Ministério da
Ciência e Tecnologia

INPE-15734-TDI/1480

SAHGA - UM ALGORITMO GENÉTICO HÍBRIDO COM REPRESENTAÇÃO EXPLÍCITA DE RELACIONAMENTOS ESPACIAIS PARA ANÁLISE DE DADOS GEOESPACIAIS

Adair Santa Catarina

Tese de Doutorado em Computação Aplicada, orientada pelos Drs. Antônio Miguel Vieira Monteiro e João Ricardo de Freitas Oliveira, aprovada em 8 de abril de 2009.

Registro do documento original:
<http://urlib.net/sid.inpe.br/mtc-m18@80/2009/03.19.13.53>

INPE
São José dos Campos
2009

PUBLICADO POR:

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) 3945-6911/6923
Fax: (012) 3945-6919
E-mail: pubtc@sid.inpe.br

Ministério da
Ciência e Tecnologia

INPE-15734-TDI/1480

SAHGA - UM ALGORITMO GENÉTICO HÍBRIDO COM REPRESENTAÇÃO EXPLÍCITA DE RELACIONAMENTOS ESPACIAIS PARA ANÁLISE DE DADOS GEOESPACIAIS

Adair Santa Catarina

Tese de Doutorado em Computação Aplicada, orientada pelos Drs. Antônio Miguel Vieira Monteiro e João Ricardo de Freitas Oliveira, aprovada em 8 de abril de 2009.

Registro do documento original:
<http://urlib.net/sid.inpe.br/mtc-m18@80/2009/03.19.13.53>

INPE
São José dos Campos
2009

Dados Internacionais de Catalogação na Publicação (CIP)

S59s

Santa Catarina, Adair.

SAHGA - Um algoritmo genético híbrido com representação explícita de relacionamentos espaciais para análise de dados geoespaciais / Adair Santa Catarina. – São José dos Campos : INPE, 2009.

120p. ; (INPE-15734-TDI/1480)

Dissertação (Mecânica Espacial e Controle) – Instituto Nacional de Pesquisas Espaciais, São José dos Campos, 2009.

Orientadores : Drs. Antônio Miguel Vieira Monteiro e João Ricardo de Freitas Oliveira.

1. Geocomputação. 2. Distribuição geográfica. 3. Recozimento simulado. 4. Sistemas de Informação Geográfica (SIG). 5. Difusão de espécies. I.Título.

CDU 004.023

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

Computação Aplicada

Dr. José Demisio Simões da Silva

Presidente / INPE / SJCampos - SP

Dr. Antonio Miguel Vieira Monteiro

Orientador(a) / INPE / SJCampos - SP

Dr. João Ricardo de Freitas Oliveira

Orientador(a) / INPE / SJCampos - SP

Dra. Silvana Amaral Kampel

Membro da Banca / INPE / SJCampos - SP

Dr. Miguel Angel Uribe Opazo

Convidado(a) / UNIOESTE / Cascavel - PR

Dr. Marcelo Azevedo Costa

Convidado(a) / UFMG / Belo Horizonte - MG

Aluno (a): Adair Santa Catarina

São José dos Campos, 08 de abril de 2009

AGRADECIMENTOS

Ao Instituto Nacional de Pesquisas Espaciais (INPE), que me proporcionou os estudos no programa de pós-graduação em Computação Aplicada e à Universidade Estadual do Oeste de Paraná (UNIOESTE) pelo suporte financeiro.

A todos os professores do programa de doutorado em Computação Aplicada, por compartilharem seus conhecimentos. Em especial, aos professores Antônio Miguel Vieira Monteiro e João Ricardo de Freitas Oliveira, pela amizade, apoio, incentivo, empenho e dedicação nas atividades de orientação.

Ao corpo técnico-administrativo do INPE, pelo zelo no exercício de suas atividades.

Aos colegas professores do Curso de Informática da UNIOESTE, pelas contribuições na realização deste trabalho.

Aos meus amigos e colegas do INPE: Danilo, Eduilson, Érica, Evaldinolia, Felipe, Flávia, Flavinha, Francisco, Fred, Gilmar, Giovana, Ilka, Joelma, Joice, Karla, Marckleuber, Missae, Murilo, Olga, Paulo, Pedro, Sérgio, Thales, Timbó, Vanessa e Vantier, pela amizade e companheirismo na horas de estudo e lazer.

Aos meus pais, Mário (*in memoriam*) e Inelde, pela vida e pela educação proporcionada. Aos meus irmãos, Ademir e Adessir, pelo incentivo e pela contribuição na minha formação educacional e humana.

À minha esposa Sirlei e aos meus filhos Mateus e Laura, pela inspiração, compreensão, apoio e incentivo incondicionais.

A todos que, de alguma forma, contribuíram para a conclusão do Curso de Doutorado em Computação Aplicada.

RESUMO

A dependência espacial é um conceito fundamental em análise geográfica. Tem sua origem na Primeira Lei da Geografia, assim denominada em homenagem ao geógrafo e matemático norte-americano Waldo Tobler, que enunciou que quando tratamos com fenômenos geográficos tudo está relacionado, mas as coisas próximas estão mais relacionadas do que coisas distantes. Este princípio mostra a importância de se considerar os relacionamentos espaciais na análise geográfica a partir dos dados geoespaciais. Entretanto, uma classe de algoritmos heurísticos utilizados na análise de dados geoespaciais, os Algoritmos Genéticos, negligenciam os relacionamentos espaciais e, conseqüentemente, os efeitos da dependência espacial. Nesta tese desenvolveu-se o SAHGA – "*Spatially Aware Hybrid Genetic Algorithm*" – um algoritmo heurístico híbrido com representação explícita de relacionamentos espaciais. O algoritmo desenvolvido incorpora a GPM (*Generalized Proximity Matrix*) para representar as associações espaciais entre objetos geoespaciais; em outras palavras os relacionamentos espaciais. O algoritmo SAHGA une duas heurísticas: Algoritmos Genéticos e *Simulated Annealing*. Desenvolveu-se dois sistemas que utilizam o algoritmo SAHGA: o SAHGA *Model Breeder* (SAHGA MB) e o SAHGA *Species Distribution Models* (SAHGA SDM). Utilizou-se o SAHGA MB na análise de dados sócio-econômicos e o SAHGA SDM na modelagem da distribuição potencial das espécies *Strix varia* Barton (1799) e *Thalurania furcata boliviana* Boucard (1894). Os dois sistemas desenvolvidos foram utilizados no ajuste de modelos ignorando, ou não, os relacionamentos espaciais. Comparou-se os resultados para averiguar os efeitos dos relacionamentos espaciais sobre os modelos ajustados. Também utilizou-se o algoritmo GARP, disponível no software openModeller Desktop, para criar modelos de distribuição de espécies que foram comparados com os modelos ajustados pelo SAHGA SDM. Os resultados obtidos mostram que os sistemas desenvolvidos ajustam modelos úteis. Os modelos ajustados pelo SAHGA MB e os modelos ajustados pelo SAHGA SDM são afetados pelos relacionamentos espaciais; os efeitos causados pelos relacionamentos espaciais dependem do conhecimento representado na GPM.

SAHGA - A SPATIALLY AWARE HYBRID GENETIC ALGORITHM FOR GEOSPATIAL DATA ANALYSIS

ABSTRACT

The spatial dependence is a fundamental concept in spatial analysis. The spatial dependence results from that “everything is related to everything else, but near things are more related than distant things”. Spatial data analysis may use heuristic algorithms as Genetic Algorithms; however, these algorithms ignore the spatial dependence. In this work we present the SAHGA – Spatially Aware Hybrid Genetic Algorithm – a Hybrid Heuristic Algorithm with explicit representation of spatial relationships. The proposed algorithm includes a Generalized Proximity Matrix (GPM) to represent the spatial association between objects in space, in other words the spatial relationships. SAHGA embody two heuristics: Genetic Algorithms and Simulated Annealing. We developed two software with SAHGA: SAHGA Model Breeder (SAHGA MB) and SAHGA Species Distribution Models (SAHGA SDM). The SAHGA MB was used in social data analysis and the SAHGA SDM was used to model the potential distribution of *Strix varia* Barton (1799) and *Thalurania furcata boliviana* Boucard (1894) species. Both software may be used to create models with and without spatial relationships. We compared the adjusted models, with and without spatial relationships, to observe the effects of spatial relationships. We also compared the created SDM with GARP models. The results obtained show that the developed software may adjust useful models. Model breeder and species distribution models were affected by spatial relationships; the effects of relationships are function of the knowledge represented by the GPM.

SUMÁRIO

Pág.

LISTA DE FIGURAS

Pág.

LISTA DE TABELAS

Pág.

LISTA DE SIGLAS E ABREVIATURAS

AG	Algoritmo Genético
AUC	*Area Under the Curve*
CCM	Coeficiente de Correlação de Matthews
FN	Falso Negativo
FP	Falso Positivo
GARP	*Genetic Algorithm for Rule-set Production*
GPM	*Generalized Proximity Matrix*
GPS	*Global Positioning System*
MB	*Model Breeder*
ROC	*Receiver Operating Characteristic*
SA	*Simulated Annealing*
SAHGA	*Spatially Aware Hybrid Genetic Algorithm*
SDM	*Species Distribution Models*
SIG	Sistemas de Informação Geográfica
VN	Verdadeiro Negativo
VP	Verdadeiro Positivo

1 INTRODUÇÃO

A capacidade para produzir, armazenar e recuperar dados espaço-temporais cresceu significativamente nos últimos anos. Contribuíram para isto o aumento na oferta de imagens de satélites em diferentes resoluções espaciais, espectrais e temporais, o uso de GPS para coleta direta de dados e o acesso a bases de dados demográficas mais detalhadas, como o censo brasileiro de 2000, que apresenta a malha censitária para os municípios brasileiros, ainda incompleta, mas representando um enorme avanço (IBGE, 2003)

Apesar dos Sistemas de Informação Geográfica (SIG) incrementarem nossa capacidade de analisar dados geoespaciais, esses necessitam integrar novos métodos de exploração e análise deste tipo de dado. Os métodos para explorar e analisar dados geoespaciais advém de áreas tradicionais como a estatística e de áreas emergentes como a inteligência computacional e sistemas complexos (Openshaw e Openshaw, 1997; Couclelis, 1998; Openshaw e Abrahart, 2000; Câmara e Monteiro, 2001).

Visando extrair informações contidas em bancos de dados geográficos, de forma automática ou semi-automática, algoritmos computacionais intensivos passaram a ser empregados, originando uma nova linha de pesquisa, a geocomputação.

O termo geocomputação foi cunhado por Openshaw e Abrahart (1996) para descrever o uso da computação intensiva na descoberta de conhecimento nas áreas de geografia humana e física. Atualmente este termo inclui também técnicas matemático-computacionais para análise de dados geoespaciais, modelos dinâmicos, visualização e modelos dinâmicos espaço-temporais (Longley *et al.*, 1998). Uma definição sucinta para geocomputação é dada por Rees e Turton (1998) como sendo o processo de aplicação da tecnologia

computacional para a solução de problemas geográficos. Para Openshaw (1999) a geocomputação oferece uma nova perspectiva e um paradigma para aplicar ciência num contexto geográfico.

Alguns métodos computacionais utilizados em geocomputação são as redes neurais, a busca heurística e os autômatos celulares (Openshaw e Openshaw, 1997; Longley *et al.*, 1998; Câmara e Monteiro, 2001). Dentre os algoritmos de busca heurística temos os algoritmos genéticos (AG) utilizados, por exemplo, em sistemas como os *model breeders* (Openshaw e Openshaw, 1997; Santa Catarina *et al.*, 2005) e o GARP - *Genetic Algorithm for Rule-set Production* (Stockwell e Peters, 1999).

Model breeders são sistemas semi-automáticos que analisam conjuntos de dados com variáveis dependentes e independentes para encontrar modelos que as relacionam. Estes modelos são conhecidos como modelos *data-driven* e possuem, como premissa essencial, a crença de que as observações futuras repetirão as relações consistentes que ocorrem nos dados observados. Um benefício significante deste tipo de modelo é que ele dispensa conhecimento profundo sobre o fenômeno modelado, um aspecto particularmente útil quando se trabalha com problemas espaciais (Solomatine, 2002; Bação, 2006; Jayawardena *et al.*, 2006; Fan, 2009).

GARP é um algoritmo utilizado na geração de modelos de distribuição de espécies, mais precisamente na predição da distribuição potencial de espécies. A área de distribuição potencial de uma espécie corresponde ao conceito de nicho fundamental da espécie. Este conceito remete àquelas áreas que apresentam os intervalos das condições ambientais necessárias para a existência da espécie, sem considerar a influência de competição interespecífica ou de predação por outras espécies (Hutchinson, 1957; Iwashita, 2007).

Um princípio essencial para compreender e analisar dados geoespaciais é originado da Primeira Lei da Geografia: "*Everything is related to everything else, but near things are more related than distant things*" (Tobler, 1970, p.236). Em outras palavras, a maior parte das ocorrências naturais, ou sociais, apresentam entre si uma relação que depende da distância (Câmara *et al.*, 2004). A expressão quantitativa deste princípio é o efeito da dependência espacial (Câmara e Monteiro, 2001).

Este princípio mostra a importância de se considerar os relacionamentos espaciais na análise geográfica a partir dos dados geoespaciais. Relacionamentos espaciais referem-se às relações que ocorrem entre objetos tanto no espaço absoluto quanto no espaço relativo, implementando o conceito de espaço próximo (Couclelis, 1997), onde a informação espacial é georreferenciada no espaço absoluto, mas está também associada a uma representação no espaço relativo do qual é parte (Pedrosa, 2003).

Entretanto, os relacionamentos espaciais, e conseqüentemente a dependência espacial, são negligenciados nos *model breeders* e no GARP; ambos ajustam modelos e realizam predições observando apenas os valores pontuais das amostras. Esta negligência ocorre não apenas nestes software; ela deve-se a ausência de um AG que incorpore, em seu mecanismos evolutivos, os relacionamentos espaciais.

Assim, a necessidade de tratar adequadamente os relacionamentos espaciais na análise de dados geoespaciais, particularmente usando sistemas semi-automáticos baseados em AG, trazem consigo questões específicas para serem respondidas:

a) É possível incorporar os relacionamentos espaciais nestes sistemas semi-automáticos de análise de dados geoespaciais?

b) Estes sistemas de análise de dados geoespaciais são capazes de operar sobre um modelo generalizado de relacionamentos espaciais?

c) Os resultados obtidos por estes sistemas, quando consideram os relacionamentos espaciais, diferem dos resultados obtidos quando os relacionamentos espaciais são ignorados?

d) O conhecimento pré-existente, acerca das formações naturais ou artificiais que afetam o problema em estudo, pode ser representado nestes sistemas?

A primeira questão remete à necessidade de computar, no sistemas semi-automáticos de análise de dados geoespaciais baseados em AG, os efeitos das inter-relações existente entre elementos espacialmente próximos, uma alusão direta à Primeira Lei da Geografia.

A segunda questão refere-se à estrutura dos relacionamentos espaciais, ou seja, com o fato de que os efeitos destes relacionamentos sobre os fenômenos espaciais não se manifestam de forma regular no espaço.

A terceira questão reflete a necessidade de averiguar se os relacionamentos espaciais, quando considerados, exercem influência sobre os resultados fornecidos pelos sistemas de análise de dados geoespaciais baseados em AG.

A quarta questão trata da inserção do conhecimento de um especialista sobre o problema e a região em estudo. Os elementos constituintes da paisagem, sejam naturais ou artificiais, podem influenciar os fenômenos espaciais de modo favorável ou desfavorável. Quando esta influência auxilia a compreensão do fenômeno estudado, este conhecimento deveria ser inserido e considerado no processo.

1.1 Hipótese Central

Os AG utilizados em sistemas de análise de dados geoespaciais, como os *model breeders* e o GARP, desconsideram os efeitos oriundos dos relacionamentos espaciais. Desconsiderar estes efeitos viola a Primeira Lei da Geografia enunciada por Tobler (1970).

A hipótese central deste trabalho é que é possível incorporar aos AG, utilizados em sistemas de análise de dados geoespaciais, uma estrutura para representação explícita de relacionamentos espaciais, a GPM – *Generalized Proximity Matrix* (Aguiar *et al.*, 2003; Pedrosa, 2003). A inserção da GPM nestes sistemas possibilita às abordagens evolutivas com AG, considerar os efeitos da dependência espacial nos fenômenos estudados.

Para validar esta hipótese desenvolveu-se, como prova de conceito, o SAHGA – *Spatially Aware Hybrid Genetic Algorithm* – um algoritmo heurístico híbrido com representação explícita de relacionamentos espaciais através da GPM. O SAHGA é híbrido pois combina, em seu núcleo de otimização, um AG com a heurística *simulated annealing* (Mahfoud e Goldberg, 1995).

Com a inserção da GPM no ciclo da abordagem por AG, além da representação explícita dos relacionamentos espaciais, também é possível inserir conhecimento prévio sobre os elementos naturais e artificiais que compõem a região e que, na perspectiva do especialista, afetam o fenômeno em estudo. Como exemplo de conhecimento prévio, representável através da GPM, tem-se a ocorrência de rios, cachoeiras e cadeias de montanhas, que são elementos naturais do relevo que afetam a distribuição de espécies animais e vegetais (Höglund e Shorey, 2004; Phillips *et al.*, 2006; Spens *et al.*, 2007).

O SAHGA é um algoritmo que se adapta para usos múltiplos. Para demonstrar sua adaptabilidade e funcionalidades desenvolveu-se dois sistemas: o SAHGA MB – *Model Breeder* e o SAHGA SDM – *Species Distribution Modeling*.

Utilizou-se o primeiro na análise de dados sócio-econômicos e o segundo na geração de modelos de distribuição potencial da espécie *Strix varia* Barton (1799) conhecida como coruja listrada, e da espécie *Thalurania furcata boliviana* Boucard (1894) conhecida como beija-flor-tesoura-verde. Para averiguar se os relacionamentos espaciais alteram os modelos ajustados, empregou-se ambos os sistemas no ajuste de modelos sem e com relacionamentos espaciais.

Também comparou-se os modelos de distribuição potencial de espécies gerados pelo SAHGA SDM com os modelos gerados por implementações do algoritmo GARP, disponíveis no software openModeller Desktop v1.0.6 (CRIA *et al.*, 2008; Santana *et al.*, 2008), objetivando avaliar a qualidade dos modelos fornecidos pelo sistema SAHGA SDM.

1.2 Organização do Texto

No Capítulo 2 tem-se o referencial teórico, que visa apresentar conceitos necessários à compreensão do trabalho desenvolvido. Dividiu-se o Capítulo 2 em 4 seções: a primeira seção aborda conceitos oriundos da geoinformática; a segunda seção apresenta os AG; a terceira seção apresenta os *model breeders* detalhando a implementação de um deles e a quarta seção apresenta os modelos de distribuição de espécies.

O Capítulo 3 detalha o algoritmo SAHGA, sua utilização na implementação do sistema SAHGA MB e a utilização deste sistema num caso de uso, realizando análise de dados sócio-econômicos.

O Capítulo 4 apresenta o sistema SAHGA SDM, uma aplicação do algoritmo SAHGA na construção de um sistema para geração de modelos de distribuição de espécies. Dois estudos de caso, onde utilizou-se o sistema desenvolvido, são apresentados: no primeiro modelou-se a distribuição potencial da espécie *Strix varia* e no segundo a distribuição potencial da espécie *Thalurania furcata boliviana*.

Finalmente, o Capítulo 5 apresenta as conclusões do trabalho.

2 REFERENCIAL TEÓRICO

Este capítulo visa apresentar conceitos necessários à compreensão do trabalho desenvolvido. Os conceitos apresentados originam-se de diversas áreas de conhecimento como a geoinformática, a inteligência computacional e a ecologia.

Na primeira seção do capítulo são apresentados métodos para representação do espaço bi-dimensional, o conceito de vizinhança espacial e a GPM, a estrutura genérica para representação de relacionamentos espaciais.

Na segunda seção do capítulo são apresentados os AG. Estes algoritmos são utilizados na construção de sistemas como os *model breeders* e o GARP. Após um breve relato histórico é apresentada a estrutura básica destes algoritmos, o mecanismo de seleção, os principais operadores genéticos e a influência dos parâmetros genéticos sobre o desempenho dos mesmos. Ao final da seção é apresentada uma tendência observada com as heurísticas utilizadas em otimização: a hibridização. No caso a hibridização dos AG com *simulated annealing*.

