BS não apresenta um aumento da área prevista, sendo neste aspecto o modelo que apresenta menor sensibilidade a erros de posicionamento. Esses resultados se opõem aos obtidos com os erros de omissão e comissão e reforçam a idéia que não pode existir uma única medida para a avaliação do modelo (Fielding e Bell, 1997).

TABELA 4.5 – Área de ocorrência prevista (%)

Área mínima						
	Bioclim		**GARP**		**Maxent**	
	Centróide	Polar	Centróide	Polar	Centróide	Polar
10 km	7,45	8,24	20,2	21,08	3,23	2,68
0,25º	11	12,17	19,9	21,56	4,55	6,57
0,5º	10,05	12,57	19,61	22,47	5,31	8,38
1º	10,01	16,73	21,62	23,57	6,6	9,16

4.4.3. Índice kappa

Todos os valores do índice Kappa obtidos da modelagem do Bioclim são classificados como razoáveis segundo a Tabela 2.3. Contudo essa classificação do índice Kappa é referencial, pois há piora do índice com o aumento dos erros de localização. O índice sofre uma queda de 0,5036/0,4456 a 10 km para 0,3502/0,3474 a 0,25º.

Com erros de posicionamento de até 10 km o Bioclim apresenta o melhor índice Kappa. Entretanto o índice Kappa tem uma queda maior no Bioclim quando os erros aumentam de 10 km para 0,25º, enquanto no GARP os valores do índice Kappa continuam próximos entre si (Tabela 4.6). O Maxent se encontra em uma situação intermediária, não possui o melhor kappa e exibe uma queda no índice para erros acima de 10 km, mas para erros de posicionamento maiores (0,25º; 0,5º e 1º) o índice apresenta pouca variação.

.TABELA 4.6 – Índice kappa

	Kappa					
	Bioclim		**GARP**		**Maxent**	
	Centróide	Polar	Centróide	Polar	Centróide	Polar
10 km	0,5036	0,4456	0,4121	0,4462	0,3994	0,4045
0,25º	0,3502	0,3474	0,3909	0,4053	0,3384	0,3442
0,5º	0,3051	0,3229	0,3841	0,3912	0,3381	0,3352
1º	0,2916	0,3004	0,3893	0,395	0,3304	0,3276

Considerando apenas o índice Kappa, o Bioclim apresenta o maior valor apenas nos erros de 10 km, a Máxima entropia encontra-se em uma situação intermediária e o GARP apresenta pouca variação nos índices, podendo ser considerado o menos sensível a erros de posicionamento.

4.4.4. ROC-plot

O ROC-plot fornece a área sob a curva (*Area Under the Curve* – AUC), quanto mais próximo de um for a AUC, melhor o desempenho do modelo. As medidas da Tabela 4.7 são consideradas indicações de bons desempenhos para os modelos, pois Elith et al. (2006) avaliaram 226 espécies e obtiveram médias que variam de 0,69 a 0,8 para a AUC.

Para o Bioclim e Maxent, esta métrica apresenta o mesmo padrão que as outras medidas de avaliação. Modelos com erros de até 10 km têm um bom

desempenho, e a partir de erros de 0,25º a medida sofre uma queda. Apesar da pequena diferença nos valores, conforme aumentam os erros de posicionamento, há um decréscimo na AUC.

TABELA 4.7 – Área sob a curva do ROC-plot

	AUC					
	Bioclim		**GARP**		**Maxent**	
	Centróides	Polar	Centróides	Polar	Centróides	Polar
10 km	0,993	0,980	0,923	0,959	0,978	0,973
0,25º	0,975	0,979	0,912	0,941	0,933	0,964
0,5º	0,978	0,973	0,907	0,933	0,923	0,963
1º	0,925	0,955	0,893	0,925	0,936	0,944

A área sob a curva do ROC-plot para o GARP *Best Subsets* exibe menor queda que os modelos Bioclim e Maxent, sendo considerado o modelo mais resiliente. Contudo, os desempenhos do Bioclim e do Maxent, com erros de 1º, estão acima da AUC do GARP com erros de 10 km.