Na terceira seção do capítulo a implementação de um *model breeder* é detalhada, visando demonstrar como codificar modelos em estruturas operáveis pelos AG, os cromossomos, bem como sua avaliação através de uma função de aptidão. A codificação e a função de aptidão são determinantes para a qualidade de solução obtida pelos AG.

Na quarta seção do capítulo são apresentados os SDM, sua estrutura e utilização. Também são apresentados os mecanismos utilizados na avaliação dos SDM, como a matriz de confusão, os índices dela extraídos e a curva ROC.

2.1 Representação do Espaço Bi-dimensional e Vizinhança Espacial

Em geoinformática, o espaço bi-dimensional pode ser representado através de um conjunto de polígonos fechados, grades regulares, grades irregulares, entre outras.

Na representação do espaço bi-dimensional através de polígonos, as divisões do espaço correspondem a unidades de área que possuem características próprias. Os limites destas regiões são estabelecidos pelo pesquisador, em função do objeto de estudo e da escala, ou correspondem a divisões políticas ou legalmente estabelecidas. Este modelo de representação corresponde a implementação do conceito de "unidade de área" de Hartshorne (1978).

Neste modelo de representação a unidade de trabalho é o polígono. Assim, quando os métodos de análise de dados geoespaciais necessitam de informações georreferenciadas num ponto, deve-se transformar os polígonos em pontos, geralmente seus centróides. A divisão política do Brasil em unidades federativas, conforme ilustrado na Figura 2.1, é um exemplo de representação do espaço bi-dimensional através de polígonos.

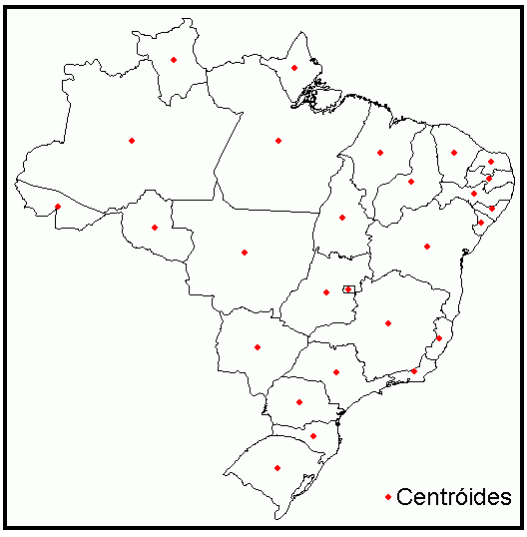

Figura 2.1 – Divisão política do Brasil

O uso de grades regulares na representação do espaço bi-dimensional também é corriqueiro. Alguns trabalhos que se utilizam desta representação são: (Stockwell e Peters, 1999), (Xiao *et al.*, 2002), (Goud, 2003), (Aukema *et al.*, 2006) e (Rangel *et al.*, 2006). As motivações para o uso deste modelo de dados são influenciadas pela natureza discreta do dado remoto e pela conveniência da programação e implementação de estruturas baseadas em grades (O'sullivan, 2002).

Nesta representação uma porção do espaço passa a ser representada por uma célula de tamanho regular, como mostrado na Figura 2.2, onde cada célula representa uma área de 64 km^2 (8 km x 8 km).

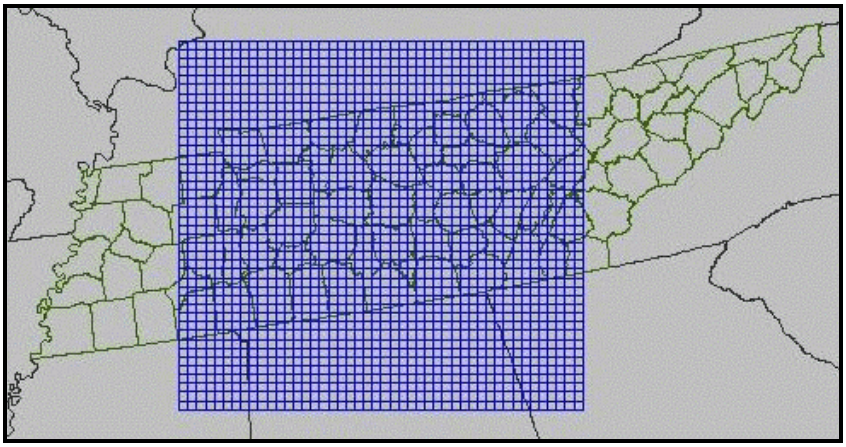

Figura 2.2 – Grade regular sobre o estado do Tennessee, EUA
Fonte: CEMPD (2008)

Cada célula é rodeada por células de mesmo tamanho, definindo uma vizinhança. Esta vizinhança pode assumir diferentes configurações, que se mantêm constante em todo o espaço. Exemplos clássicos de configurações são apresentadas na Figura 2.3.

Porém, esta representação não é suficiente para modelar diversos fenômenos do mundo real como o processo de mudança do uso de solo da Amazônia (Pedrosa, 2003) ou a distribuição de espécies de peixes ao longo de um rio.

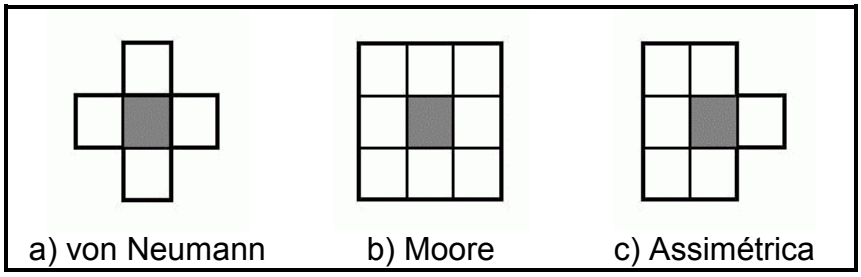

Figura 2.3 – Exemplos de vizinhança
Fonte: Adaptado de Pedrosa (2003)

No processo de mudança do uso de solo da Amazônia há um fator condicionante, a presença de aglomerados urbanos. A localização destes aglomerados é fortemente influenciada pela rede de transporte da região.

Ao longo de um rio a rede de afluentes, a localização das fontes poluidoras e as cachoeiras são fatores que influenciam a distribuição das espécies de peixes; por exemplo, espécies à jusante de uma cachoeira podem não ser encontradas a montante desta barreira natural.

Assim, uma representação mais adequada, para estudar estes fenômenos, seria aquela capaz de representar as relações de vizinhança entre as células. Em modelos celulares a matriz de proximidade mostra-se uma solução apropriada para dar a estes modelos a flexibilidade necessária para capturar as relações de vizinhança entre células.

Uma representação ainda mais flexível é a GPM, vista em maiores detalhes na seção seguinte.

2.1.1 *Generalized Proximity Matrix* (GPM)

A GPM, ou matriz de vizinhança generalizada, é uma variação da matriz de proximidade. Os pesos são calculados a partir de relações espaciais no espaço

absoluto como distância euclidiana e adjacência, ou com base em relações espaciais no espaço relativo, que levam em conta a conectividade de objetos em uma rede de transporte ou de comunicação, por exemplo (Aguiar *et al.*, 2003; Pedrosa, 2003).

Uma GPM é composta por um conjunto de objetos geoespaciais *O*, um grafo *G* e uma matriz de proximidade *V*:

a) Os objetos geoespaciais (*O*) são representados por células regulares ou polígonos, de acordo com a representação espacial utilizada.

b) O grafo (*G*) é constituído por um conjunto de nós e arcos; cada nó representa um objeto (célula ou polígono) e os arcos representam os relacionamentos de vizinhança entre dois nós.

c) A matriz de proximidade (*V*) é composta por um conjunto de elementos W_{ij} que serve para indicar o quanto dois objetos O_i e O_j estão próximos; geralmente é representada em termos de adjacência ou distância euclidiana. As opções mais comuns para definir W_{ij} são:

   - W_{ij} = 1 se O_i é vizinho de O_j; caso contrário W_{ij} = 0;
   - W_{ij} = 1, se a distância(O_i, O_j) < max; caso contrário W_{ij} = 0;
   - W_{ij} = 1/(distância(O_i, O_j))^2, se $i = j$, W_{ij} = 0.

A GPM independe do modelo utilizado na representação do espaço bi-dimensional; ela pode ser utilizada tanto no modelo de representação com polígonos quanto no modelo de representação com grades regulares.

2.2 Algoritmos Genéticos

Os AG são algoritmos heurísticos de busca, que utilizam regras baseadas numa metáfora do processo evolutivo proposto por Charles Darwin, operando sobre um espaço de soluções codificado (Goldberg, 1989; Holland, 1992). Os AG simulam o mecanismo evolucionário dos sistemas biológicos naturais, onde os indivíduos mais aptos têm maior probabilidade de se reproduzir gerando descendentes.

2.2.1 Histórico

No começo dos anos 70, John Holland, quando pesquisava as características da evolução natural, acreditava que se estas características fossem adequadamente incorporadas a algoritmos computacionais, poder-se-ia produzir uma técnica para solucionar problemas difíceis da mesma forma que a natureza fazia para resolver os seus problemas, ou seja, usando a evolução.

Acreditando nisto ele deu início a uma pesquisa sobre algoritmos que manipulavam cadeias de 0 e 1, às quais deu o nome de cromossomos. Os algoritmos de Holland realizavam a evolução simulada de populações destes cromossomos. Desta forma, imitando a natureza, seus algoritmos resolviam muito bem o problema de encontrar bons cromossomos, através da manipulação do material contido nos cromossomos.

Outro ponto interessante nas técnicas desenvolvidas por Holland é que, assim como na natureza, estes cromossomos não têm conhecimento algum sobre o tipo de problema que estão resolvendo. A única informação que eles dispunham era uma avaliação de cada cromossomo produzido. O objetivo desta avaliação era verificar quais cromossomos estavam mais adaptados e,

com base nisto, aumentar suas chances de serem selecionados para a reprodução.

Quando Holland começou os seus estudos sobre estes algoritmos, eles ainda não tinham um nome. Foi apenas quando esta técnica começou a demonstrar o seu potencial que houve a necessidade de se dar um nome adequado e significativo a ela. Como uma referência às suas origens na biologia, Holland os batizou de Algoritmos Genéticos.

2.2.2 Estrutura Clássica de um AG

Em um AG clássico, como apresentado por Goldberg (1989), o algoritmo inicia gerando um conjunto de soluções aleatórias codificadas em cadeias de dígitos binários, chamado população. Cada indivíduo da população recebe o nome de cromossomo, representando uma solução candidata para o problema. A avaliação de cada cromossomo determina seu índice de aptidão; indivíduos mais aptos têm maior probabilidade de gerarem filhos. Os cromossomos evoluem através de iterações sucessivas chamadas de gerações.

O processo de seleção e os operadores de cruzamento e mutação são os responsáveis por criar as novas gerações. O processo de seleção visa escolher indivíduos, de acordo com seu índice de aptidão, para combiná-los através do operador de cruzamento, gerando novos indivíduos que mantém características de seus pais. Posteriormente estes novos indivíduos podem ser modificados pelo operador de mutação. Após várias gerações o AG pode produzir soluções aceitáveis para um problema.

O elemento de ligação entre o AG e o problema a ser resolvido é a função de aptidão. Esta toma como entrada um cromossomo e retorna um número, ou

uma lista de números, que representam a medida de performance do cromossomo com relação ao problema a ser resolvido. O bom desempenho de um AG está condicionado à qualidade da avaliação da aptidão dos indivíduos (Oliveira, 2000).

De maneira geral, um AG clássico pode ser descrito através do fluxograma apresentado na Figura 2.4.

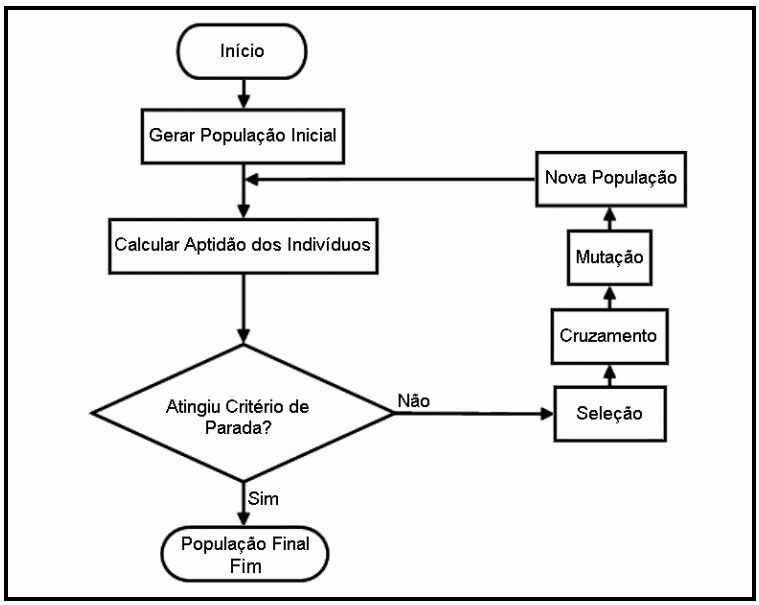

Figura 2.4 – Fluxograma de um AG clássico

2.2.3 Esquemas de Codificação Empregados nos AG

Os esquemas usados na codificação dos cromossomos variam de problema para problema e de AG para AG. A codificação clássica usada no trabalho de Holland, e até hoje a mais usada, utiliza cadeias de 0 e 1. Com o passar do tempo, outros pesquisadores apresentaram outras formas de codificação.

A codificação clássica, quando utilizada em problemas que possuem variáveis contínuas e cujas soluções requeridas necessitam boa precisão numérica,

torna os cromossomos longos. Para cada ponto decimal acrescentado na precisão, é necessário adicionar 3,3 dígitos (0 ou 1) na cadeia (Lacerda e Carvalho, 1999).

A conseqüência imediata do aumento da cadeia, que representa o cromossomo, é o aumento no tempo necessário para calcular o equivalente decimal deste cromossomo. Por este motivo, formas não clássicas de codificação dos cromossomos foram desenvolvidas, gerando codificações adequadas para problemas específicos (Herrera *et al.*, 1998).

Uma das formas não clássicas de codificação mais utilizada é a codificação real. Esta forma de codificação consiste em representar variáveis numéricas contínuas através de seu próprio valor real. Um cromossomo pode ser composto por múltiplos genes quando o problema a ser resolvido envolve duas ou mais variáveis.

As primeiras aplicações da codificação real foram propostas por Lucasius e Kateman (1989) e Davis (1989). A partir de então a codificação real tornou-se padrão em problemas de otimização numérica com variáveis contínuas.

Castro (1999) afirma que, com certeza, nenhuma forma de codificação funcionaria igualmente bem em todas as situações e que, para cada caso, deve-se fazer uma escolha cuidadosa do tipo de codificação a ser utilizada, pois uma codificação ruim pode não levar ao resultado esperado.

2.2.4 Seleção e Elitismo

A finalidade da seleção é escolher os elementos da população que devem se reproduzir. Em problemas de maximização, esta escolha deve ser feita de tal

forma que dê maior chance de reprodução aos membros da população mais adaptados ao meio ambiente, isto é, àqueles que apresentam um valor de aptidão mais elevado. O mecanismo de seleção mais utilizado é a roleta.

Neste método de seleção cada indivíduo da população é representado na roleta proporcionalmente ao seu índice de aptidão. Assim, aos indivíduos com alta aptidão é dada uma porção maior da roleta, enquanto aos de aptidão mais baixa é dada uma porção relativamente menor da roleta. Finalmente, a roleta é girada um determinado número de vezes, dependendo do tamanho da população, e são escolhidos, como indivíduos que participarão da próxima geração, aqueles sorteados na roleta. O método da roleta é apresentado na Figura 2.5.

S_i	Indivíduo S_i	*Fitness* $f(S_i)$	*Fitness* Relativo
S1	10110	3,43	0,18
S2	11000	6,83	0,35
S3	11110	4,52	0,24
S4	01001	1,99	0,10
S5	00110	2,45	0,13

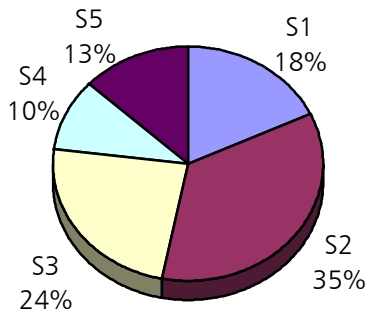

Figura 2.5 – A seleção através do método da roleta

O elitismo é uma técnica utilizada para melhorar a convergência dos AG. Ele foi primeiramente introduzido por De Jong (1975) como uma adição aos métodos de seleção. O elitismo força os AG a reter um certo número de “melhores” indivíduos em cada geração. Tais indivíduos podem ser perdidos se não forem selecionados para reprodução ou se forem destruídos por cruzamento ou mutação.

Em outras palavras, o elitismo seleciona os melhores cromossomos de uma população e transporta-os à geração seguinte. Esta técnica consiste

basicamente em realizar o processo de seleção em duas etapas:

a) Seleciona-se uma elite de *r* membros entre os melhores da população inicial, os quais são incorporados diretamente na população final;

b) O restante da população final é obtida a partir dos (*n* - *r*) elementos restantes da população inicial de tamanho *n*.

Em geral a elite tem um tamanho reduzido, com *r* = 1 ou 2 para um *n* = 50. Quando é utilizada a técnica do elitismo, o algoritmo converge mais rapidamente. Como na natureza, os indivíduos mais aptos podem, além de reproduzirem-se mais, ter uma vida mais longa, muitas vezes sobrevivendo de uma geração para a outra e se reproduzindo. O efeito negativo desta estratégia prende-se ao fato de que a população inicial pode convergir para uma população homogênea de superindivíduos, não explorando outras soluções.

2.2.5 Operadores Genéticos

Nos AG a seleção e os operadores genéticos são utilizados para criar as novas gerações, fazendo a evolução da população acontecer.

Holland (1992) define três operadores para criar filhos diferentes dos pais: cruzamento, mutação e inversão. O objetivo final destes operadores é fazer com que os cromossomos criados durante o processo de reprodução sejam diferentes dos cromossomos dos pais. O operador de cruzamento é responsável por combinar os cromossomos dos pais na criação dos cromossomos filhos. O operador de mutação é responsável pela introdução de pequenas mudanças aleatórias nos cromossomos dos filhos. O operador de inversão inverte partes da cadeia cromossômica e seu uso é indicado quando existe ordem entre os elementos da cadeia cromossômica (Ichihara, 1998).

Diversos operadores de cruzamento e mutação foram desenvolvidos por pesquisadores, alguns adequados a um tipo específico de codificação dos cromossomos, outros com intenção de serem mais genéricos. Serão aqui mencionados apenas os mais utilizados.

O operador de cruzamento em um ponto é o método de cruzamento mais simples e o mais utilizado. Este método consiste em dividir os cromossomos selecionados num ponto de sua cadeia, ponto este escolhido aleatoriamente. Posteriormente copia-se para os novos cromossomos uma parte de cada um dos cromossomos selecionados - cromossomos pais - formando assim os novos cromossomos, os cromossomos filhos. Nas implementações mais tradicionais, é comum um par de cromossomos selecionados dar origem a dois filhos, mas este não é um fator restritivo. A princípio, pode-se criar qualquer quantidade de filhos, desde que, é claro, o tamanho do cromossomo permita o número desejado de combinações diferentes. A Figura 2.6 apresenta um exemplo do operador de cruzamento em um ponto.

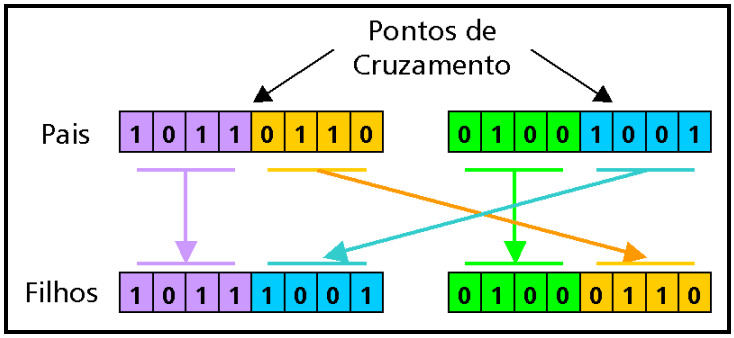

Figura 2.6 – Operador de cruzamento em um ponto

Outro método de cruzamento, um pouco menos utilizado que o cruzamento em um ponto, é o cruzamento em múltiplos pontos. Este método divide o cromossomo em vários pontos e os recombina para formar os filhos, assemelhando-se mais ao processo que ocorre na vida real; possui a vantagem de assegurar uma variedade genética maior.

Nos AGs com codificação real estes operadores de cruzamento não são adequados, pois apenas trocam os valores dos genes, não criando novos valores. Assim, os operadores de cruzamento aritméticos são mais indicados. Alguns operadores de cruzamento aritméticos são: média aritmética (Davis, 1991), média geométrica, *BLX-*α (Eshelman e Schaffer, 1993), aritmético e heurístico (Michalewicz, 1996).

Os cruzamentos média aritmética e média geométrica consistem em gerar um novo cromossomo usando a média aritmética simples e a média geométrica de dois cromossomos pais, respectivamente. O cruzamento *BLX-*α consiste em gerar um novo cromossomo a partir da Equação 2.1.

$$c = p_1 + \beta(p_2 - p_1) \tag{2.1}$$

onde *c* é o novo cromossomo gerado, p_1 e p_2 são os cromossomos pais e $\beta \in U(-\alpha, 1 + \alpha)$. α é um pequeno valor que estende os limites para a definição de *c* e *U* é a distribuição de probabilidade uniforme. Caso o cromossomo seja formado por múltiplos genes a Equação 2.1 é aplicada a cada par de genes de p_1 e p_2.

O cruzamento aritmético consiste em gerar dois cromossomos filhos (c_1 e c_2) a partir de dois cromossomos pais (p_1 e p_2), usando a Equação 2.2.

$$\begin{aligned} c_1 &= \beta p_1 + (1-\beta)p_2 \\ c_2 &= (1-\beta)p_1 + \beta p_2 \end{aligned} \tag{2.2}$$

onde $\beta \in U(0, 1)$.

O cruzamento heurístico consiste em gerar um cromossomo filho a partir de uma interpolação linear entre os pais usando a informação da aptidão. Dados

dois cromossomos p_1 e p_2 em que p_1 é melhor do que p_2 em termos de aptidão, obtém-se um cromossomo c através da Equação 2.3.

$$c = p_1 + r(p_1 - p_2), \text{ onde } f(p_1) > f(p_2) \tag{2.3}$$

onde $r \in U(0, 1)$ e f é a função de aptidão.

O operador mutação de *bit* é aplicável em todas as formas binárias de representação de cromossomos. O processo de mutação de *bit* é bem simples, e normalmente é realizado da seguinte maneira: dada uma certa probabilidade de mutação, normalmente muito baixa e determinada de forma empírica, cada *bit* na cadeia do cromossomo é avaliado para saber se deverá sofrer uma mutação; caso deva sofrer mutação, seu valor é simplesmente trocado por outro.

A Tabela 2.1 mostra 3 cromossomos de comprimento 4 e os números aleatórios gerados para cada um dos bits no cromossomo, os novos bits que demonstram as possibilidades de mutação e o resultado final após a mutação. Os números em negrito na coluna "N^{os} aleatórios" indicam probabilidades muito baixas (< 2%) e, portanto, serão os genes que sofrerão mutação. Os dígitos em negrito na coluna "Novo cromossomo" são os genes alterados.

Tabela 2.1 – Exemplos de mutação de bit

Cromossomo anterior	N^{os} aleatórios				Novo *bit*	Novo cromossomo
0011	0,653	**0,007**	0,287	0,373	1	0**1**11
1001	0,721	0,432	0,043	0,840	-	1001
1110	**0,012**	0,076	0,934	0,471	0	**0**110

Quando se utiliza a codificação em números reais a mutação pode ser realizada de diversas formas: uniforme, gaussiana, *creep*, limite, não-uniforme

e não-uniforme múltipla. As três últimas formas de mutação foram propostas por Michalewicz (1996).

A mutação uniforme consiste em substituir o gene selecionado do cromossomo por outro gene gerado aleatoriamente, segundo uma distribuição uniforme, entre os limites mínimo e máximo permitidos. A mutação gaussiana consiste em substituir o gene selecionado por outro gerado a partir de uma distribuição $N(pi, \sigma^2)$, onde *pi* é igual ao valor de gene a ser substituído e a variância σ^2 é definida pelo usuário. Lacerda e Carvalho (1999) mencionam que o valor da variância σ^2 pode ser diminuído à medida que aumenta o número de gerações do algoritmo genético.

A mutação *creep* consiste em acrescentar ou subtrair um pequeno número aleatório obtido de uma distribuição $N(0, \sigma^2)$ onde a variância σ^2 assume um valor pequeno. Esta mutação é usada para explorar localmente o espaço de busca.

A mutação não-uniforme consiste na simples substituição de um gene por um número extraído de uma distribuição não-uniforme. A mutação não-uniforme múltipla consiste em aplicar a mutação não-uniforme em todos os genes do cromossomo selecionado.

Se compararmos a reprodução sexuada (cruzamento) e a assexuada (mutação), veremos que na reprodução sexuada é necessário haver mais de um indivíduo. Os indivíduos devem desprender uma boa parcela de seu tempo e energia para encontrar um parceiro certo para realizar a reprodução sexuada; isto representa um custo a mais para o indivíduo/algoritmo. Porém, como a reprodução sexuada parece ter vencido esta guerra, pode-se concluir que este talvez seja um preço pequeno a pagar, comparado aos benefícios que traz consigo. Um destes benefícios é a rápida combinação de características benéficas, o que não é possível no caso da reprodução assexuada.

Uma das formas de vida que mais demonstra possuir uma alta capacidade de adaptação reproduz-se assexuadamente, o vírus. O alto poder de adaptação dos vírus vem do fato de que eles são altamente mutáveis, o que pode nos levar a concluir que a capacidade de sofrer mutações também é uma determinante nos organismos naturais. Ainda que não tenhamos cruzamento, se tivermos uma taxa de mutação bastante elevada, nossa população poderá ser capaz de comportar-se como os vírus, mudando sempre para se adaptar ao seu meio ambiente, e reproduzindo-se de forma assexuada.

2.2.6 Parâmetro Genéticos

É importante também, analisar de que maneira alguns parâmetros influem no comportamento dos AGs, para que se possa estabelecê-los conforme as necessidades do problema e dos recursos disponíveis.

a) Tamanho da População: o tamanho da população determina o número de cromossomos na população, afetando o desempenho global e a eficiência dos AG. Com uma população pequena o desempenho pode cair, pois a população fornecerá uma pequena cobertura do espaço de busca do problema. Uma grande população geralmente fornece uma cobertura representativa do domínio do problema, além de prevenir convergências prematuras para soluções locais ao invés de globais. No entanto, para se trabalhar com grandes populações, são necessários mais recursos computacionais, ou que o algoritmo trabalhe por um período de tempo muito maior;

b) Taxa de Cruzamento: determina a probabilidade com que um cruzamento ocorrerá. Quanto maior for esta taxa, mais rapidamente novas estruturas serão introduzidas na população. Mas se esta for muito alta, a maior parte da população será substituída, e pode

ocorrer perda de estruturas de alta aptidão. Com um valor baixo, o algoritmo pode tornar-se lento;

c) Taxa de Mutação: determina a probabilidade de ocorrência de uma mutação. Uma baixa taxa de mutação previne a convergência prematura para um ótimo local, possibilitando ao algoritmo explorar melhor todo o espaço de busca. Uma taxa de mutação muito alta faz com que o processo de busca torne-se essencialmente aleatório;

d) Intervalo de Geração: controla a porcentagem da população que será substituída durante a próxima geração. Com um valor alto, a maior parte da população será substituída, podendo ocorrer perda de estruturas de alta aptidão. Com um valor baixo, o algoritmo pode tornar-se lento.

Na implementação dos AG é possível utilizar uma grande variedade de opções, desde a codificação do cromossomo, a implementação dos operadores genéticos até a escolha dos parâmetros para o AG, como tamanho da população inicial, taxas de cruzamento e mutação entre outros. Existe uma grande discussão sobre como configurar os parâmetros de um AG, pois estes tipicamente interagem de forma não linear, sendo assim eles não podem ser otimizados um de cada vez (Mitchell, 1998). Isto implica que muitas vezes tais parâmetros devam ser ajustados empiricamente ou baseados em trabalhos correlatos.

Dada a robustez dos AG, mesmo com uma escolha inadequada dos parâmetros genéticos, eles são capazes de convergir para uma solução otimizada. Os reflexos desta escolha inadequada manifestam-se, principalmente, no tempo necessário para a convergência (Santa Catarina e Bach, 2003).

2.2.7 Hibridização

A hibridização resulta na integração de uma boa maneira convencional de resolver um problema aos conceitos usuais de AG (Bittencourt, 1998). O resultado costuma ser melhor que o obtido com qualquer uma das duas técnicas isoladamente (Davis, 1991).

A hibridização agrega a representação usual de dados no domínio original, bem como as técnicas de otimização já existentes. Isto permite a incorporação de heurísticas otimizadoras ao conjunto de operadores genéticos (cruzamento e mutação) que passam portanto a ser dependentes do domínio. Nesse sentido, o algoritmo genético passa a ser muito mais uma filosofia de otimização do que um método pronto para utilização (Bittencourt, 1998).

Um exemplo de hibridização, na representação dos dados, é a codificação dos cromossomos usando número reais e não números binários. Alguns conceitos teriam que ser adaptados: por exemplo, a mutação não seria mais a troca simples de um *bit*, mas a geração de um novo real, possivelmente dentro de um intervalo dado. Já a recombinação de dois reais poderia ser qualquer número compreendido entre eles, ou talvez a sua média aritmética.

Em relação às heurísticas de otimização, os AG podem ser combinados com *simulated annealing*, *hill climbing, ant colony* e busca tabu, por exemplo, definindo um algoritmo heurístico híbrido (AHH) (Mahfoud e Goldberg, 1995; Gwee e Chang, 2003; Lee *et al.*, 2008).

Uma das heurísticas utilizada em conjunto com os AG é o *Simulated Annealing*. A seção a seguir apresenta esta heurística em maiores detalhes.

2.2.8 *Simulated Annealing* (SA)

Esta heurística é uma metáfora de um processo térmico utilizado para obtenção de estados de baixa energia num sólido. O processo consiste de duas etapas: na primeira a temperatura do sólido é aumentada para um valor máximo no qual ele se funde; na segunda o resfriamento deve ser realizado lentamente até que o material se solidifique. Nesta segunda fase, executada lentamente, os átomos que compõem o material organizam-se numa estrutura uniforme com energia mínima.

O processo de recozimento (*annealing*) pode ser visto como um processo estocástico de determinação da organização dos átomos num sólido que apresente energia mínima. Em altas temperaturas os átomos movem-se livremente, com grande probabilidade de se moverem para posições que incrementarão a energia total do sistema.

Quando a temperatura baixa, os átomos gradualmente movem-se em direção a uma estrutura regular; somente com pequena probabilidade incrementarão suas energias. Esse processo foi simulado em computador, com sucesso, por Metropolis *et al.* (1953).

O algoritmo utilizado baseava-se em métodos de Monte Carlo e gerava uma seqüência de estados de um sólido da seguinte maneira: dado um estado corrente *i* do sólido com energia E_i, um estado subseqüente era gerado pela aplicação de um mecanismo de perturbação, o qual transformava o estado corrente em um próximo estado por uma pequena distorção, por exemplo, pelo deslocamento de uma única partícula. A energia do próximo estágio passa a ser E_j.

Se a diferença de energia fosse menor ou igual a zero, o estado *j* era aceito

como estado corrente. Se a variação fosse maior que zero, o estado *j* era aceito com uma probabilidade dada por: $exp((E_i - E_j)/(KB * T))$ onde *T* representa a temperatura atual do sistema e *KB* é uma constante física conhecida como constante de Boltzmann. Essa regra de aceite é conhecida como critério de Metropolis e o algoritmo também leva o seu nome.

Kirkpatrick *et al.* (1983) desenvolveram um algoritmo, de utilização genérica, análogo ao de Metropolis, denominado Algoritmo *Simulated Annealing* (Algoritmo de Recozimento Simulado). Nesse algoritmo, utilizaram como critério de aceite uma nova solução, a Equação 2.4.

$$P_{c_k}(aceitar\ j) = \begin{cases} 1 & se\ g_j \le g_i \\ \exp\left(\dfrac{-(g_j - g_i)}{c_k}\right) & se\ g_j > g_i \end{cases} \qquad (2.4)$$

onde *g* é a função a ser otimizada (no caso minimizada), *i* é a solução corrente, *j* é uma solução candidata e c_k um parâmetro representando a temperatura *T*.

Segundo a Equação 2.4, se uma solução candidata *j* é melhor que a solução corrente *i*, ou seja ($g_j \le g_i$), esta é aceita com probabilidade 1. Caso contrário, a solução candidata é aceita com uma dada probabilidade. Essa probabilidade é maior na medida em que for menor a variação de energia, definida por ($g_j - g_i$).

Ao mesmo tempo, à medida que há um decréscimo da temperatura c_k, o algoritmo torna-se mais seletivo, passando a aceitar, com menor freqüência, soluções que apresentem grande aumento na variação de energia, isto é, soluções que sejam muito piores que a solução corrente. Essa probabilidade tende a zero à medida que a temperatura se aproxima do ponto de congelamento.

O algoritmo SA pode ser considerado como uma extensão do método original de busca local. A busca local requer somente a definição de um esquema de vizinhança, e um método de avaliação do custo de uma solução em particular, sempre apresentando uma solução final (De Bona e Algeri, 2005).

Entende-se por esquema de vizinhança o mecanismo através do qual se obtém uma nova solução, também pertencente ao espaço de soluções do problema, realizando uma pequena alteração na solução corrente.

O método de busca local é ineficiente quanto à armadilha do ótimo local, fazendo desse método uma heurística pobre para muitos problemas de otimização combinatória.

Uma propriedade desejável de qualquer algoritmo é a habilidade de encontrar uma boa solução, independente do ponto de partida. Um ótimo local se caracteriza quando o algoritmo atinge uma região correspondente ao fundo de um vale, em se tratando de um problema de minimização, que não contém a solução ótima e dele não consegue sair, uma vez que todas as soluções naquela vizinhança possuem valores maiores que a solução corrente.

Uma estratégia para escapar da armadilha do ótimo local é executar diversas vezes o algoritmo com diferentes soluções iniciais, sendo adotado como solução a melhor solução encontrada. Entretanto, esse procedimento conduz a um novo problema que é o de determinar quando parar o algoritmo, além de poder ser inviável em se tratando de grandes problemas (Araujo, 2001).

O algoritmo SA consegue escapar de um ótimo local uma vez que o aceite de uma nova solução não depende única e exclusivamente do seu valor. Mesmo apresentando um valor pior que o da solução corrente, uma nova solução pode ser aceita de forma probabilística. Outra característica muito interessante do algoritmo SA é a simplicidade de sua implementação computacional, conforme

mostrado na Figura 2.7.

```
Ler α e NR; //Constante de resfriamento e número de repetições
S = S();    //Conjunto aleatório de soluções iniciais
T = LS;     //Limite superior
TMIN = LI;  //Limite inferior

Enquanto (T > TMIN) faça
   Para i = 1 até NR faça
      Gerar uma solução S' a partir de S; //perturbação de S
      Avaliar a variação de energia;      //ΔE = g(S') - g(S);
      Se (variação de energia <= 0) então S = S'
      Senão
         Gerar aleatoriamente Rnd;  //no intervalo (0, 1)
         Se (Rnd < exp(-variação / T) então S = S';
      Fim se;
   Fim Para;
   T = T * α;
Fim enquanto;
```

Figura 2.7 – O algoritmo SA

Para evitar a convergência precoce para um mínimo local, o algoritmo inicia com um valor de *T* relativamente alto. Esse parâmetro é gradualmente diminuído e, para cada um dos seus valores, são realizadas várias tentativas (*NR*) de se alcançar uma melhor solução, nas vizinhanças da solução corrente.

A expressão *T = T ** α corresponde ao processo de diminuição da temperatura, normalmente o parâmetro α é uma constante positiva menor que um.

2.3 *Model Breeders*

Model Breeders são sistemas para modelagem automática (Openshaw e Openshaw, 1997). Estes sistemas são capazes de encontrar um modelo matemático que relaciona variáveis independentes a uma variável dependente, seguindo a expressão geral representada na Equação 2.5.

$$y = f\left(x_1, x_2, \cdots, x_n\right) \tag{2.5}$$

Na ferramenta proposta por Openshaw e Openshaw (1997), o mecanismo utilizado para automatizar este processo foi um Algoritmo Genético. Evoluindo um conjunto de soluções iniciais, obtidas aleatoriamente, este algoritmo é capaz de encontrar um modelo matemático que explique o comportamento de uma variável dependente em função de um conjunto de variáveis independentes.

Esta ferramenta, segundo o referido autor, apresenta como vantagens a simplicidade, a capacidade de produzir modelos simples de compreender e a eficiência computacional, quando comparada com métodos tradicionais de modelagem. Como desvantagens tem-se uma representação muito simples dos fenômenos observados e o grande consumo de tempo para obtenção de respostas ótimas.

Outro *model breeder* foi implementado por Santa Catarina *et al.* (2005). O diferencial desta implementação diz respeito à forma de codificação utilizada nos cromossomos genéticos. Na proposta de Openshaw e Openshaw (1997) utilizou-se codificação binária, com as informações codificadas em cadeias de 0 e 1. Nessa implementação utilizou-se uma codificação híbrida envolvendo cadeias binárias e números reais em base decimal para representar o polinômio geral descrito na Equação 2.6.

$$y = c_1 \cdot x_1^{Exp_1}\, op_1\, c_2 \cdot x_2^{Exp_2} op_2 \cdots op_{n-1} c_n \cdot x_n^{Exp_n} \tag{2.6}$$

onde y é a variável dependente, c_i são os coeficiente de cada termo do polinômio, x_i são as n variáveis independentes, Exp_i são os expoentes das variáveis independentes e op_i são os operadores que relacionam os termos do polinômio (+, -, x, /). O cromossomo que representa um polinômio como o da Equação 2.6 é apresentado na Figura 2.8.

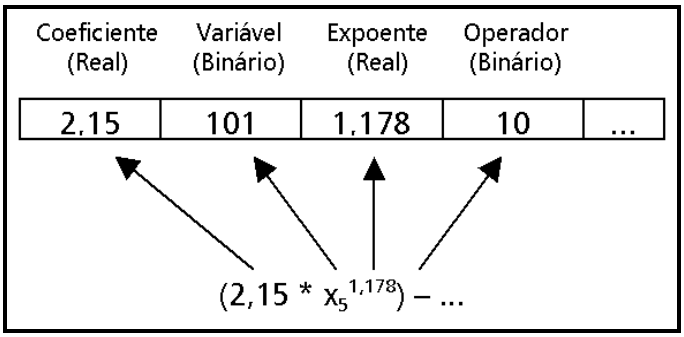

Figura 2.8 – Codificação empregada no *model breeder*
Fonte: Santa Catarina *et al.* (2005)

A função que mede a aptidão do cromossomo (Equação 2.7), e que deve ser maximizada, baseou-se na soma dos quadrados dos desvios. O modelo de maior aptidão é aquele que apresenta o menor valor para a referida soma.

$$Fitness_k = \frac{Min\left(SQT_1, SQT_2, \cdots, SQT_{Tp}\right)}{SQT_k} \tag{2.7}$$

onde $Fitness_k$ é o grau de aptidão da k-ésima solução, com k = 1..Tp, Tp é o tamanho da população avaliada e SQT é o somatório dos quadrados dos desvios total, calculado através da Equação 2.8.

$$SQT = \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 \tag{2.8}$$

onde Y_i é o valor assumido pela variável dependente na amostra i, $\hat{Y}_i$ é o valor estimado para a variável dependente na amostra i e n é o número total de amostras.

Segundo o autor, a escolha da função (Equação 2.7) foi motivada por sua simplicidade e pela capacidade de medir adequadamente o ajuste do modelo encontrado. Os testes realizados conduziram a bons resultados, permitindo concluir que a ferramenta era adequada para realizar análise exploratória dos dados.

2.4 *Species Distribution Models*

Cada vez mais ecólogos utilizam *Species Distribuition Models* (SDM) para estudar problemas teóricos e práticos como estudos sobre perda de biodiversidade (Polasky e Solow, 2001), predição dos efeitos das mudanças climáticas sobre espécies (Midgley *et al.*, 2002), busca por novas populações de espécies conhecidas ou espécies similares (Raxworthy *et al.*, 2003), identificação de regiões com potencial para conservação de espécies raras ou ameaçadas (Austin e Meyers, 1996; Araújo e Williams, 2000; Engler *et al.*, 2004) além de determinar os melhores locais para reintrodução de espécies (Hirzel e Guisan, 2002).