Todas as métricas de avaliação do Bioclim mostram um bom desempenho do modelo com erros de até 10 km. Entretanto, o Bioclim mostra uma queda de desempenho a partir de erros de 0,25º, onde todas as medidas de avaliação pioram.

O GARP-BS mostra menor variação para área mínima prevista, kappa e ROC-plot e pode ser considerado o modelo mais robusto. Porém, quando comparado ao modelo de treinamento o desempenho do modelo é baixo nestas três métricas.

O Maxent é menos sensível a erros de posicionamento quando se considera apenas os erros de omissão e comissão. Para a área mínima, o kappa e o ROC-plot, o modelo apresentou uma sensibilidade a erros intermediária.

5 CONCLUSÕES

Apesar dos inúmeros trabalhos que avaliam os mais diversos aspectos dos modelos de distribuição de espécies, os erros de posicionamento não foram considerados em nenhum deles. Esta dissertação foi elaborada para que os erros de posicionamento e suas conseqüências sobre os resultados dos modelos sejam incorporados no processo de modelagem de distribuição de espécies

Os erros de posicionamento em pontos de ocorrência e sua influência nos modelos de distribuição de espécies foram avaliados através de dados artificiais. Foram simulados o nicho fundamental e os pontos de ocorrência de uma espécie vegetal hipotética. Dois métodos de introdução de erros foram utilizados, a projeção das coordenadas das amostras para centróides de células e erros com distribuição normal com parâmetros em coordenadas polares.

Todos os modelos analisados apresentam sensibilidade aos erros de posicionamento. Para fins de aplicações, o principal problema detectado foi um aumento de área de ocorrência prevista pelos modelos.

Da análise dos resultados, ressalta-se a necessidade de mais de uma métrica de avaliação. As métricas, assim como os modelos, possuem sensibilidade e restrições de uso.

São necessárias métricas de avaliação diversas para os sistemas e algoritmos que modelam a distribuição de espécies para evitar conclusões equivocadas. O modelo GARP *Best Subsets*, por exemplo, apresenta baixos erros de omissão e comissão, e sob o critério de área predita tem o desempenho mais baixo entre os três métodos avaliados, mas é considerado o mais resiliente a erros de posicionamento.

O Bioclim mostra-se sensível a erros iguais ou maiores que um quarto de grau (aproximadamente 27,75 km na linha do Equador). A superestimativa no caso do Bioclim torna impraticável sua utilização para seleção de áreas prioritárias para conservação (Araújo et al., 2004; Polasky e Solow, 2001). Este problema inviabiliza também sua aplicação em trabalhos que analisam os impactos de mudanças climáticas (Randin et al., 2006) sobre os padrões de distribuição de espécies.

O GARP-BS apresenta uma superestimativa mesmo considerando-se o resultado do modelo de treinamento comparado ao nicho simulado. Quando o modelo de treinamento é comparado aos resultados dos modelos com erros o GARP-BS apresenta robustez se o número de amostras decresce (no caso da projeção em centróides), e até mesmo quando a dimensão dos erros aumenta.

O modelo de máxima entropia também apresenta um aumento de área de ocorrência prevista nos modelos gerados com amostras que possuem erros de posicionamento. Porém, as áreas com alta probabilidade de ocorrência dos modelos com erros se aproximam mais do modelo de treinamento (quando os erros têm distribuição normal) se comparados ao GARP e ao Bioclim.