Os SDM utilizam-se da modelagem matemática, aliada às ferramentas computacionais, para prever a ocorrência de espécies através da geração de superfícies temáticas, indicando presença ou ausência (Guisan e Thuiller, 2005). A Figura 2.9 apresenta a estrutura geral de um sistema para geração de SDM.

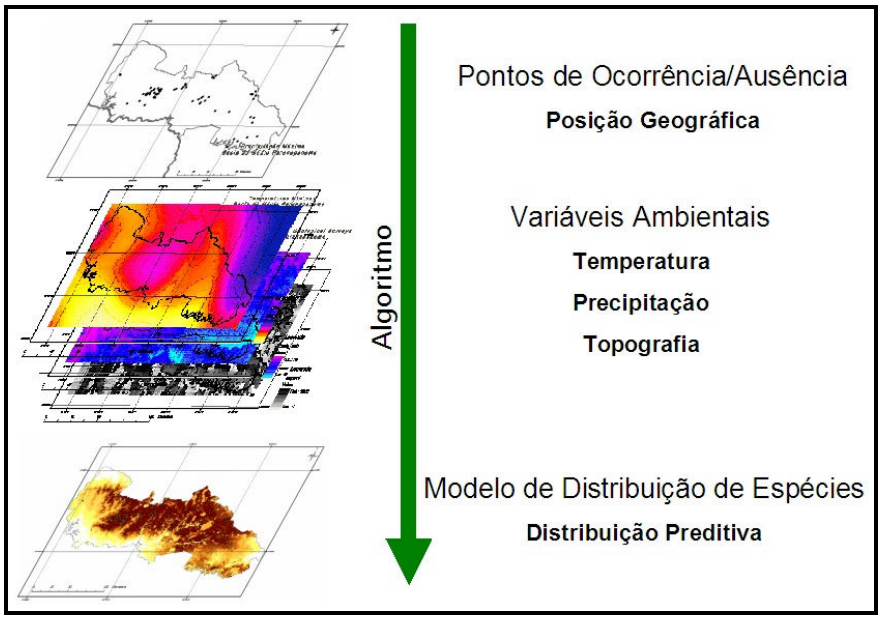

Figura 2.9 – Estrutura geral de um sistema para geração de SDM
Fonte: Siqueira (2005)

Há três pilares no estudo de modelos matemáticos aplicados à ecologia: generalidade, realidade e precisão (Guisan e Zimmermann, 2000). Desses são derivados três grupos de modelos, onde em cada grupo dois desses aspectos devem ser enfocados em detrimento do terceiro, conforme ilustrado na Figura 2.10.

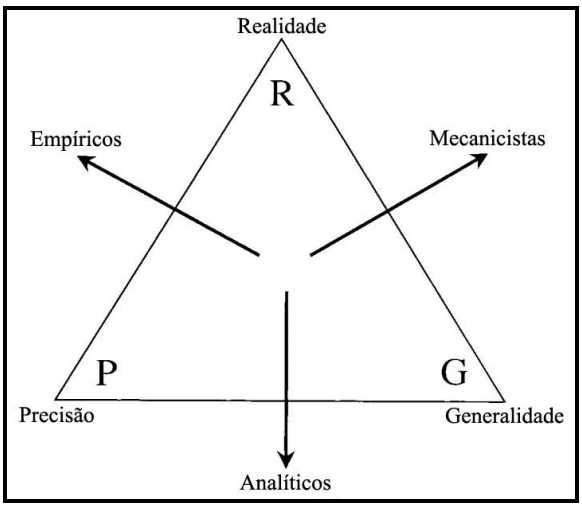

Figura 2.10 – Classificação dos modelos matemáticos aplicados em ecologia
Fonte: Adaptado de Guisan e Zimmerman (2000)

Os modelos do primeiro grupo são chamados analíticos e focam a generalidade e a precisão, como as equações de crescimento populacional logístico e as de Lotka-Volterra (Volterra, 1926). Os do segundo grupo são chamados mecanicistas, fisiológicos, casuais ou modelos de processos; são desenvolvidos visando ser realista e generalista e suas previsões são baseadas nas relações reais de causa e efeito. Os modelos do terceiro grupo sacrificam a generalidade pela precisão e realidade, são os chamados modelos empíricos, estatísticos ou fenomenológicos.

Os SDM são geralmente empíricos, pois são baseados em amostras de campo (realidade) e são aplicados especificamente para modelar a ocorrência de uma espécie numa determinada área de estudo, através de métodos estatísticos e/ou computacionais (Guisan e Zimmermann, 2000; Iwashita, 2007).

A formulação matemática utilizada nos modelos empíricos não objetiva descrever de modo realista as relações de causa e efeito entre os parâmetros do modelos e a resposta predita, nem descrever os mecanismos e funções ecológicas básicas, sendo seu principal objetivo aglutinar fatos empíricos (Wissel, 1992).

Os estudos que envolvem SDM possuem três componentes básicos: 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).

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 as condições ambientais sejam adequadas; c) as condições ambientais são realmente inadequadas 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).

Um dos sistemas para geração de SDM mais utilizados é o GARP (vide Anexo A). Como dados de entrada, o GARP usa um conjunto de pontos amostrais onde a espécie ocorre e um conjunto de *layers* geográficos que representam os parâmetros ambientais que podem delimitar a sobrevivência da espécie.

O algoritmo GARP utiliza métodos de modelagem baseados em inteligência

artificial para desenvolver modelos que consistem num conjunto de regras, ou sentenças se-então, que descrevem o nicho potencial da espécie. As regras representam um conjunto de relacionamentos multivariados entre as ocorrências da espécie e as variáveis ambientais, incluindo envelopes ambientais (regras BIOCLIM), regressão logística e regras categóricas (Stockwell, 1999; Stockwell e Peters, 1999; Payne e Stockwell, 2001; Stockwell e Peterson, 2002).

2.4.1 Avaliação de Modelos

O método de avaliação mais utilizado nos sistemas de geração de SDM é a matriz de confusão de acertos e erros associados à previsão dos modelos. Esta matriz é utilizada para quantificar a qualidade do modelo ajustado e seu formato é apresentado na Figura 2.11.

	Amostras	
Previsão (Modelo)	Presente	Ausente
Presente	VP	FP
Ausente	FN	VN

Figura 2.11 – Matriz de confusão

Os valores VP (Verdadeiro Positivo) e VN (Verdadeiro Negativo) são predições corretas. FP (Falso Positivo) e FN (Falso Negativo) são considerados erros de predição. Os erros do tipo FP também são conhecidos como erros de comissão ou superestimativa, enquanto os erros do tipo FN são conhecidos como erros de omissão. A Figura 2.12 apresenta a simulação de um modelo onde os dois tipos de erros estão presentes.

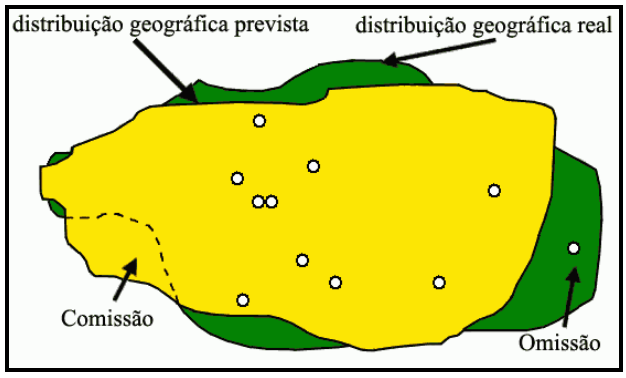

Figura 2.12 – Simulação de um modelo com erros de comissão e omissão
Fonte: Adaptado de Siqueira (2005)

A análise da matriz de confusão é essencial para evitar modelos com superestimativa (Figura 2.13a), super-ajuste (Figura 2.13b) ou alta taxa de erros de omissão (Figura 2.13c) (Meyer, 2005).

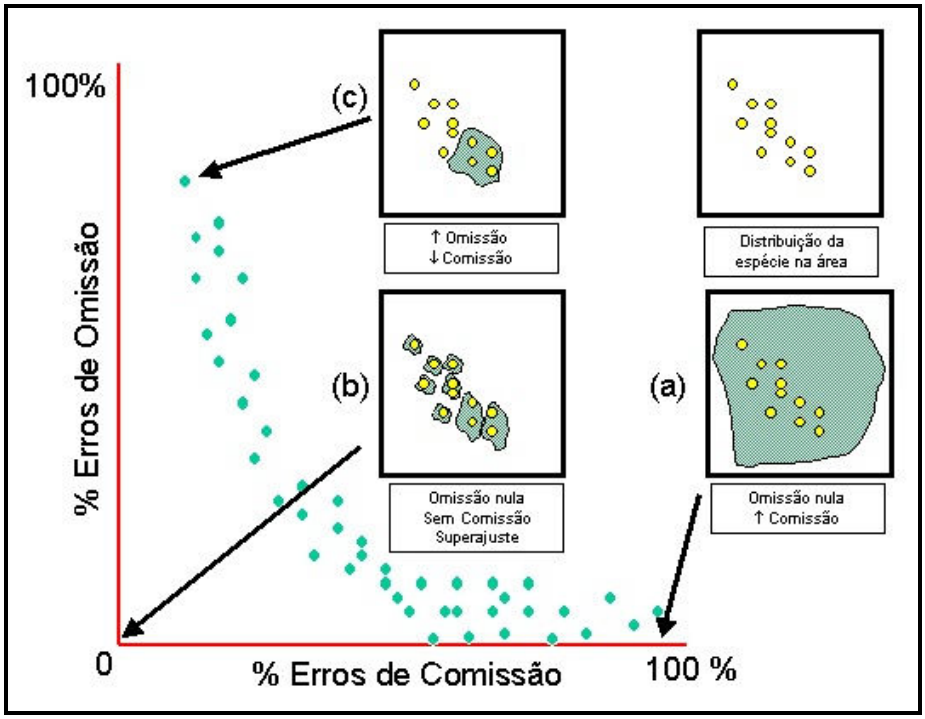

Figura 2.13 – Avaliação dos modelos quanto aos erros de comissão e omissão
Fonte: Adaptado de Meyer (2005)

Os erros de comissão não são considerados erros graves podendo ser causados por diversos fatores:

a) As condições ambientais da área são adequadas mas a espécie não foi observada; a espécie pode ser encontrada na área;

b) As condições ambientais da área são adequadas à espécie, mas fatores topológicos e/ou biológicos impedem que a espécie ocupe a área;

c) A área é mesmo inadequada, caso de um verdadeiro erro do tipo FP.

Os erros do tipo FN (omissão) são considerados erros graves. Ou seja, um local onde se sabe que a espécie é encontrada e está sendo predito como ausente. O super-ajuste também prejudica a utilidade do modelo, visto que muitos trabalhos visam projetar modelos para outras áreas ou condições climáticas (Iwashita, 2007).

Algumas métricas, derivadas da matriz de confusão, e utilizadas na avaliação de SDM são apresentadas na Tabela 2.2.

Tabela 2.2 – Métricas derivadas da matriz de confusão

Métrica	**Cálculo**
Acurácia	$(VP+VN)/(VP+FP+FN+VN)$
Sensibilidade	$VP/(VP+FN)$
Especificidade	$VN/(FP+VN)$
Taxa de falso positivo (comissão)	$FP/(FP+VN)$
Taxa de falso negativo (omissão)	$FN/(VP+FN)$
Coeficiente de correlação de Matthews	$\frac{(VP\cdot VN-FP\cdot FN)}{\sqrt{(VP+FP)(VP+FN)(VN+FP)(VN+FN)}}$

Fonte: Matthews (1975) e Fielding e Bell (1997)

A acurácia mede o acerto global do modelo. A sensibilidade é uma medida que descreve a probabilidade de uma amostra ser corretamente classificada como presença. Especificidade é a probabilidade de uma amostra ser corretamente classificada como ausência (Fielding e Bell, 1997; Guisan e Zimmermann,

2000; Segurado e Araújo, 2004).

O coeficiente de correlação de Matthews (CCM) é utilizado em aprendizagem de máquina como uma medida de qualidade em classificações binárias (Matthews, 1975). Este coeficiente considera todas as informações advindas da matriz de confusão e é uma medida balanceada que pode ser utilizada mesmo que as classes possuam diferentes tamanhos. Ele retorna um valor entre –1 e +1. CCM igual a +1 equivale a uma predição perfeita, 0 corresponde à predição aleatória e –1 significa uma predição inversa. Apesar de não existir métrica perfeita para descrever a matriz de confusão com um único número, o CCM é considerado uma das melhores métricas com este objetivo (Baldi *et al.*, 2000).

Outro método utilizado na avaliação de SDM é a curva ROC (*Receiver Operating Characteristic*). A curva ROC é representado num gráfico de sensibilidade x (1 – especificidade), ou Taxa de VP x Taxa de FP, conforme observado na Figura 2.14.

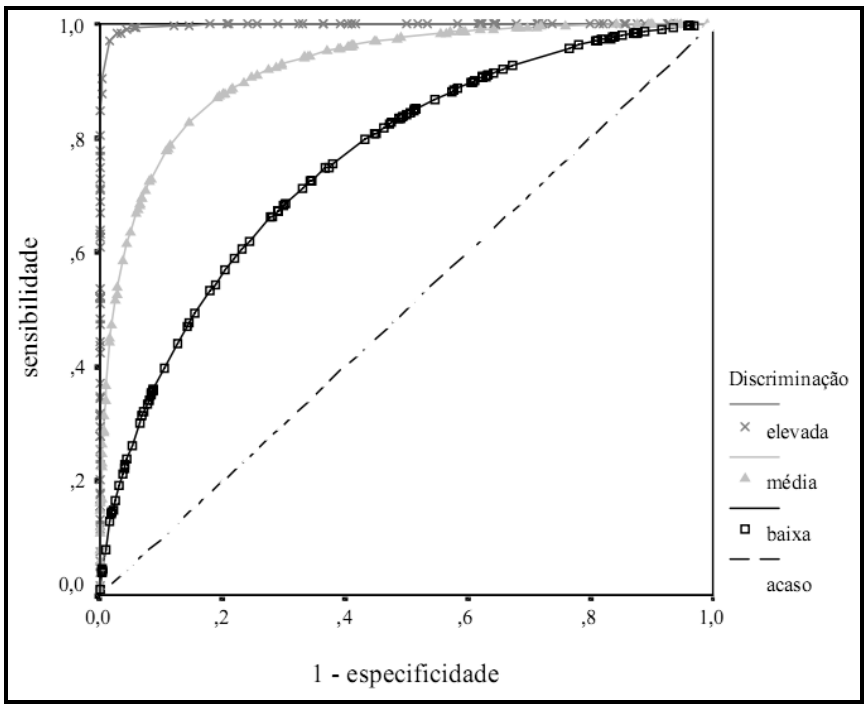

Figura 2.14 – Curvas ROC para três graus de capacidade de discriminação
Fonte: Adaptado de Braga (2000)

A área sob a curva ROC (AUC – *Area Under the Curve*) é a medida utilizada para sumarizar a qualidade da curva. Quanto mais a AUC aproximar-se de 1 melhor o desempenho, sendo que AUC = 0,5 equivale a uma predição aleatória (Braga, 2000; Rushton *et al.*, 2004; Phillips *et al.*, 2006). Este método é bastante utilizado porque é uma medida global de desempenho independente de limites de corte, geralmente empregados na construção da matriz de confusão (Deleo, 1993).

3 SAHGA MB – *MODEL BREEDER*

Model Breeders são sistemas semi-automáticos que analisam conjuntos de dados com variáveis dependentes e independentes para encontrar modelos que as relacionam. Os sistemas implementados por Openshaw e Openshaw (1997) e Santa Catarina *et al.* (2005) utilizavam AG para encontrar os modelos otimizados; entretanto ambos ignoravam os relacionamentos espaciais presentes nos dados analisados.

Para incorporar relacionamentos espaciais aos AG, utilizados na construção de *Model Breeders*, desenvolveu-se o SAHGA, um algoritmo heurístico híbrido adaptável para usos múltiplos, onde os relacionamentos espaciais são representados explicitamente através de uma GPM.

Neste capítulo é apresentado o SAHGA MB, um sistema que utiliza o algoritmo SAHGA na construção de um *Model Breeder* que considera os relacionamentos espaciais existentes nos dados analisados. Também é apresentado um estudo de caso, onde aplicou-se o sistema desenvolvido na análise de dados sócio-econômicos, visando averiguar a influência que os relacionamentos espaciais exercem sobre os modelos ajustados.

3.1 Estrutura Geral do Sistema SAHGA MB

O sistema SAHGA MB foi concebido para analisar conjuntos de dados com variáveis dependente e independentes procurando modelos que as relacionam. O diferencial deste *Model Breeder* está na incorporação de uma GPM, utilizada para representar explicitamente os relacionamentos espaciais presentes nos dados analisados.

A estrutura geral do sistema SAHGA MB é apresentada na Figura 3.1. Nesta figura observa-se 1 *layer* com 11 objetos geoespaciais, correspondente à variável dependente, 3 pares de *layers* correspondentes aos coeficientes do modelo e às variáveis independentes e 1 *layer* correspondente aos valores estimados pelo modelo ajustado. Para cada objeto geoespacial, numericamente identificado, está associado o valor da variável dependente, os valores das variáveis independentes e seus relacionamentos espaciais. O conjunto dos objetos geoespaciais e seus relacionamentos constituem a GPM.

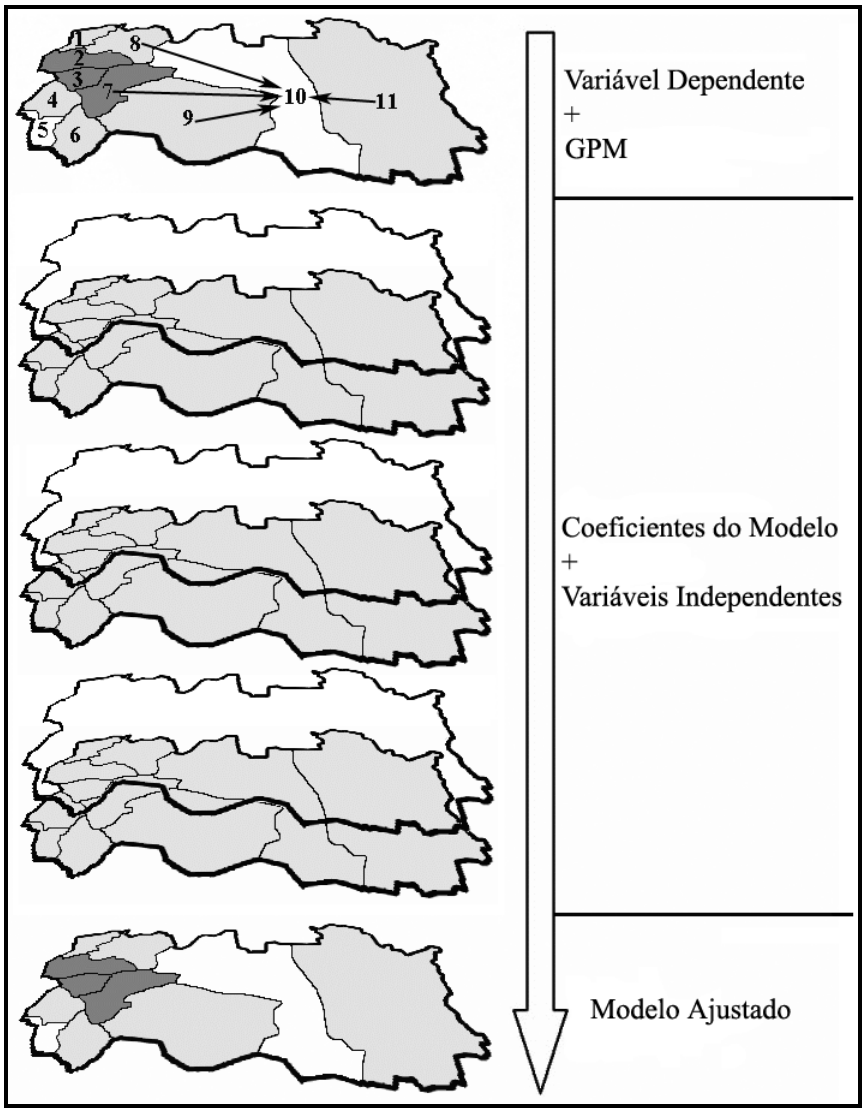

Figura 3.1 – Estrutura geral do sistema SAHGA MB

O sistema recebe como dados de entrada a GPM e os valores das variáveis dependente e independentes. Durante o processamento o conjunto de coeficientes do modelo é otimizado segundo a função de aptidão empregada no SAHGA MB. Como resultado final tem-se o conjunto de coeficientes que descrevem o modelo multivariado que relaciona a variável dependente com as variáveis independentes.

A GPM empregada no SAHGA MB atende a dois objetivos: incorporar os relacionamentos espaciais existentes entre os objetos geoespaciais e representar, através das relações de vizinhança W_{ij}, o conhecimento prévio acerca dos relacionamentos entre os objetos geoespaciais. Por exemplo, se a presença de uma via expressa limitar o relacionamento entre dois objetos geoespaciais, então a relação de vizinhança será $W_{ij} = 0$. Porém, se a existência de uma passarela entre ambos fortalecer seu relacionamento, então a relação de vizinhança será $W_{ij} = 2$.

Os coeficientes do modelo, representados pelos *layers* associados às variáveis independentes (Figura 3.1), quantificam o efeito destas variáveis sobre a variável dependente.

3.2 Representação dos Dados de Entrada

Os dados de entrada para o SAHGA MB são: a GPM e os valores das variáveis dependente e independentes. Para cada objeto geográfico haverá um conjunto de dados de entrada como representado na Figura 3.2.

i	*NRE*	*REsp*	W_{ij}	y_i	x_{1i}	x_{2i}	...	x_{ni}

Figura 3.2 – Estrutura dos dados de entrada

i é um descritor numérico que identifica o objeto geoespacial O_i; *NRE* é o número de objetos aos quais o objeto O_i está relacionado; *REsp* é um conjunto com *NRE* objetos espacialmente relacionados com o objeto O_i; W_{ij} é um conjunto com *NRE* valores, onde cada valor W_{ij} quantifica a relação de vizinhança entre os objetos O_i e O_j; y_i é o valor da variável dependente e x_{ki} são os valores das variáveis independentes, com k = 1..n, onde n é o número de variáveis independentes.

A Figura 3.3 exemplifica a utilização da estrutura dos dados de entrada, considerando o objeto geoespacial de número 10 na Figura 3.1.

i	*NRE*	*REsp*	W_{ij}	y_i	x_{1i}	x_{2i}	x_{3i}
10	5	7; 8; 9; 10; 11	1; 2; 1; 1; 2	1,318	0,508	0,705	1,103

Figura 3.3 – Um exemplo de uso da estrutura dos dados de entrada

O objeto com i = 10 possui *NRE* = 5 relacionamentos de vizinhança, com os objetos enumerados em *REsp* = {7; 8; 9; 10; 11}. Os pesos destes relacionamentos são W_{ij} = {1; 2; 1; 1; 2}. O valor da variável dependente para o objeto i = 10 é y_{10} = 1,318 e os valores das variáveis independentes x_{k10} = {0,508; 0,705; 1,103}.

3.3 Codificação, Avaliação da Aptidão e Operadores Genéticos

Num AG o processo evolutivo ocorre sobre os cromossomos. Os cromossomos codificam as possíveis soluções para o problema. No algoritmo SAHGA cada cromossomo é formado pelo conjunto de coeficientes de um modelo multivariado; portanto, são estes coeficientes que sofrem o processo evolutivo. A GPM é um dado de entrada estático; ela representa o conhecimento sobre os relacionamentos espaciais que não mudam durante o processo evolutivo.

A codificação utilizada é a real; cada gene do cromossomo corresponde ao coeficiente (β_k) associado à uma variável independente. O último gene do cromossomo corresponde à constante do modelo (β_0). A Figura 3.4 mostra a codificação empregada nos cromossomos do algoritmo SAHGA.

β_1	β_2	…	β_n	β_0
1,453	2,317	…	-2,112	0,015

Figura 3.4 – Codificação cromossômica empregada no algoritmo SAHGA

Para cada objeto geoespacial *i* o valor estimado para a variável dependente ($\hat{y}_i$), é calculado através da Equação 3.1.

$$\hat{y}_i = \beta_0 + \sum_{k=1}^{n}\left(\beta_k \cdot \left(\sum_{j=1}^{NRE_i}\left(W_{ij} \cdot x_{kj}\right) \Big/ \sum_{j=1}^{NRE_i} W_{ij} \right)\right) \tag{3.1}$$

onde β_k é o coeficiente da variável independente x_k; W_{ij} é o peso da relação de vizinhança entre os objetos O_i e O_j; x_{kj} é o valor da *k*-ésima variável independente associada ao objeto O_j; β_0 é a constante do modelo; *n* é o número de variáveis independentes; NRE_i é o número de relacionamentos espaciais do objeto O_i e *j* corresponde ao *j*-ésimo elemento em $REsp_i$.

De acordo com a Equação 3.1, para cada objeto geográfico as variáveis independentes x_k são estimadas através de uma média ponderada, considerando as relações de vizinhança W_{ij} descritas na GPM.

No SAHGA MB a aptidão de um cromossomo é calculada pela soma dos erros ao quadrado, conforme a Equação 3.2, onde *m* é o número total de objetos geográficos. O objetivo final do SAHGA MB é minimizar o valor da Aptidão.

$$Aptidão = \sum_{i=1}^{m} (y_i - \hat{y}_i)^2 \quad (3.2)$$

A soma dos erros ao quadrado é essencial na avaliação da qualidade de um modelo ajustado. Quanto menor a soma dos erros ao quadrado, melhor o ajuste do modelo.

A Figura 3.5 apresenta o núcleo de otimização do algoritmo SAHGA, mostrando a hibridização do AG com a heurística SA.

```
void AGSA::Run(){
  CreatePop(); //Criando a população inicial
  EvalPopAG(); //Avaliando a população inicial

  for (i = 0; i < NumCiclos; ++i){
    EvolPopAG(); //Gerando nova população numa iteração do AG
    EvalPopAG(); //Avaliando a nova população na iteração do AG

    while (TAtual > TMinima){ //Núcleo do SA
      EvolPopSA(); //Executando uma iteração do SA
      EvalPopSA(); //Avaliando a população na iteração do SA
    }
    População(0) = GBest; //Restaura o melhor indivíduo
    ResetTAtual(); //Ajusta a temperatura no SA --> TAtual = TMax
  }
}
```

Figura 3.5 – O núcleo de otimização do SAHGA

O primeiro processo realizado no núcleo de otimização é a geração aleatória de uma população inicial. O próximo processo é a avaliação da aptidão de cada indivíduo da população criada. Na seqüência iniciam-se os ciclos de otimização; em cada ciclo há uma iteração do AG seguido de uma execução do algoritmo SA.

A iteração do AG realiza a evolução da população através da seleção, seguida pelas operações de cruzamento e mutação. Os operadores genéticos utilizados foram o cruzamento aritmético de Michalewicz (1996) e a mutação uniforme,

com limites entre –4 e 4. O método de seleção empregado foi a roleta, onde a probabilidade de seleção, por tratar-se de um problema de minimização, é inversamente proporcional a aptidão de cada cromossomo. Para evitar a perda dos melhores indivíduos, uma elite é preservada; estes elementos não sofrem cruzamento ou mutação e são copiados diretamente para a nova população. Após evoluir a população numa iteração do AG, novamente estima-se a aptidão dos indivíduos da população.

Em cada execução do SA, dentro de um ciclo de otimização, novos indivíduos são gerados e avaliados. O esquema de busca local empregado no SA também está baseado na mutação uniforme; a cada coeficiente representado no cromossomo adiciona-se um valor aleatório entre -0,5 e 0,5.

Na conclusão de cada ciclo de otimização a melhor solução, encontrada até o momento, retorna à população e alguns parâmetros do algoritmo SA são reajustados para os valores iniciais. O algoritmo SAHGA pára após a realização de um número pré-definido de ciclos, definido no sistema.

Qualquer sistema que utilize o algoritmo SAHGA usará o núcleo de otimização apresentado na Figura 3.5; somente a função de aptidão deverá ser rescrita, refletindo os objetivos para os quais o sistema foi construído.

Os parâmetros genéticos: tamanho da população, taxas de cruzamento e mutação, bem como os parâmetros do algoritmo SA: temperatura mínima, temperatura máxima, constante de resfriamento e número de repetições, devem ser ajustados no sistema.

Alguns conjuntos pré-definidos para estes parâmetros estão disponíveis no algoritmo SAHGA. Estes conjuntos foram criados seguindo recomendações obtidas em Grefenstette (1986), De Jong e Spears (1991) e Santa Catarina e

Bach (2003) e refinados através dos testes realizados com o algoritmo. Os conjuntos pré-definidos e seus valores são apresentado na Tabela 3.1.

Tabela 3.1 – Conjuntos de parâmetros pré-definidos do algoritmo SAHGA

Parâmetros	**Conjunto de parâmetros**				
	Default	***Fast***	***Hard***	***Ultra***	***HighPop***
Tamanho da população	50	20	100	200	500
Número de ciclos	10	5	20	50	20
Temp. mínima	0,001	0,001	0,001	0,001	0,001
Temp. máxima	3	3	3	3	3
Constante de resfriamento	0,9	0,9	0,9	0,9	0,75
Número de repetições	5	3	5	10	5
Tamanho Elite	1	1	1	1	1
Taxa cruzamento	80%	80%	80%	80%	80%
Taxa mutação	1%	1%	1%	1%	2%

3.4 Estudo de Caso

Neste estudo de caso aplicou-se o sistema desenvolvido, o SAHGA MB, na análise de um conjunto de dados sócio-econômicos. Os conjuntos de dados utilizados provêm do Censo Demográfico 2000 (IBGE, 2007) e são apresentados na Tabela 3.2.

Os testes aqui realizados visam encontrar um modelo que relacione a variável dependente y = “número de filhos nascidos vivos”, com as variáveis independentes x_1 = “número de domicílios com banheiro”, x_2 = “número de domicílios cujo responsável tem 8 ou mais anos de estudo” e x_3 = “número de domicílios onde a renda é maior que 3 salários mínimos”.

Convém ressaltar, neste momento, que os modelos ajustados não pretendem explicar a variação da variável dependente em função do conjunto de variáveis independentes utilizado. Os estudos de caso visam, tão somente, demonstrar a

aplicabilidade do sistema SAHGA MB e verificar se os relacionamentos espaciais exercem alguma influência sobre os modelos ajustados.

Tabela 3.2 – Dados utilizados no estudo de caso com o SAHGA MB

Estado	Id	NRE	REsp				W_{ij}				y	x_1	x_2	x_3
RO	1	2	1	2			1	1			1202863	209451	274875	174857
AC	2	2	2	1			1	1			519328	49142	106099	53380
AM	3	1	3				1				2423692	300476	611136	250182
RR	4	1	4				1				265621	47634	83363	39898
PA	5	1	5				1				5538281	580556	1211265	537327
AP	6	1	6				1				405330	52668	126069	51216
TO	7	1	7				1				1063302	159651	232773	101387
MA	8	1	8				1				5577060	403502	912504	307044
PI	9	1	9				1				2926783	316964	457217	175372
CE	10	1	10				1				7693171	1013537	1433765	552265
RN	11	1	11				1				2935527	481658	612604	244271
PB	12	2	12	13			1	1			3847081	595675	619956	260108
PE	13	2	13	12			1	1			8015103	1446109	1753714	703468
AL	14	1	14				1				2982700	427277	443524	188763
SE	15	1	15				1				1770399	335513	344433	144054
BA	16	2	16	17			1	1			12761579	2095830	2499704	1049684
MG	17	4	17	16	19	20	1	1	1	1	15763185	4329948	4840801	2574913
ES	18	1	18				1				2677347	786522	921201	468228
RJ	19	3	19	17	20		1	1	1		11131946	4114484	5496427	2797601
SP	20	4	20	17	19	21	1	1	1	1	28499542	10176557	13662174	7543605
PR	21	3	21	20	22		1	1	1		8313533	2419203	2992480	1569988
SC	22	3	22	21	23		1	1	1		4538950	1397872	1651501	1025313
RS	23	2	23	21			1	1			8266032	2780508	3252382	1909129
MS	24	2	24	25			1	1			1790394	502762	546199	296185
MT	25	2	25	24			1	1			2075588	508834	622448	356853
GO	26	1	26				1				4218333	1243066	1329209	729428
DF	27	1	27				1				1499905	518354	865115	401411

Para utilização no SAHGA MB, os dados da Tabela 3.2, referentes às variáveis dependentes e independentes, foram padronizados através da Equação 3.3. A padronização uniformiza a escala das variáveis, fazendo com que elas sejam tratadas igualitariamente pelo algoritmo SAHGA.

$$x_p = \frac{x - \bar{x}}{s} \tag{3.3}$$

onde x_p é a variável padronizada, x é o valor da variável de entrada, $\bar{x}$ e s são, respectivamente, a média e o desvio padrão amostral da variável de entrada x.

Os relacionamentos espaciais e os pesos W_{ij} apresentados na Tabela 3.2, correspondentes à GPM, foram definidos arbitrariamente. Imaginou-se que alguns Estados da Federação exercem algum tipo de influência sobre seus vizinhos mais próximos. Por exemplo, São Paulo exerce influência sobre os Estados de Minas Gerais, Rio de Janeiro e Paraná, bem como é influenciado por estes.

Realizou-se dois testes com os dados apresentados na Tabela 3.2. No primeiro deles ajustou-se um modelo multivariado desconsiderando os relacionamentos espaciais representados na GPM; assumiu-se que cada estado relaciona-se apenas consigo mesmo, com peso W_{ii} = 1. No segundo teste ajustou-se um modelo multivariado considerando a GPM apresentada. Em ambos os testes os parâmetros do SAHGA MB foram definidos como *Default* (Tabela 3.1) .Os resultados obtidos são apresentados nas seções seguintes.

3.4.1 Teste 1: Modelo Multivariado Desconsiderando a GPM

O modelo multivariado ajustado para os dados da Tabela 3.2, padronizados pela Equação 3.3, desconsiderando a GPM é apresentado na Equação 3.4.

$$\hat{y}_p = 0{,}0025 + 1{,}5605 \cdot x_{1p} + 2{,}7114 \cdot x_{2p} - 3{,}3185 \cdot x_{3p} \tag{3.4}$$

Este modelo apresentou aptidão mínima igual a 0,9298, correspondente à soma dos desvios quadráticos.

A interpretação direta do modelo não é trivial pois as variáveis envolvidas estão padronizadas; ou seja, possuem médias nulas e variâncias iguais a 1. Resumidamente, pode-se dizer que nos estados onde as variáveis x_1 e x_2 são maiores que suas médias e a variável x_3 é menor que sua média, o modelo prediz um número de filhos nascidos vivos maior que a média nacional; nos estados onde as variáveis x_1 e x_2 são menores que suas médias e a variável x_3 é maior que sua média, o modelo prediz um número de filhos nascidos vivos menor que a média nacional.

3.4.2 Teste 2: Modelo Multivariado Considerando a GPM

O modelo multivariado ajustado para os dados da Tabela 3.2, padronizados pela Equação 3.3, considerando a GPM é apresentado na Equação 3.5.

$$\hat{y}_p = -0{,}011 + 1{,}5397 \cdot x_{1p} + 3{,}3491 \cdot x_{2p} - 4{,}0 \cdot x_{3p} \qquad (3.5)$$

A aptidão mínima para este modelo, correspondente ao valor da soma dos desvios ao quadrado, foi igual a 1,0737.

Observando as Equações 3.4 e 3.5 percebe-se que as relações entre as variáveis independentes e a variável dependente se mantém. Entretanto, as proporções destas variações são distintas mostrando que, de fato, os relacionamentos espaciais representados na GPM exercem influência sobre os modelos ajustados.

As variáveis da Tabela 3.2, padronizadas pela Equação 3.3, os valores estimados pelos modelos ajustados (Equação 3.4 e Equação 3.5) são apresentados na Tabela 3.3.

Tabela 3.3 – Dados padronizados e valores estimados nos testes 1 e 2

Estado	y_p	x_{1p}	x_{2p}	x_{3p}	$\hat{y}_n$ (Teste 1)	$\hat{y}_n$ (Teste 2)
RO	-0,70862	-0,55628	-0,54452	-0,48229	-0,74157	-0,76196
AC	-0,82114	-0,63238	-0,60581	-0,56225	-0,76120	-0,76461
AM	-0,50765	-0,51307	-0,42239	-0,43272	-0,50753	-0,48474
RR	-0,86291	-0,6331	-0,61407	-0,57112	-0,75526	-0,75787
PA	0,005067	-0,38011	-0,20444	-0,24372	-0,33624	-0,30604
AP	-0,83991	-0,63071	-0,59856	-0,56367	-0,73420	-0,73204
TO	-0,73159	-0,57992	-0,55981	-0,53065	-0,65945	-0,65615
MA	0,01145	-0,46416	-0,31294	-0,39529	-0,25864	-0,19257
PI	-0,42483	-0,50524	-0,47829	-0,48195	-0,48349	-0,46295
CE	0,3598	-0,17456	-0,12363	-0,23389	0,17100	0,24174
RN	-0,42339	-0,42706	-0,42186	-0,43661	-0,35895	-0,33496
PB	-0,27334	-0,37293	-0,41919	-0,42618	-0,30183	-0,28437
PE	0,412796	0,03079	-0,00743	-0,13437	0,47627	0,54900
AL	-0,41563	-0,45287	-0,48327	-0,47314	-0,44450	-0,43423
SE	-0,61519	-0,49644	-0,51925	-0,50257	-0,51240	-0,50412
BA	1,19415	0,339227	0,263493	0,093499	0,93601	1,01978
MG	1,688268	1,399814	1,113729	1,097372	1,56512	1,48479
ES	-0,46589	-0,28233	-0,30978	-0,2892	-0,31837	-0,32639
RJ	0,925884	1,297528	1,351838	1,243941	1,56473	1,53848
SP	3,7849	4,175332	4,31746	4,36766	3,73074	3,40670
PR	0,461923	0,49274	0,442459	0,435952	0,52440	0,48570
SC	-0,15944	0,007891	-0,04456	0,077458	-0,36307	-0,45790
RS	0,454103	0,664259	0,53685	0,659167	0,30727	0,17305
MS	-0,6119	-0,41704	-0,44598	-0,40244	-0,52209	-0,53698
MT	-0,56495	-0,41416	-0,41829	-0,36251	-0,57502	-0,59952
GO	-0,21222	-0,0656	-0,1616	-0,11729	-0,14887	-0,18408
DF	-0,65972	-0,40964	-0,33015	-0,33318	-0,42633	-0,41471

3.5 Conclusões

O sistema SAHGA MB foi utilizado para ajustar modelos multivariados sem e com relacionamentos espaciais representados numa GPM. Os resultados encontrados (Tabela 3.3) mostram que os modelos ajustados são adequados para estimar o valor padronizado da variável “número de filhos nascidos vivos”, em função do conjunto de variáveis independentes utilizado.

A principal vantagem apresentada pelo SAHGA MB é o fato dele trabalhar com diferentes tipos de modelos, utilizando a mesma estrutura de codificação. Ou seja, mostrou-se uma solução genérica para um problema de análise de dados geoespaciais, procurando relações existentes entre variáveis dependentes e independentes, com e sem relacionamentos espaciais.

O algoritmo SAHGA apresenta outra característica favorável; ele é extensível. Nos testes realizados ajustou-se apenas modelos multivariados lineares, mas ele pode ser utilizado para ajustar modelos multivariados não lineares. Para tanto basta mudar a estrutura do cromossomo (Figura 3.4), incluindo um número maior de coeficientes, e reescrever a Equação 3.1, considerando termos não lineares.

A Figura 3.6 e a Equação 3.6 mostram as adaptações necessárias para que o SAHGA MB seja capaz de ajustar modelos com termos quadráticos.

β^2_1	β_1	β^2_2	β_2	…	β^2_n	β_n	β_0
1,453	2,144	0,654	2,317	…	1,498	6,322	0,015

Figura 3.6 – Codificação proposta para um *model breeder* com termos quadráticos

$$\hat{y}_i = \beta_0 + \sum_{k=1}^{n} \left(\beta_k^2 \cdot \left(\frac{\sum_{j=1}^{NRE_i} \left(W_{ij} \cdot x_{kj}^2 \right)}{\sum_{j=1}^{NRE_i} W_{ij}} \right) + \beta_k \cdot \left(\frac{\sum_{j=1}^{NRE_i} \left(W_{ij} \cdot x_{kj} \right)}{\sum_{j=1}^{NRE_i} W_{ij}} \right) \right) \tag{3.6}$$

Outra adaptação possível envolve apenas a substituição da Equação 3.1 pela Equação 3.7. Neste modelo os relacionamentos espaciais são aplicados sobre a variável dependente e não mais sobre as variáveis independentes.

$$\hat{y}_i = \beta_0 + \sum_{k=1}^{n}\left(\beta_k x_{ik}\right) + \lambda \frac{\sum_{j=1}^{NRE_i}\left(W_{ij} \cdot y_j\right)}{\sum_{j=1}^{NRE_i} W_{ij}}, \; com \; j \neq i \qquad (3.7)$$

onde λ é um fator de ponderação.