Este trabalho demonstra a importância do analista considerar que os erros de posicionamento influenciam os resultados da modelagem. Como a maioria dos dados de ocorrência disponíveis no Brasil provém de herbários ou museus, a utilização de dados com erros é quase inevitável. É necessário conhecer os níveis de erros de posicionamento dos dados que serão utilizados na modelagem. Para modelar distribuição de espécies com dados passíveis de erros de localização deve-se considerar principalmente que:

- Se os erros forem de até 10 km é possível aplicar qualquer um dos modelos sem grande perda no desempenho do modelo;

- Admitindo como premissa da modelagem que os erros possuem uma distribuição normal há um ganho ao utilizar o Maxent que apresenta uma superestimativa menor;

- Quando os erros são de dimensão desconhecida e/ou se o conjunto de amostras for pequeno recomenda-se o uso do GARP-BS, sob a mesma premissa de normalidade dos erros e consciente que o GARP-BS apresenta os maiores valores de área predita.

A principal contribuição deste trabalho, além dos resultados discutidos, reside na metodologia desenvolvida de simulação de nicho fundamental e de erros de posicionamento. Novos testes em diferentes escalas e com diferentes padrões de distribuição poderão ser realizados. Nesse sentido, como trabalhos futuros recomendam-se:

- O emprego de medidas locais para avaliação de desempenho de modelos;

- Verificar a influencia de erros de posicionamento em diferentes padrões de distribuição (restrito e disjunto);

- Trabalhar com resoluções espaciais mais finas analisando a influência da autocorrelação espacial das variáveis sobre os modelos.

REFERÊNCIAS BIBLIOGRÁFICAS

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, Apr. 2003.

ARAÚJO, M. B.; CABEZA, M.; THUILLER, W.; HANNAH, L.; WILLIAMS, P. H. Would climate change drive species out of reserves? An assessment of existing reserve-selection methods. **Global Change Biology**, v. 10, n. 9. p. 1618-1626, Sept. 2004.

ARAÚJO, M. B.; GUISAN, A. Five (or so) challenges for species distribution modelling. **Journal of Biogeography**, v. 33, n.10. p. 1677-1688, Oct. 2006.

ARAÚJO, M. B.; WILLIAMS, P. H. Selecting areas for species persistence using occurence data. **Biological Conservation**, v. 96, n. 3. p. 331-345, Dec. 2000.

AUSTIN, M. P., BELBIN, L. MEYERS, J. A., DOHERTY, M. D., LUOTO, M. Evaluation of statistical models used for predicting plant species distributions: Role of artificial data and theory. **Ecological Modelling**, v. 199, n. 2. p. 197-216, Nov. 2006.

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, Nov. 2002.

BARRY, S.; ELITH, J. Error and uncertainty in habitat models. **Journal of Applied Ecology**, v. 43, n. 3. p. 413-423, June. 2006.

BAZZAZ, F. A. **Plants in changing environments**: Linking physiological, population, am community ecology. Cambridge: Cambridge University Press, 1998.

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, Aug. 2005.

BONACORSO E.; KOCH, I.; PETERSON, A. T. Pleistocene fragmentation of Amazon species' ranges. **Diversity and Distributions**. v. 12, n. 2. p. 157-164, Mar. 2006.

Brown, J. H., Gibson, A. C. **Biogeography**. Missouri: Mosby Company, 1983.

CARPENTER, G.; GILLISON, A. N.; WINTER, J. DOMAIN: a flexible modeling procedure for mapping potential distributions of plants, animals. **Biodiversity and Conservation**, v. 2, n. 6. p. 667-680, Dec. 1993

CHAPMAN, A. D.; MUÑOZ, M. E. S.; KOCH, I. Environmental information: Placing biodiversity phenomena in a ecological and environmental context. **Biodiversity Informatics**, v. 2, p. 24-41, 2005.

COLLINGHAM, Y. C.; WADSWORTH, R. A.; HUNTLEY, B.; HULME, P. E. Predicting the spatial distribution of non-indigenous riparian weeds: issues of spatial scale and extent. **Journal of Applied Ecology**, v. 37, n. 1. p. 13-27, Feb. 2000.

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

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, Apr. 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, Apr. 2004.

EDWARDS JR. T. C.; CUTLER, D. R.; ZIMMERMMANN, N. E.; GEISER, L.; MOISEN, G. G. Effects of sample survey design on the accuracy of classification tree models in species distribution models. **Ecological Modelling**, v. 199, n. 2. p. 132-141, Nov. 2006.