O capítulo seguinte relata o uso do algoritmo SAHGA no desenvolvimento de um sistema para modelagem de distribuição de espécies. A incorporação da GPM é a inovação apresentada neste sistema, permitindo computar os efeitos dos relacionamentos espaciais sobre a distribuição potencial das espécies estudadas.

4 SAHGA SDM – *SPECIES DISTRIBUTION MODELS*

Os SDM utilizam-se da modelagem matemática, aliada às ferramentas computacionais, para prever a presença ou ausência de uma espécie numa determinada área de estudo (Guisan e Thuiller, 2005; Iwashita, 2007).

Um dos sistemas para geração de SDM mais utilizados é o GARP. Como dados de entrada, o GARP usa um conjunto de pontos de presença da espécie e um conjunto de *layers* geográficos, representando os parâmetros ambientais que podem delimitar a sobrevivência da espécie. Internamente o GARP utiliza um AG para construir um conjunto de regras, ou sentenças se-então, que descrevem o nicho potencial da espécie. Assim como nos *model breeder*s, o AG utilizado no GARP ignora os relacionamentos espaciais presentes nos dados analisados.

Neste capítulo é apresentado o SAHGA SDM, um sistema para criação de SDM que emprega o algoritmo SAHGA. Também são apresentados dois estudos de caso, onde aplicou-se o sistema desenvolvido na modelagem da distribuição potencial das espécies *Strix varia* Barton (1799) e *Thalurania furcata boliviana* Boucard (1894). Em ambos os casos modelou-se a distribuição da espécie sem e com os relacionamentos espaciais e comparou-se seus resultados com os fornecidos pelos algoritmos GARP *Single Run* e GARP *Best Subset*. O objetivo é averiguar a influência que os relacionamentos espaciais exercem sobre os modelos ajustados e avaliar a qualidade dos SDM obtidos pelo sistema implementado comparando-os com os resultados fornecidos pelo GARP.

4.1 Estrutura Geral do Sistema SAHGA SDM

O sistema SAHGA SDM é um sistema que emprega o algoritmo SAHGA para

modelar a distribuição potencial de espécies. O diferencial deste sistema está na sua capacidade de construir SDM que considerem os relacionamentos espaciais presentes nos dados de entrada, representando-os através de uma GPM. A estrutura geral do sistema SAHGA SDM é apresentada na Figura 4.1.

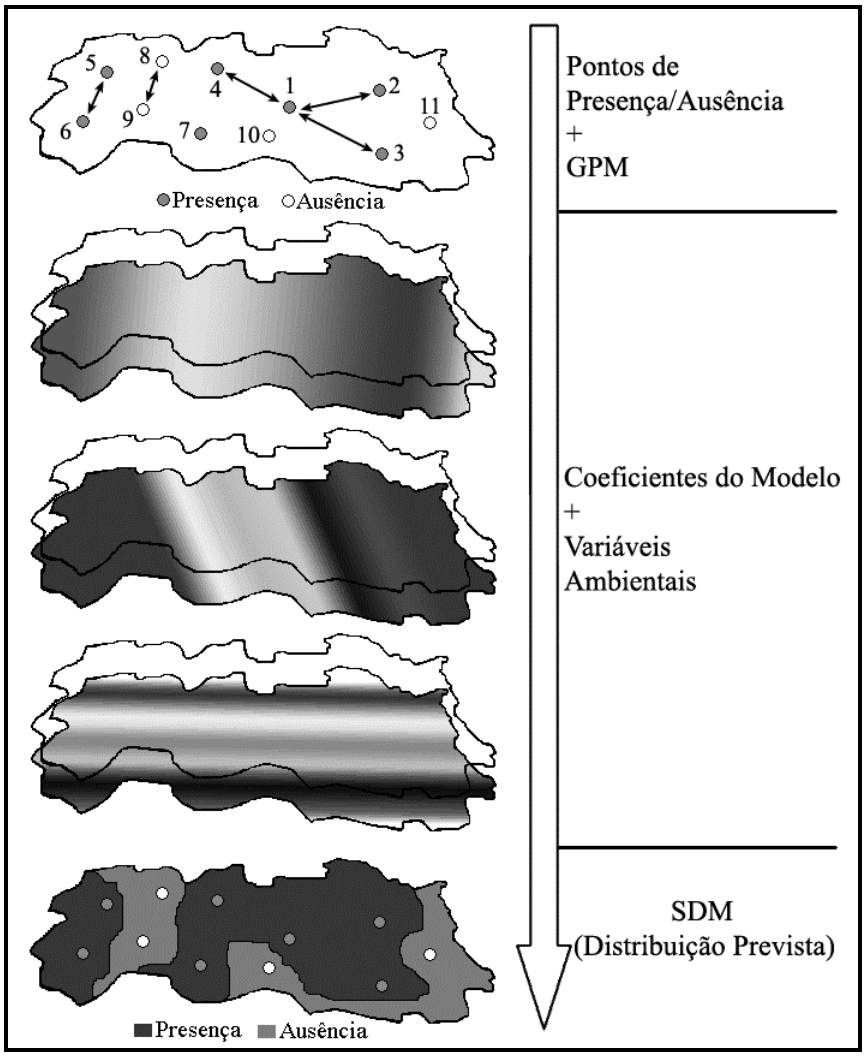

Figura 4.1 – Estrutura geral do sistema SAHGA SDM

Os coeficientes associados às variáveis ambientais quantificam o efeito destas variáveis sobre a distribuição prevista da espécie. A GPM atende a dois objetivos: incorporar os relacionamentos espaciais existentes entre os pontos e a representação do conhecimento pré-existente sobre os elementos naturais e artificiais presentes no espaço, cujos efeitos são significativos na distribuição

potencial da espécie modelada, como estradas, rios, cadeias de montanhas, etc.

Os relacionamentos são representados pela associação entre dois pontos P_i e P_j, indicados pelas setas na Figura 4.1. O conhecimento pré-existente, sobre os elementos naturais e artificiais presentes no espaço, são quantificados nos pesos W_{ij} da GPM.

A distância entre pontos amostrais é um fator que pode ser representado nos pesos W_{ij} ; quanto maior a distância entre os pontos, menor o peso do relacionamento. Outro exemplo de conhecimento representado na GPM é o caso de dois pontos amostrais separados por um rio. Se o rio dificulta a distribuição da espécie, os dois pontos podem estar relacionados com peso $W_{ij} = 0{,}5$. Caso o rio seja altamente restritivo, pode-se atribuir ao relacionamento um peso $W_{ij} = 0{,}1$. Porém, se o rio facilitar a dispersão da espécie, pode-se atribuir ao relacionamento um peso $W_{ij} = 2$.

4.2 Representação dos Dados de Entrada

Os dados de entrada para o SAHGA SDM são: pontos amostrais de presença ou ausência da espécie, com seus relacionamentos espaciais (GPM), e o conjunto de *layers* geográficos que representam as variáveis ambientais que podem delimitar a sobrevivência da espécie.

Para cada ponto de presença ou ausência da espécie haverá um conjunto de dados de entrada conforme apresentado na Figura 4.2.

P_i	*NRE*	*REsp*	W_{ij}	y	x_1	x_2	...	x_n

Figura 4.2 – Estrutura dos dados de entrada

P_i é um número que identifica o ponto amostral (ponto de presença ou ausência); *NRE* é o número de pontos aos quais o ponto P_i está relacionado; *REsp* é um conjunto com *NRE* pontos espacialmente relacionados com o ponto P_i ; W_{ij} é um conjunto com *NRE* valores, onde cada valor W_{ij} quantifica a relação de vizinhança entre os objetos P_i e P_j ; y é igual a 0 para um ponto de ausência e 1 para um ponto de presença; x_k são os valores das variáveis ambientais nos pontos P_i, com $i = 1..m$ e $k = 1..n$, onde m é o número total de pontos amostrais e n é o número de variáveis ambientais.

A Tabela 4.1 mostra um exemplo com os dados de entrada para alguns pontos da Figura 4.1. O ponto amostral P_1, por exemplo, possui *NRE* = 4 relacionamentos, com os pontos P_1, P_2, P_3 e P_4. Os pesos associados aos 4 relacionamentos são W_{ij} = {1; 0,8; 0,5; 0,8}. Este ponto equivale a uma presença da espécie (y = 1) e os valores das três variáveis ambientais a ele associadas são x_k = {12,18; 33,6; 413,5}.

Tabela 4.1 – Dados de entrada para alguns pontos da Figura 4.1

P_i	*NRE*	*REsp*	W_{ij}	y	x_1	x_2	x_3
1	4	1; 2; 3; 4	1; 0,8; 0,5; 0,8	1	12,18	33,6	413,5
2	2	1; 2	0,7; 1	1	13,12	60,2	389,1
3	2	1; 3	0,2; 1	1	14,21	55,9	425,8
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
11	1	11	1	0	6,12	80,3	130,6

Os valores das variáveis ambientais (x_k) são extraídos dos *layers* geográficos, nas coordenadas dos pontos amostrais de presença ou ausência da espécie.

4.3 Codificação, Avaliação da Aptidão e Operadores Genéticos

O sistema SAHGA SDM utiliza o algoritmo SAHGA; portanto a codificação e os operadores genéticos são os mesmos apresentados na Seção 3.3. Apenas a

função de avaliação da aptidão dos cromossomos foi adaptada para ser utilizada no SAHGA SDM. A avaliação do cromossomo para cada ponto ($\hat{y}_i$) é realizada pela Equação 4.1.

$$\hat{y}_i = \beta_0 + \sum_{k=1}^{n}\left(\beta_k \cdot \left(\sum_{j=1}^{NRE_i}\left(W_{ij} \cdot x_{kj}\right) \Big/ \sum_{j=1}^{NRE_i} W_{ij} \right)\right) \quad (4.1)$$

onde β_k é o coeficiente da variável ambiental x_k; W_{ij} é a peso da relação de vizinhança entre os pontos P_i e P_j ; x_{kj} é o valor da *k*-ésima variável ambiental associada ao ponto P_j ; β_0 é a constante do modelo; *n* é o número de variáveis ambientais; NRE_i é o número de relacionamentos espaciais do ponto P_i e *j* corresponde ao *j*-ésimo elemento em $REsp_i$.

De acordo com a Equação 4.1, para cada ponto de presença ou ausência, as variáveis ambientais x_k são ponderadas pelos pesos advindos das relações de vizinhança descritas na GPM. O valor final da Equação 4.1 corresponde à estimativa do modelo para o ponto avaliado. Se $\hat{y}_i < 0{,}5$ então assume-se que P_i será um ponto de ausência; mas se $\hat{y}_i \geq 0{,}5$ então assume-se que P_i é um ponto de presença da espécie.

A aptidão de um cromossomo é estimada pela Equação 4.2, onde *m* é o número total de pontos de presença ou ausência. O objetivo final do SAHGA SDM é minimizar o valor de Aptidão. A aptidão é uma ponderação da soma dos desvios quadráticos entre valores observados e valores estimados pelo modelo e do número total de pontos amostrais estimados como FP ou FN.

$$Aptidão = \sum_{i=1}^{m}\left(\left(y_i - \hat{y}_i\right)^2 + 0{,}1 \cdot \begin{cases} 0, se\left(\left(y_i = 0\, e\, \hat{y}_i < 0{,}5\right) ou \left(y_i = 1\, e\, \hat{y}_i \geq 0{,}5\right)\right) \\ 1, se\left(\left(y_i = 0\, e\, \hat{y}_i \geq 0{,}5\right) ou \left(y_i = 1\, e\, \hat{y}_i < 0{,}5\right)\right) \end{cases} \right) \quad (4.2)$$

Na avaliação do cromossomo acrescenta-se 0,1 à aptidão para cada ponto amostral cuja estimativa discorda do valor observado. Este valor foi definido empiricamente e corresponde a uma penalização aplicada ao cromossomo quando este comete um erro do tipo FP ou FN. Deseja-se um SDM que seja capaz de maximizar sua capacidade preditiva minimizando o número de erros na avaliação dos pontos amostrais.

4.4 Estudos de Caso

Aplicou-se o sistema SAHGA SDM na modelagem da distribuição potencial de duas espécies: *Strix varia* e *Thalurania furcata boliviana*. Para cada espécie ajustou-se dois modelos de distribuição potencial: o primeiro modelo ignora os relacionamentos espaciais entre os pontos amostrais enquanto que o segundo os considera. Objetiva-se, com os estudos de caso, averiguar a influência que os relacionamentos espaciais exercem sobre os modelos ajustados e avaliar a qualidade dos SDM obtidos pelo sistema implementado comparando-os com os resultados fornecidos pelos algoritmos GARP *Single Run* e GARP *Best Subset*, ambos implementados no software openModeller Desktop v1.0.6 (CRIA *et al.*, 2008).

Em ambos os estudos de caso construiu-se a GPM, estrutura responsável por representar os relacionamentos espaciais, utilizando-se a seguinte regra: dois pontos de presença, ou ausência, distantes entre si até 50 km estão espacialmente relacionados com peso W_{ij} = 1; se a distância entre eles for de 50 – 100 km o peso do relacionamento será W_{ij} = 0,5; pontos afastados por distâncias superiores a 100 km não estão relacionados. A Figura 4.3 exemplifica a regra empregada na construção da GPM.

Os pontos de presença 5 e 6 estão afastados por uma distância inferior a 50 km, portanto os relacionamentos entre ambos terão pesos W_{56} = W_{65} = 1. O

mesmo ocorre com os pontos de ausência 8 e 9. Já os pontos de presença 1 e 2 estão afastados por distância superior a 50 km e inferior a 100 km, portanto os relacionamentos entre eles terão pesos $W_{12} = W_{21} = 0,5$.

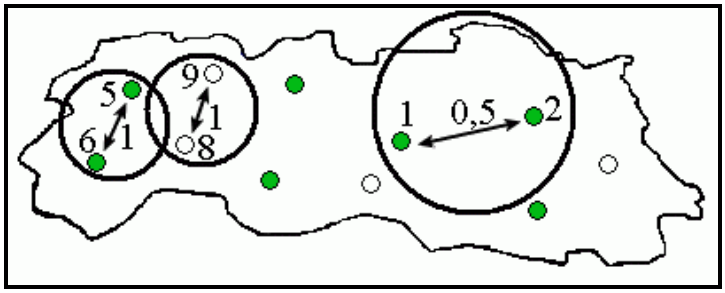

Figura 4.3 – Regra empregada na construção da GPM

Os limiares de 50 km e 100 km são meramente experimentais e visam tão somente proporcionar um critério para construir uma GPM. Um especialista, detentor de conhecimento acerca da espécie modelada e da região em estudo, pode inserir, via GPM, seu conhecimento sobre as relações existentes entre os pontos observados e os elementos que exercem influência sobre a distribuição da espécie modelada, como rios, estradas e montanhas, por exemplo.

Nos estudos de caso os parâmetros do SAHGA SDM foram definidos como *Hard* (Tabela 3.1). Os parâmetros empregados nos algoritmos GARP Single Run e GARP Best Subset foram os parâmetros padrão do software openModeller Desktop. Os resultados obtidos são apresentados nas seções seguintes.

4.4.1 Espécie *Strix varia* Barton, 1799

A base de dados *Strix varia* é a base exemplo fornecida junto ao instalador do software DesktopGarp (Kansas University, 2007). A base contém 1218 pontos de presença da referida espécie; também são disponibilizados 7 *layers*

geográficos correspondentes às variáveis: temperatura média diária, variação da temperatura, precipitação anual, número de dias úmidos, elevação do terreno, declividade do terreno e exposição. A Figura 4.4 mostra um exemplar da espécie, conhecida popularmente como coruja listrada.

Figura 4.4 – Um exemplar da espécie *Strix varia*
Fonte: Mojtahedi (2005)

Para realização dos testes selecionou-se, ao acaso, 100 pontos de presença dos 1218 fornecidos na base. Com estes 100 pontos, utilizou-se o software openModeller Desktop para gerar o SDM usando o algoritmo BIOCLIM (Nix, 1986). De posse do mapa da distribuição potencial prevista, segundo o algoritmo BIOCLIM (vide Anexo A), selecionou-se 100 pontos de pseudo-ausência, chamados a partir de agora simplesmente de ausência. A Figura 4.5 apresenta o mapa da distribuição prevista com os pontos de presença e ausência.

Dividiu-se, aleatoriamente, o conjunto com 200 pontos de presença/ausência em dois subconjuntos: treino e teste. Cada subconjunto restou com 100 pontos,

onde 50% deles são pontos de presença e os outros 50% são pontos de ausência.

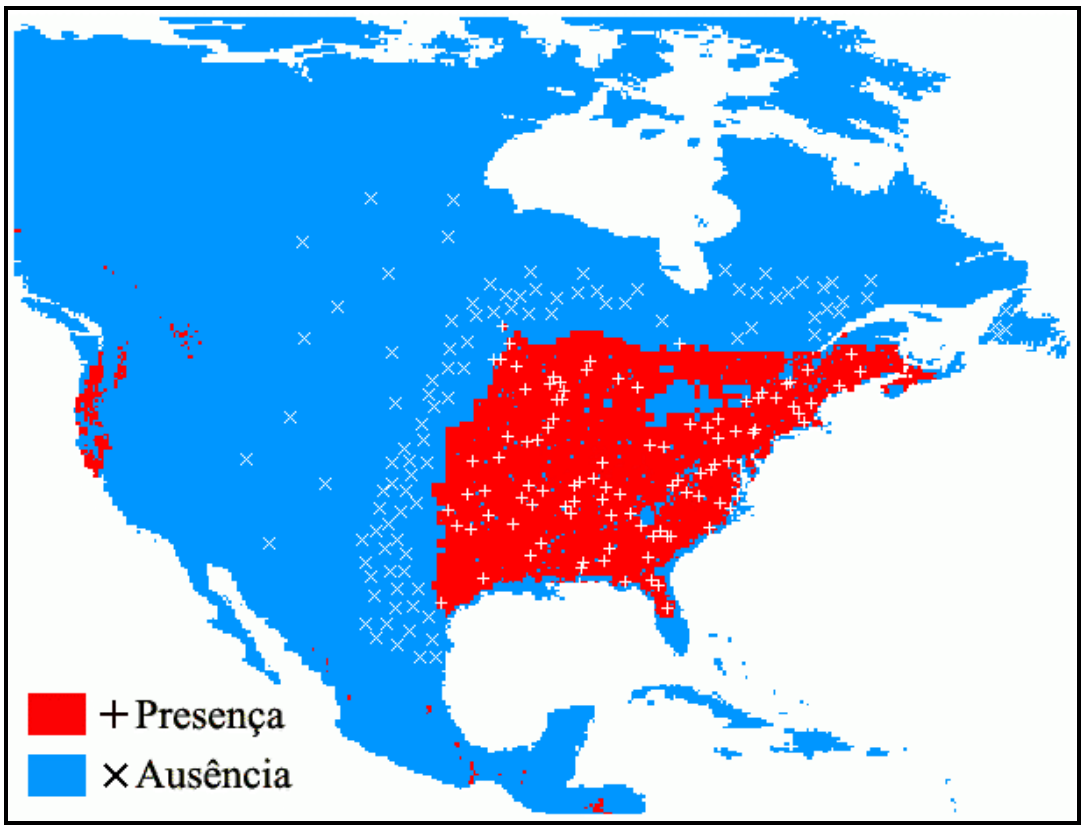

Figura 4.5 – Distribuição potencial da espécie *Strix varia* (BIOCLIM)

Com os pontos do sub-conjunto de treino extraiu-se, dos 7 *layers* geográficos, os valores das variáveis ambientais, posteriormente padronizados pela Equação 4.3.

$$xp_{ij} = \frac{x_{ij} - \bar{x}_i}{s_i} \tag{4.3}$$

onde xp_{ij} é o valor padronizado da *i*-ésima variável ambiental correspondente ao *j*-ésimo ponto; x_{ij} é o valor da *i*-ésima variável ambiental correspondente ao *j*-ésimo ponto; $\bar{x}_i$ é a média amostral da *i*-ésima variável ambiental e s_i é o desvio-padrão amostral da *i*-ésima variável ambiental.