ESCADA, M. I. S.; VIEIRA, I. C. G.; KAMPEL, S. A.; ARAÚJO, R.; VEIGA, J. B. D.; AGUIAR, A. P. D.; VEIGA, I.; OLIVEIRA, M.; PEREIRA, J. L. G.; FILHO, A. C.; FEARNSIDE, P. M.; VENTURIERI, A.; CARRIELLO, F.; THALLES, M.; CARNEIRO, T. S. G.; MONTEIRO, A. M. V.; CÂMARA, G. Processos de ocupação nas novas fronteiras da Amazônia. **Estudos Avançados**, v. 19, n. 54, p. 9-23, 2005.

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. 1, p. 38-49, Mar. 1997.

GRINNELL, J. Field tests of theories concerning distributional control. **American Naturalist** v. 51, p. 115-128, 1917.

GUISAN, A.; THOMAS C. EDWARDS, J.; HASTIE, T. Generalized linear and generalized additive models in studies of species distributions: setting the scene. **Ecological Modelling**, v. 157, n.2. p. 89-100, Nov. 2002.

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

GUISAN, A.; WEISS, S. B.; WEISS, A. D. GLM versus CCA spatial modeling of plants species distributions. **Plant Ecology**, v. 143, n. 1. p. 107-122, July. 1999.

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

HIJMANS, R.J., S.E. CAMERON, J.L. PARRA, P.G. JONES AND A. JARVIS. Very high resolution interpolated climate surfaces for global land areas. **International Journal of Climatology,** v. 25, n. 15. p. 1965-1978, Dec. 2005

HIRZEL, A.; GUISAN, A. Which is the optimal sampling strategy for habitat suitability modelling. **Ecological Modelling**, v. 157, n. 3. p. 331-341, Nov. 2002.

HIRZEL, A. H.; HELFER, V.; METRAL, F. Assessing habitat-suitability models with a virtual species. **Ecological Modelling**, v. 145, n. 2. p. 111-121, Nov. 2001

HIRZEL, A. H.; HAUSSER, J.; CHESSEL, D.; PERRIN, N. Ecological-niche factor analysis: how to compute habitat suitability maps without absence data? **Ecology**, v. 83, n. 7. p. 2027-2036, Jul. 2002.

HUTCHINSON, G. E. Concluding Remarks. Cold Spring Harbor Symp. **Quantitative Biol.**, v. 22, p. 415-427, 1957.

ISAAKS, E. H; SRIVASTAVA R. M. **Applied Geostatistics**. 2ª ed. Oxford: Oxford University Press, 1989. 561p.

Kansas applied remote sensing program. Ecological Niche Modelling: Inter-model Variation. Best-subset Models Selection. *In* **Global Biodiversity Information Facility Data Modelling Workshop**. Mexico City – Faculty of Sciences of UNAM University. 2005

LANDIS, J.R.; KOCH, G.G. The measurement of observer agreement for categorical data. **Biometrics**, v.33, n.1, p.159-174, Mar. 1977.

LARCHER, W. **Ecofisiologia vegetal**. São Carlos: RiMa – Artes e Textos. 2000. 531p.

LASSUEUR, T.; JOOST, S.; RANDIN, C. F. Very high resolution digital elevation models: Do they improve models of plants species distributions? **Ecological Modelling**, v. 198, n. 2. p. 139-153, Sep. 2006.

LOCATELLI, M.; SILVA FILHO, E. P.; VIEIRA, A. H.; MARTINS, P.; PEQUENO, P. L. L. Castanha do Brasil – Opção para solo de baixa fertilidade na Amazônia. In: SEMINÁRIO NACIONAL DE DEGRADAÇÃO E RECUPERAÇÃO AMBIENTAL, 2003, Foz do Iguaçu. **Anais...** Foz do Iguaçu: [s.n], 2003.