Com o sistema SAHGA SDM ajustou-se dois SDM para a espécie *Strix varia*. No primeiro deles, chamado modelo S1, criou-se uma GPM onde não há relacionamentos espaciais entre os pontos amostrais, ou seja, cada ponto está relacionado apenas consigo mesmo, com W_{ii} = 1. No segundo, chamado modelo S2, criou-se a GPM através da regra apresentada na Seção 4.4. A GPM construída para o modelo S2 busca considerar os efeitos oriundos da dependência espacial que, neste caso, limita-se a 100 km ao redor de cada ponto de presença/ausência. Ao comparar os modelos S1 e S2 objetiva-se demonstrar que os relacionamentos espaciais influenciam na predição da distribuição potencial da espécie.

As métricas para avaliação dos modelos S1 e S2 são apresentadas na Tabela 4.2. As curvas ROC são apresentadas na Figura 4.6.

Tabela 4.2 – Métricas para avaliação dos modelos S1 e S2

Métrica	Modelo S1	Modelo S2
Acurácia	91%	91%
Erro de Omissão	6%	6%
CCM	0,821	0,821
AUC	0,957	0,964

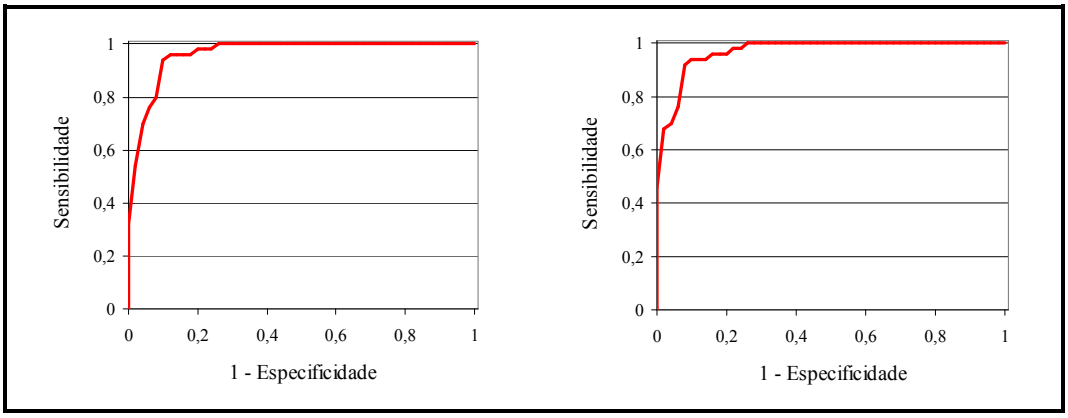

Figura 4.6 – Curvas ROC para os modelos S1 e S2

Ao observar as métricas de avaliação dos modelos S1 e S2, apresentadas na Tabela 4.2, percebe-se que eles têm qualidades similares. Ambos os modelos

possuem boa acurácia e baixo índice de erros de omissão. Quanto ao CCM observa-se que ambos os modelos têm boa capacidade preditiva, pois os valores de CCM são superiores a 0,8.

Quanto à curva ROC e ao índice AUC há uma pequena diferença entre os modelos. Nota-se que o modelo S2 possui AUC ligeiramente superior ao modelo S1; entretanto, não se pode assegurar que o modelo S2 é melhor que o modelo S1. Ambos os modelos podem ser considerados excelentes classificadores, pois apresentam AUC superiores a 0,95.

Os mapas da distribuição potencial para a espécie *Strix varia*, prevista pelos modelos S1 e S2, são apresentados na Figura 4.7 e na Figura 4.8, respectivamente.

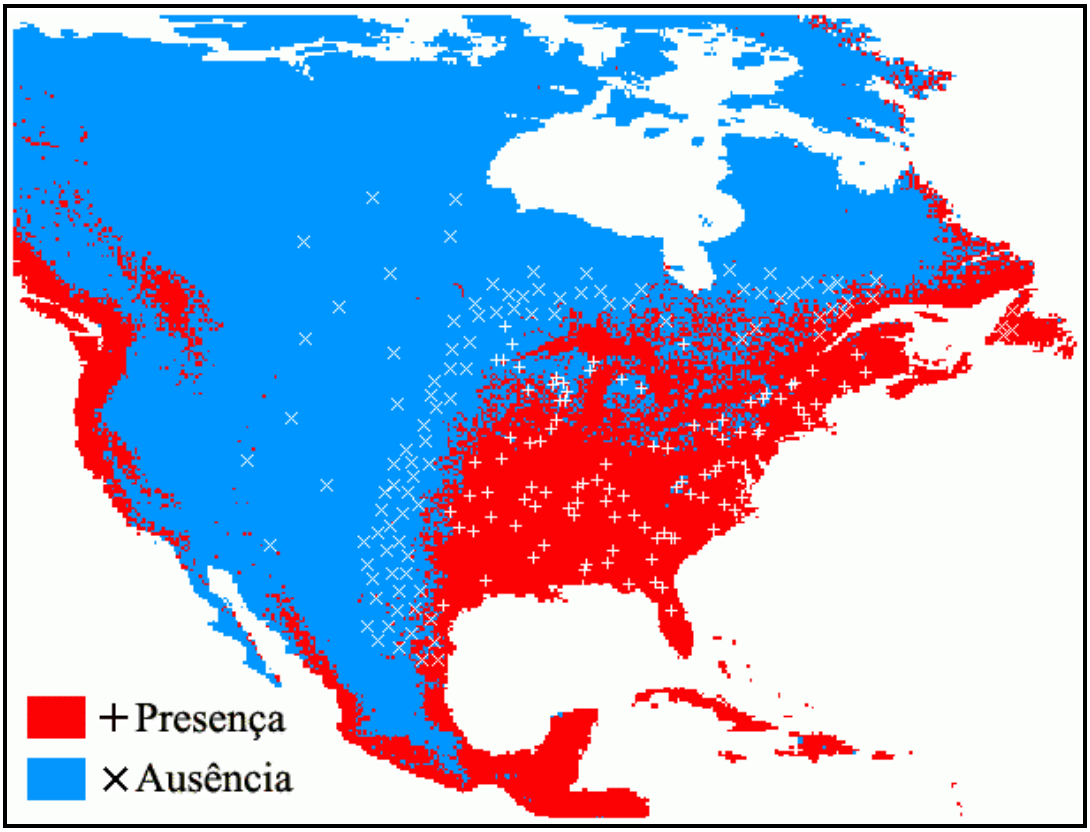

Figura 4.7 – Distribuição potencial da espécie *Strix varia* (modelo S1)

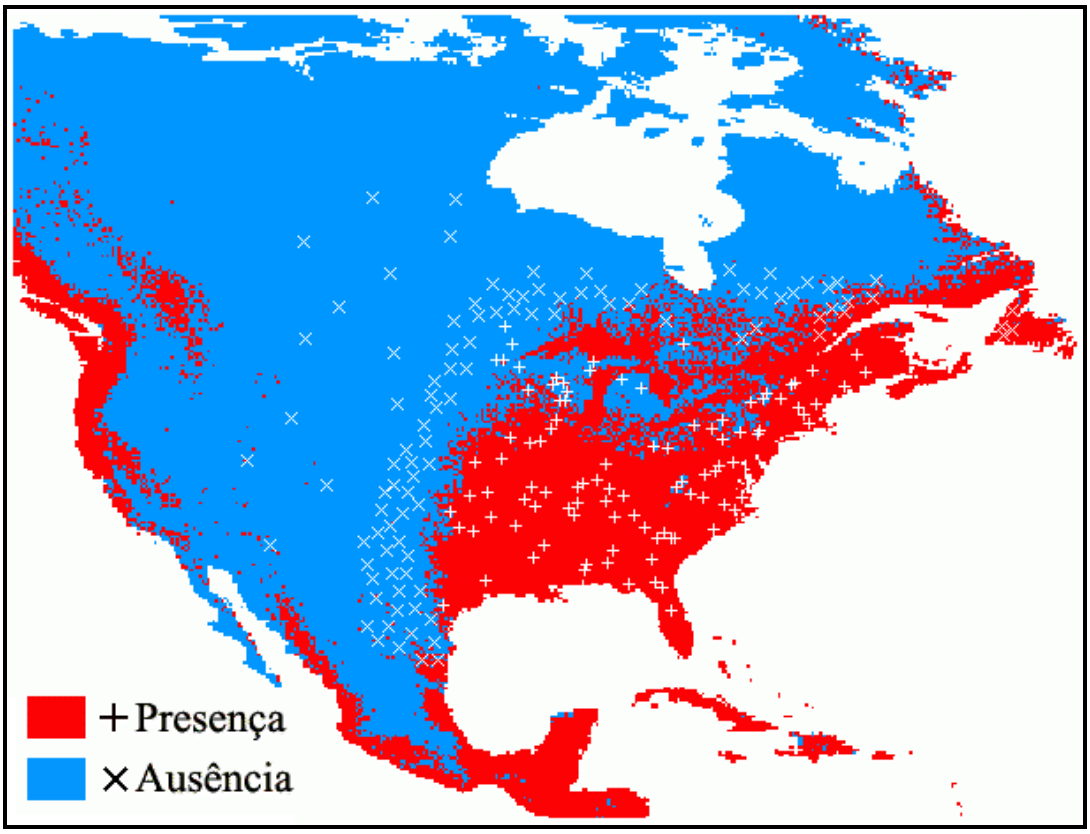

Figura 4.8 – Distribuição potencial da espécie *Strix varia* (modelo S2)

A análise visual da Figura 4.7 e da Figura 4.8 não permite verificar quais efeitos os relacionamentos espaciais, representados na GPM do modelo S2, exercem sobre a distribuição potencial prevista para a espécie *Strix varia*. Para verificar estes efeitos calculou-se as taxas de células preditas como presença ou ausência nos dois mapas de distribuição potencial.

Percebe-se, através das taxas apresentadas na Tabela 4.3, que o modelo S2 produz um mapa com menor área de presença. Esse resultado deve-se à escolha do algoritmo utilizado na construção da GPM; os relacionamentos nela representados estão baseados no espaço absoluto, ou seja, na distância geográfica entre os pontos amostrais, conferindo maiores pesos aos relacionamentos entre pontos mais próximos. Outras estruturas de conhecimento, baseadas no espaço relativo por exemplo, gerariam resultados distintos.

Tabela 4.3 – Taxas de presença e ausência segundo os modelos S1 e S2

Classes	Modelo S1		Modelo S2	
	Número células	%	Número células	%
Presença	16542	28,5%	15837	27,3%
Ausência	41523	71,5%	42228	72,7%
Total	58065	100%	58065	100%

Através do software openModeller Desktop gerou-se outros dois SDM para a espécie *Strix varia*: GARP *Single Run* (SGSR) e GARP *Best Subset* (SGBS). Objetiva-se agora comparar a qualidade dos modelos gerados pelo sistema SAHGA SDM com dois algoritmos freqüentemente utilizados na modelagem de distribuição de espécies. Escolheu-se os algoritmos GARP *Single Run* e GARP *Best Subset* pois ambos utilizam-se de AG em seus núcleos de otimização e desconsideram efeitos oriundos dos relacionamentos espaciais entre as amostras.

A Tabela 4.4 apresenta as métricas para avaliação dos modelos SGSR e SGBS. As curvas ROC são apresentadas na Figura 4.9.

Tabela 4.4 – Métricas para avaliação dos modelos SGSR e SGBS

Métrica	Modelo SGSR	Modelo SGBS
Acurácia	86%	98%
Erro de Omissão	14%	2%
AUC	0,85	0,91

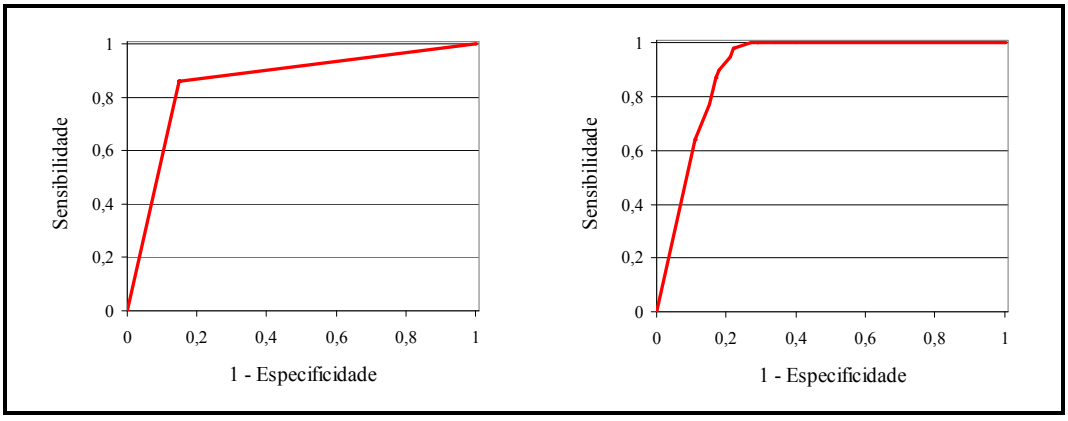

Figura 4.9 – Curvas ROC para os modelos SGSR e SGBS

Ao observar as métricas de avaliação dos modelos SGSR e SGBS apresentadas na Tabela 4.4, nota-se, em todos as métricas, que o modelo SGBS possui desempenho superior ao SGSR.

Essa diferença entre os dois algoritmos já era esperada pois o algoritmo GARP *Best Subset* ajusta vários modelos GARP *Single Run*. Ao final do processo, um número pré-determinado de melhores modelos *Single Run* são selecionados para compor o mapa de distribuição potencial final (Meyer, 2005).

Os mapas de distribuição potencial para a espécie *Strix varia*, prevista pelos modelos SGSR e SGBS, são apresentados na Figura 4.10 e na Figura 4.11. Nota-se que a área predita como presença no algoritmo SGSR é menor do que no algoritmo SGBS. Afirmação ratificada pelas taxas apresentadas na Tabela 4.5.

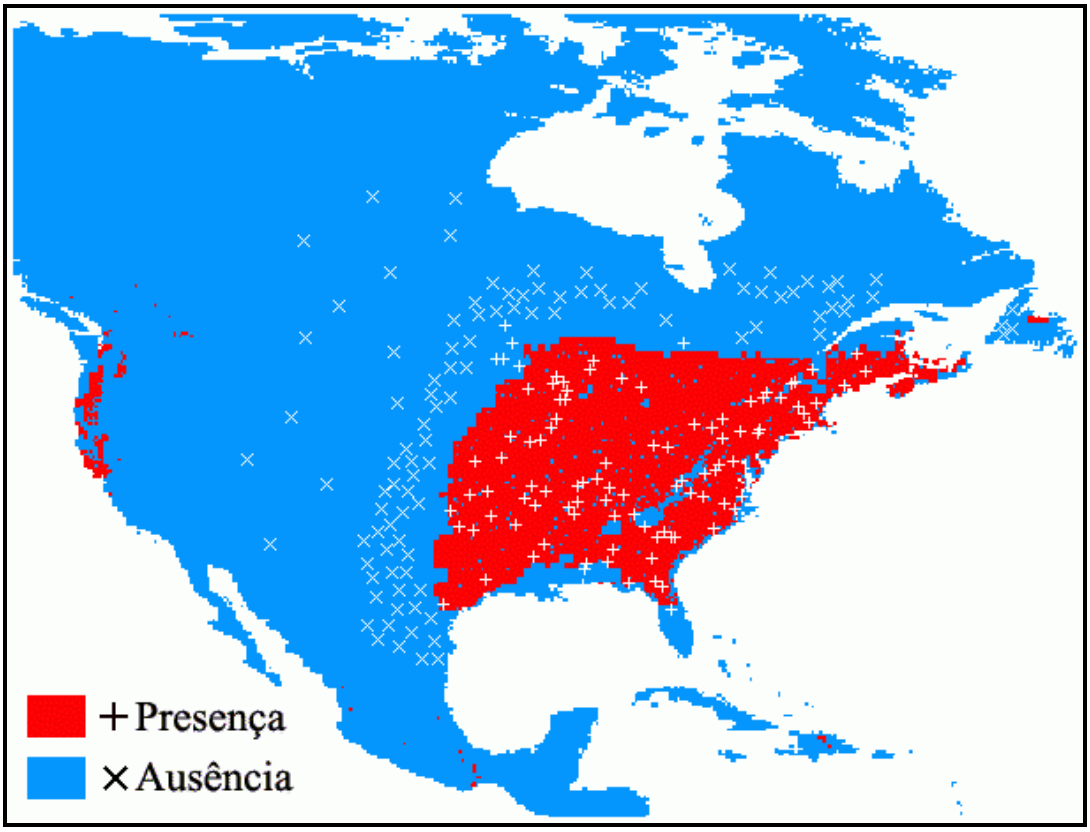

Figura 4.10 – Distribuição potencial da espécie *Strix varia* (modelo SGSR)

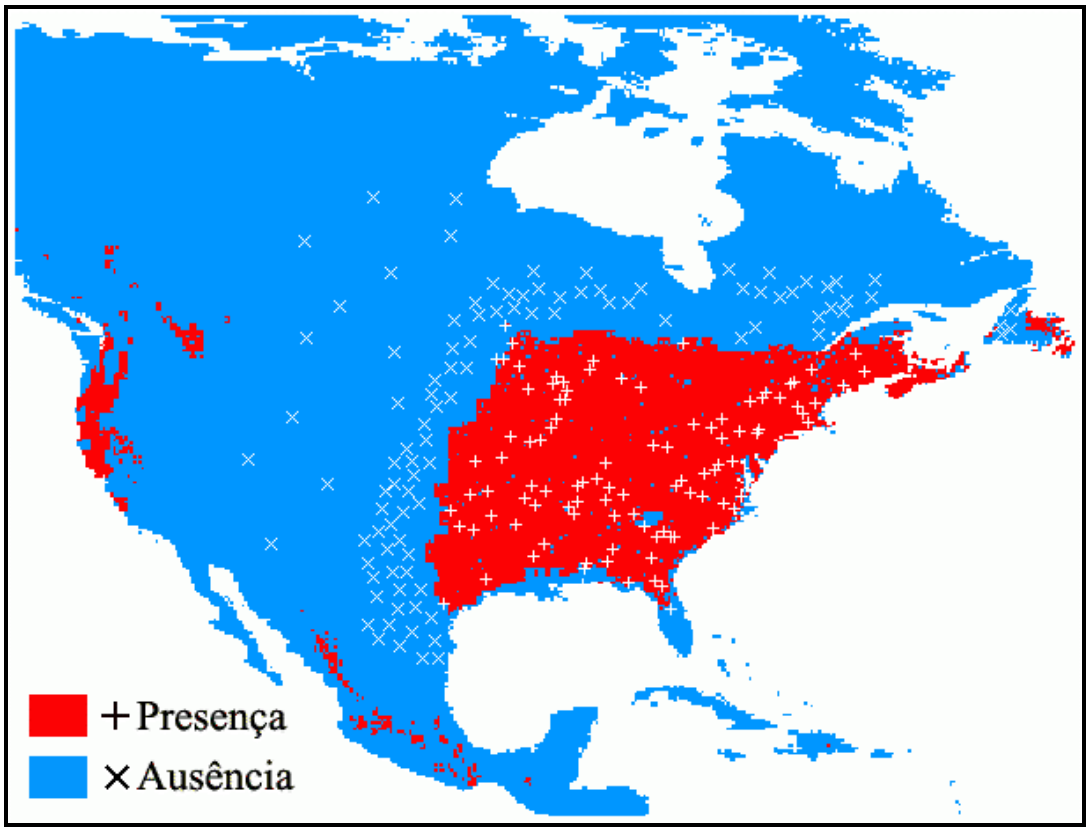

Figura 4.11 – Distribuição potencial da espécie *Strix varia* (modelo SGBS)

Tabela 4.5 – Taxas de presença e ausência segundo os modelos SGSR e SGBS

Classes	Modelo SGSR		Modelo SGBS	
	Número células	%	Número células	%
Presença	8992	15,5%	10789	18,6%
Ausência	49073	84,5%	47276	81,4%
Total	58065	100%	58065	100%

Os dois modelos ajustados através do SAHGA SDM, modelo S1 e modelo S2, apresentaram excelentes resultados. A acurácia de ambos foi superior a 90% com AUC superiores a 0,95. Comparando-os com o modelo SGSR, produzido pelo algoritmo GARP *Single Run*, ambos apresentam métricas de avaliação superiores. Em relação ao modelo SGBS, produzido pelo algoritmo GARP *Best Subset*, a acurácia dos modelos S1 e S2 foram ligeiramente inferiores, 98% contra 91%; entretanto, na comparação do índice AUC, os modelos S1 e S2

apresentaram valores superiores, 0,96 contra 0,91.