LUOTO, M.; PÖYRY, J.; HEIKKINEN, R. K.; SAARINEN, K. Uncertainty of bioclimate envelope models based on the geographical distribuition of species. **Global Ecology and Biogeography**, v. 14, n. 5. p. 575-584, Sep. 2005.

MANEL, S.; WILLIAMS, H. C.; ORMEROD, S. J. Evaluating presence-absence models in ecology: the need to account for prevalence. **Journal of Applied Ecology**, v. 38, n. 5. p. 921-931, Oct. 2001.

MILLER, J.; FRANKLIN, J. Modeling the distribution of four vegetation alliances using generalized linear models and classification tress with spatial dependence. **Ecological Modelling**, v. 157, n. 3. p. 227-247, Nov. 2002.

NETER, J.; KUTNER, M. N.; NACHTSSHEIM, C. J.; WASSERMAN, W. **Applied linear statistical models.** Boston: WCB/McGraw-Hill, 4ª Ed. 1996, 791 p.

OKSANEN, J.; MINCHIN, P. R. Continuum theory revisited: what shapes are species responses along ecologial gradients? **Ecological Modelling**, v. 157, n. 2. p. 119-129, Nov. 2002.

PEARSON, R. G.; THUILLER, W.; ARAÚJO, M. B.; MARTINEZ-MEYER, E.; BROTONS, L.; MCCLEAN, C. J.; MILES, L.; SEGURADO, P.; DAWSON, T. P.; LEES, D. C. Model-based uncertainty in species range prediction. **Journal of Biogeography**, v. 33, n. 10. 1704-1711, Oct. 2006.

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

POLASKY, S.; SOLOW, A. R. The value of information in reserve site selection. **Biodiversity and Conservation**, v. 10, n. 7. p. 1051-1058, July. 2001.

PONTIUS Jr., G. Quantification error versus location error in comparison of categorical maps **Photogrammetric engineering and remote sensing**, v. 66, n. 8. p. 1011-1016, Aug, 2000.

PONTIUS Jr., G.; Schneider, L. C. Land-cover change model validation by an ROC method for the Ipswich watershed, Massachusetts, USA. **Agriculture, Ecosystems and Environment**. v. 85, n. 1. p. 239-248, Jun. 2001.

Instituto Nacional de Pesquisas Espaciais. Coordenação-Geral de Observação da Terra (INPE/OBT). **Projeto Prodes** – monitoramento da Floresta Amazônica Brasileira por satélite. São José dos Campos, 2006. Disponível em: <www.obt.inpe.br/prodes>. Acesso em Dez. de 2006.

RANDIN, C. F.; DIRNBÖCK, T.; DULLINGER, S.; ZIMMERMANN, N. E.; ZAPPA, M.; GUISAN, A. Are niche-based species distributions models transferable in space? **Journal of Biogeography**, v. 33, n. 10. p. 1689-1703, Oct. 2006.

RAVEN, P. H; EVERT, R. F; EICHHORN, S. E. **Biologia Vegetal**. 6ª ed. Rio de Janeiro: Guanabara Koogan, 2001.

REDDY, S.; DÁVALOS, L. M. Geographical sampling bias and its implications for conservation priorities in Africa. **Journal of Biogeography**, v. 30, n. 11. p. 1719-1727, Nov. 2003.

ROSSATO, L.; ALVALÁ, R. C. S.; TOMASELLA, J. Variação espaço-temporal da umidade do solo no Brasil: Análise das condições médias para o período de 1971-1990. **Revista Brasileira de Meteorologia**, v.19, n. 2, p. 113-122, 2004.

RUSHTON, S. P.; ORMEROD, S. J.; KERBY, G. New paradigms for modelling species distributions? **Journal of Applied Ecology**, v. 41, n. 2. 193-200, Apr. 2004.

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

SEGURADO, P.; ARAÚJO, M. B.; KUNIN, W. E. Consequences of spatial autocorrelation for niche-based models. **Journal of Applied Ecology**, v. 43, n. 3. p. 433-44, June. 2006.

SIQUEIRA, M. F. d. **Uso de modelagem de nicho fundamental na avaliação do padrão de distribuição geográfica de espécies vegetais**. 107 p. Tese de doutorado (Escola de Engenharia de São Carlos da Universidade de São Paulo), São Carlos, 2005.

STOCKWELL, D. **Desktop GARP**: Users manual. 19 pp. Disponível em: <www.lifemapper.org/desktopgarp>. Acesso em 10 de Dez. de 2006.

STOCKWELL, D.; PETERS, D. The GARP modelling system: problems and solutions to automated spatial prediction. **International Journal of Geographical Information Science**, v. 13, n. 2, p. 143-158, 1999.

STOCKWELL, D. ; PETERSON, A. T. Effects of sample size on accuracy of species distribution models. **Ecological Modelling**, v. 148, n. 1, p. 1-13, Feb, 2002.

STOCKWELL, D. ; BEACH, J. H.; STEWART, A.; VORONTSOV, G.; VIEGLAIS, D.; PEREIRA, R. S. The use of the GARP genetic algorithm and Internet grid computing in the Lifemapper world atlas of species biodiversity. **Ecological Modelling**, v. 195, n. 1-2, p. 139-145, June, 2005.

TOBLER, M.; HONORIO, E.; JANOVEC, J.; REYNEL, C. Implications of collection patterns of botanical specimens on their usefulness for conservation planning: an example of two neotropical plant families (Moraceae and Myristicaceae) in Peru. **Biodiversity and Conservation**, v. 16, n. 3. p. 659-677, Mar. 2007.

VARGAS, J. H.; CONSIGLIO, T.; JØRGENSEN, P. M.; CROAT, T. B. Modelling distribution patterns in a species-rich plant genus,Anthurium (Araceae), in Ecuador. **Diversity and Distributions**, v. 10, n. 3. p. 211-216, May 2004.

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

ZANIEWSKI, A. E.; LEHMANN, A.; OVERTON, J. M. Predicting species spatial distributions using presence-olny data: a case study of native New Zealand ferns. **Ecological modelling**, v. 157, n.2 p. 261-280, Nov, 2002.

PUBLICAÇÕES TÉCNICO-CIENTÍFICAS EDITADAS PELO INPE

Teses e Dissertações (TDI)

Teses e Dissertações apresentadas nos Cursos de Pós-Graduação do INPE.

Manuais Técnicos (MAN)

São publicações de caráter técnico que incluem normas, procedimentos, instruções e orientações.

Notas Técnico-Científicas (NTC)

Incluem resultados preliminares de pesquisa, descrição de equipamentos, descrição e ou documentação de programa de computador, descrição de sistemas e experimentos, apresentação de testes, dados, atlas, e documentação de projetos de engenharia.

Relatórios de Pesquisa (RPQ)

Reportam resultados ou progressos de pesquisas tanto de natureza técnica quanto científica, cujo nível seja compatível com o de uma publicação em periódico nacional ou internacional.

Propostas e Relatórios de Projetos (PRP)

São propostas de projetos técnico-científicos e relatórios de acompanhamento de projetos, atividades e convênios.

Publicações Didáticas (PUD)

Incluem apostilas, notas de aula e manuais didáticos.

Publicações Seriadas

São os seriados técnico-científicos: boletins, periódicos, anuários e anais de eventos (simpósios e congressos). Constam destas publicações o Internacional Standard Serial Number (ISSN), que é um código único e definitivo para identificação de títulos de seriados.

Programas de Computador (PDC)

São a seqüência de instruções ou códigos, expressos em uma linguagem de programação compilada ou interpretada, a ser executada por um computador para alcançar um determinado objetivo. São aceitos tanto programas fonte quanto executáveis.

Pré-publicações (PRE)

Todos os artigos publicados em periódicos, anais e como capítulos de livros.