Para que o algoritmo GARP *Best Subset* obtenha resultados tão expressivos ele ajusta diversos modelos GARP *Single Run* e, na seqüência, seleciona um conjunto de melhores modelos. Essa característica pode ser vista como uma desvantagem em relação ao SAHGA SDM pois, neste sistema, um único modelo é ajustado.

Os relacionamentos espaciais entre os pontos amostrais, representados na GPM do modelo S2, ocasionou uma ligeira redução na área predita como presença, quando comparados os modelos S1 e S2. Para o modelo S1 a área predita como presença foi de 28,5%, enquanto para o modelo S2 a área predita foi de 27,3%.

Os dois modelos, S1 e S2, acabam por predizer áreas de presença maiores que os modelos SGSR e SGBS, indicando que os primeiros tendem a cometer mais erros do tipo FP (comissão) do que os últimos. Entretanto este indicativo, para ser tomado como verdade, deve ser validado por especialistas sobre a espécie *Strix varia*, pois estes tipos de erros nem sempre são erros verdadeiros.

4.4.2 Espécie *Thalurania furcata boliviana* Boucard, 1894

A base de dados *Thalurania furcata boliviana* é um dos conjuntos de dados fornecidos com o instalador do software openModeller Desktop. A base contém 65 pontos de presença da referida espécie; também são disponibilizados 8 *layers* geográficos correspondentes às variáveis: precipitação acumulada no trimestre mais úmido, precipitação acumulada no trimestre mais quente, precipitação anual, temperatura média anual, temperatura média no trimestre

mais frio, temperatura média no trimestre mais seco, temperatura média no trimestre mais quente e temperatura média no trimestre mais úmido. A Figura 4.12 mostra um exemplar da espécie *Thalurania furcata boliviana*, conhecida popularmente beija-flor-tesoura-verde.

Figura 4.12 – Um exemplar da espécie *Thalurania furcata boliviana*
Fonte: Tobias *et al.* (2008)

Com os 65 pontos de presença utilizou-se o software openModeller Desktop para gerar o SDM usando o algoritmo BIOCLIM. A partir do SDM gerado selecionou-se ao acaso 50 pontos de pseudo-ausência, chamados a partir de agora simplesmente de ausência. A Figura 4.13 apresenta o mapa da distribuição prevista com os pontos de presença e ausência.

Na seqüência, dividiu-se aleatoriamente o conjunto com 115 pontos de presença/ausência em dois subconjuntos: treino e teste. O conjunto de treino possui 70 pontos, sendo 40 de presença e 30 de ausência; o conjunto de teste

possui 45 pontos, sendo 25 de presença e 20 de ausência.

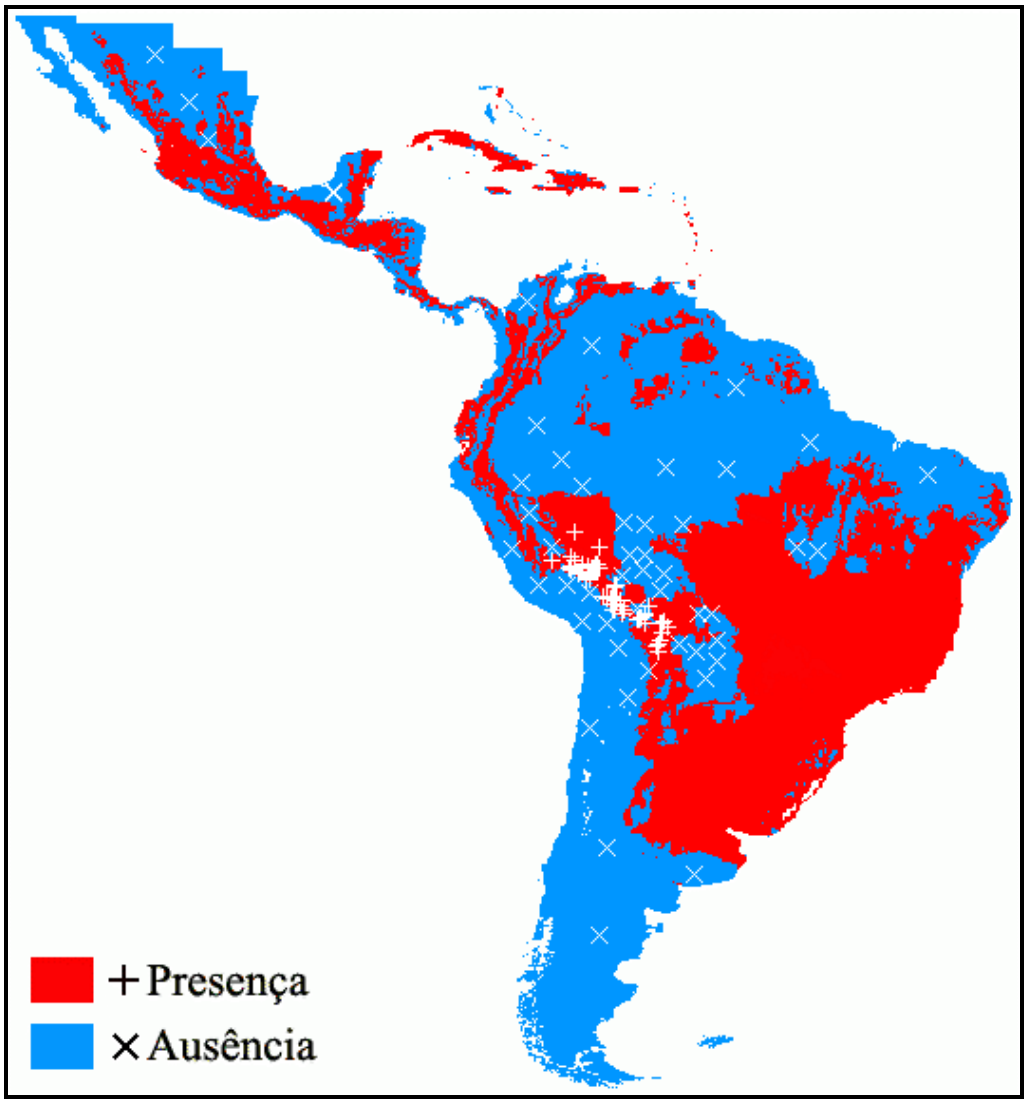

Figura 4.13 – Distribuição potencial da espécie *Thalurania furcata boliviana* (BIOCLIM)

Conduziu-se o experimento com a espécie *Thalurania furcata boliviana* de forma análoga ao experimento com a espécie *Strix varia*. Ajustou-se dois modelos com o sistema SAHGA SDM: modelo T1 e modelo T2. Através do software openModeller Desktop ajustou-se outros dois modelos: GARP *Single Run* (TGSR) e GARP *Best Subset* (TGBS). Os resultados são apresentados na seqüência.

As métricas para avaliação dos modelos T1 e T2 são apresentadas na Tabela 4.6. As curvas ROC são apresentadas na Figura 4.14.

Tabela 4.6 – Métricas para avaliação dos modelos T1 e T2

Métrica	Modelo T1	Modelo T2
Acurácia	82,2%	80%
Erro de Omissão	16%	24%
CCM	0,640	0,606
AUC	0,886	0,892

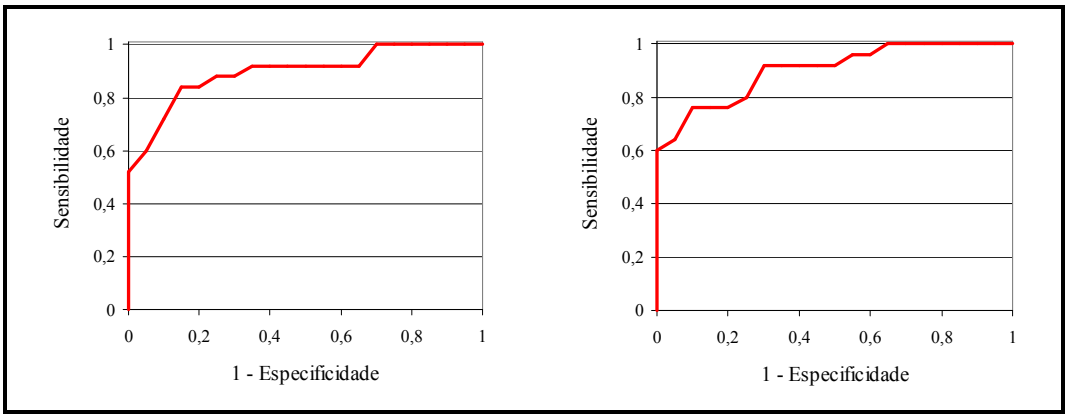

Figura 4.14 – Curvas ROC para os modelos T1 e T2

Ao observar as métricas de avaliação dos modelos T1 e T2, apresentadas na Tabela 4.6, percebe-se que eles têm qualidades similares. Ambos os modelos possuem boa acurácia, mas o modelo T2 apresenta taxa de erros de omissão superior ao modelo T1; conseqüentemente, o CCM do modelo T2 é inferior ao do modelo T1. Os valores de AUC são praticamente iguais e ambos os modelos podem ser considerados bons classificadores, pois apresentam AUC superiores a 0,88.

Os mapas de distribuição potencial para a espécie *Thalurania furcata boliviana*, prevista pelos modelos T1 e T2, são apresentados na Figura 4.15 e na Figura 4.16, respectivamente.

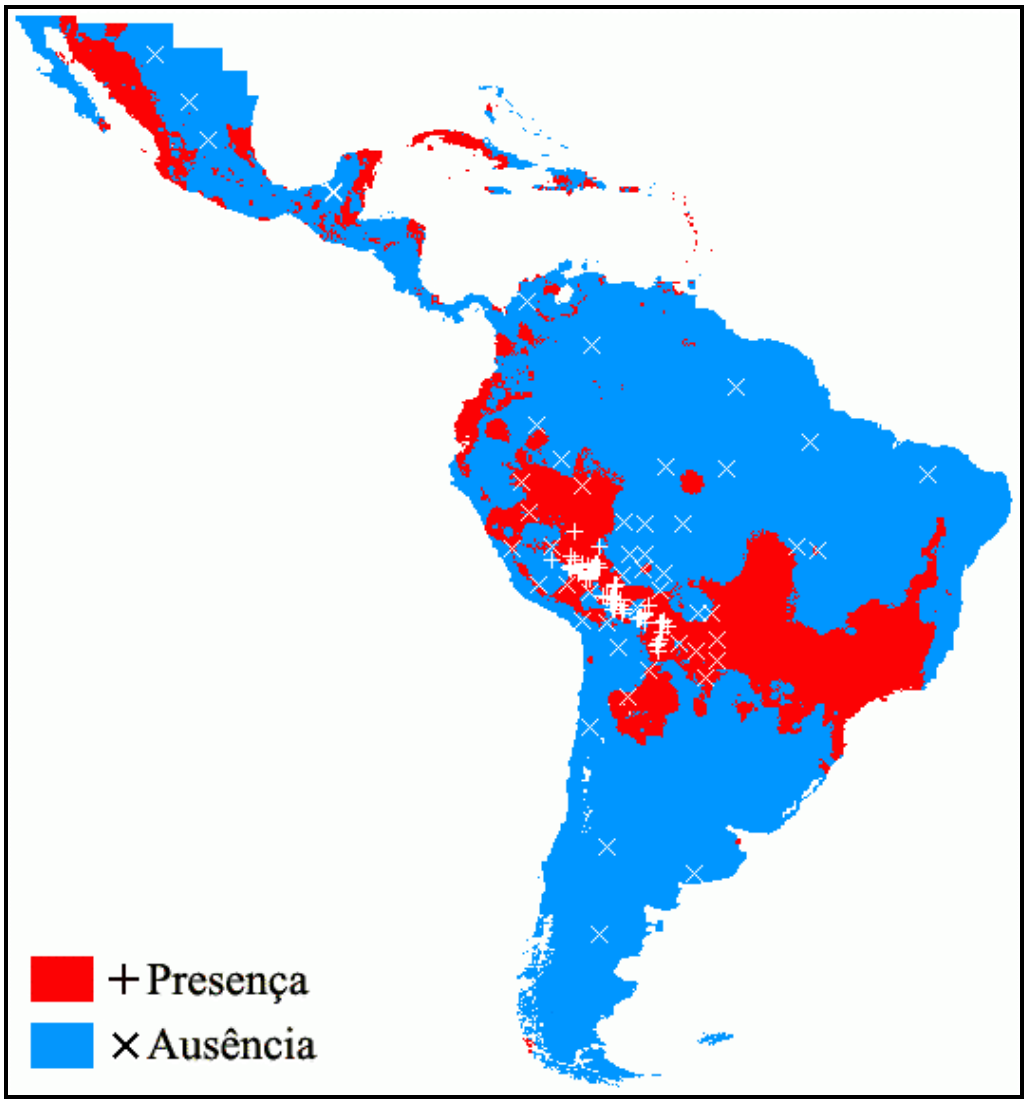

Figura 4.15 – Distribuição potencial da espécie *Thalurania furcata boliviana* (modelo T1)

Na análise visual da Figura 4.15 e da Figura 4.16 percebe-se nitidamente que os relacionamentos espaciais, considerados no modelo T2, afetam a previsão da distribuição potencial da espécie *Thalurania furcata boliviana*. As regiões onde o modelo T2 prevê presença são menores. O modelo T2 prevê menor área de presença que o modelo T1 tanto na região da América Central, quanto na região da América do Sul. As taxas de células preditas como presença ou ausência nos dois mapas de distribuição potencial são apresentadas na Tabela 4.7.

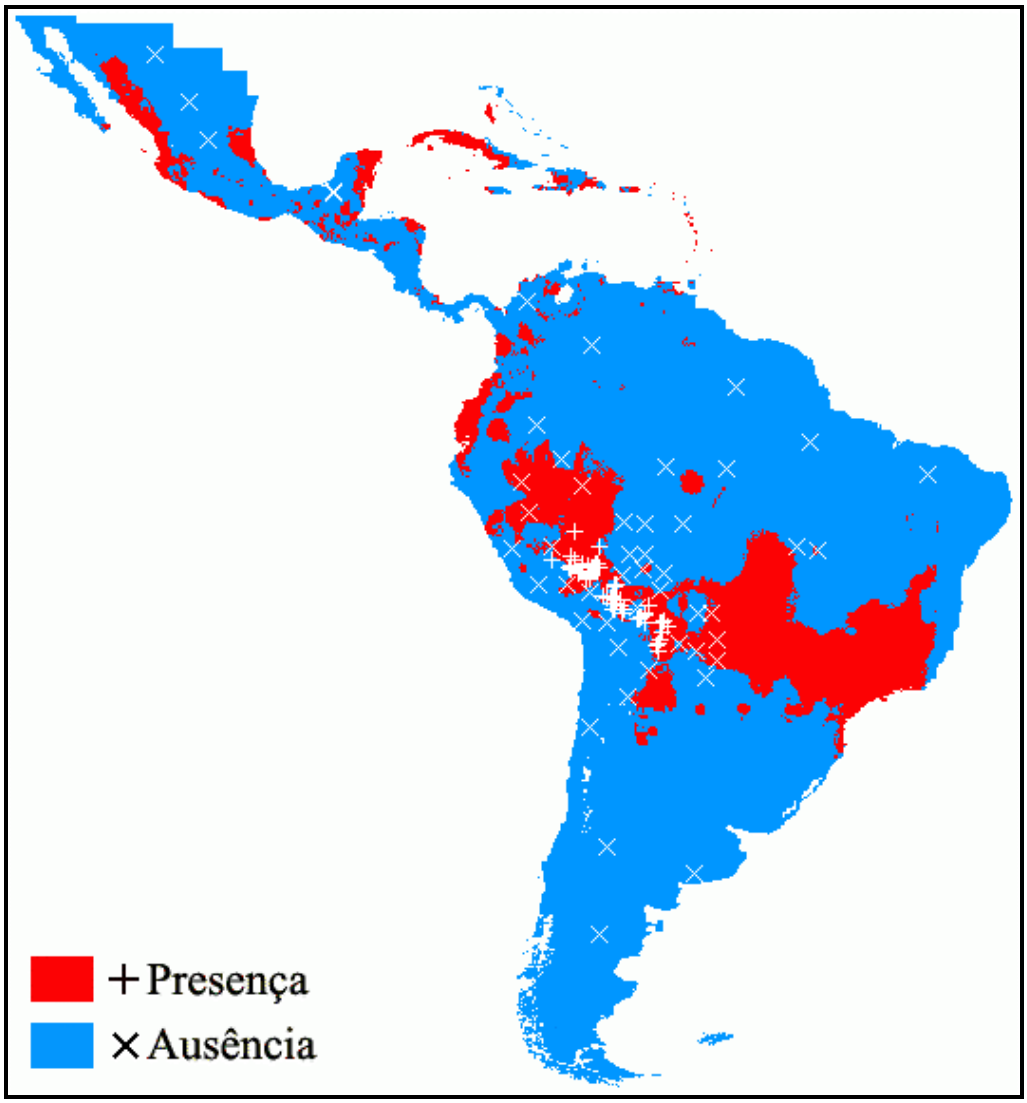

Figura 4.16 – Distribuição potencial da espécie *Thalurania furcata boliviana* (modelo T2)

Percebe-se, através das taxas apresentadas na Tabela 4.7, que o modelo T2 produz um mapa com área de presença inferior ao modelo T1, repetindo o ocorrido com os modelos S1 e S2, ajustados para a espécie *Strix varia*.

Tabela 4.7 – Taxas de presença e ausência segundo os modelos T1 e T2

Classes	Modelo T1		Modelo T2	
	Número células	%	Número células	%
Presença	14816	22,9%	12609	19,5%
Ausência	49865	77,1%	52072	80,5%
Total	64681	100%	64681	100%

A Tabela 4.8 apresenta as métricas para avaliação dos modelos TGSR e TGBS. As curvas ROC são apresentadas na Figura 4.17.

Tabela 4.8 – Métricas para avaliação dos modelos TGSR e TGBS

Métrica	Modelo TGSR	Modelo TGBS
Acurácia	90,8%	89,2%
Erro de Omissão	9,2%	10,8%
AUC	0,85	0,88

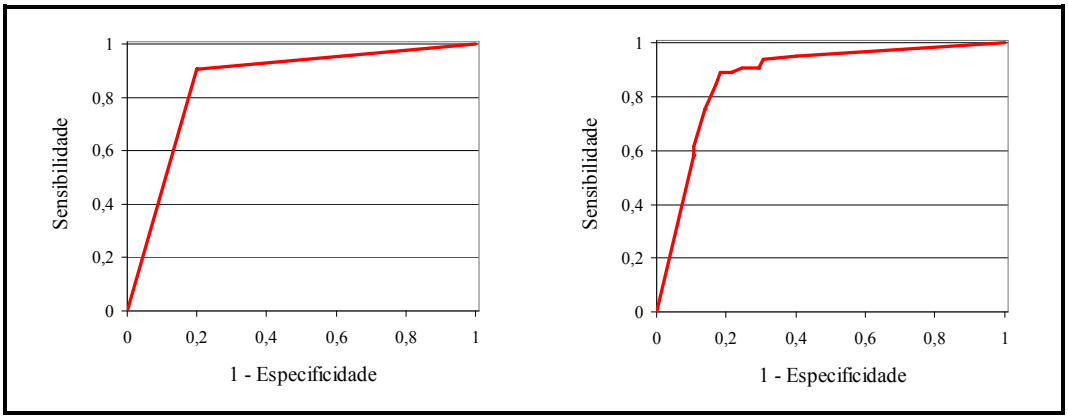

Figura 4.17 – Curvas ROC para os modelos TGSR e TGBS

Ao observar as métricas de avaliação dos modelos TGSR e TGBS, apresentadas na Tabela 4.8, nota-se que o modelo TGSR possui desempenho superior ao modelo TGBS, exceto pelo valor do AUC. Esperava-se o oposto; o modelo TGBS deveria apresentar desempenho nitidamente superior ao TGSR, pois o processo *Best Subset* considera apenas o conjunto com os melhores modelos *Single Run*. O fato do modelo TGSR possuir desempenho muito superior à média dos modelos utilizados na construção do TGBS pode justificar esta contradição.

Os mapas de distribuição potencial para a espécie *Thalurania furcata boliviana* previstas pelos modelos TGSR e TGBS são apresentados na Figura 4.18 e na Figura 4.19. Nota-se que a área predita como presença no algoritmo TGSR é

maior do que no algoritmo TGBS. Afirmação ratificada pelos números apresentados na Tabela 4.9.

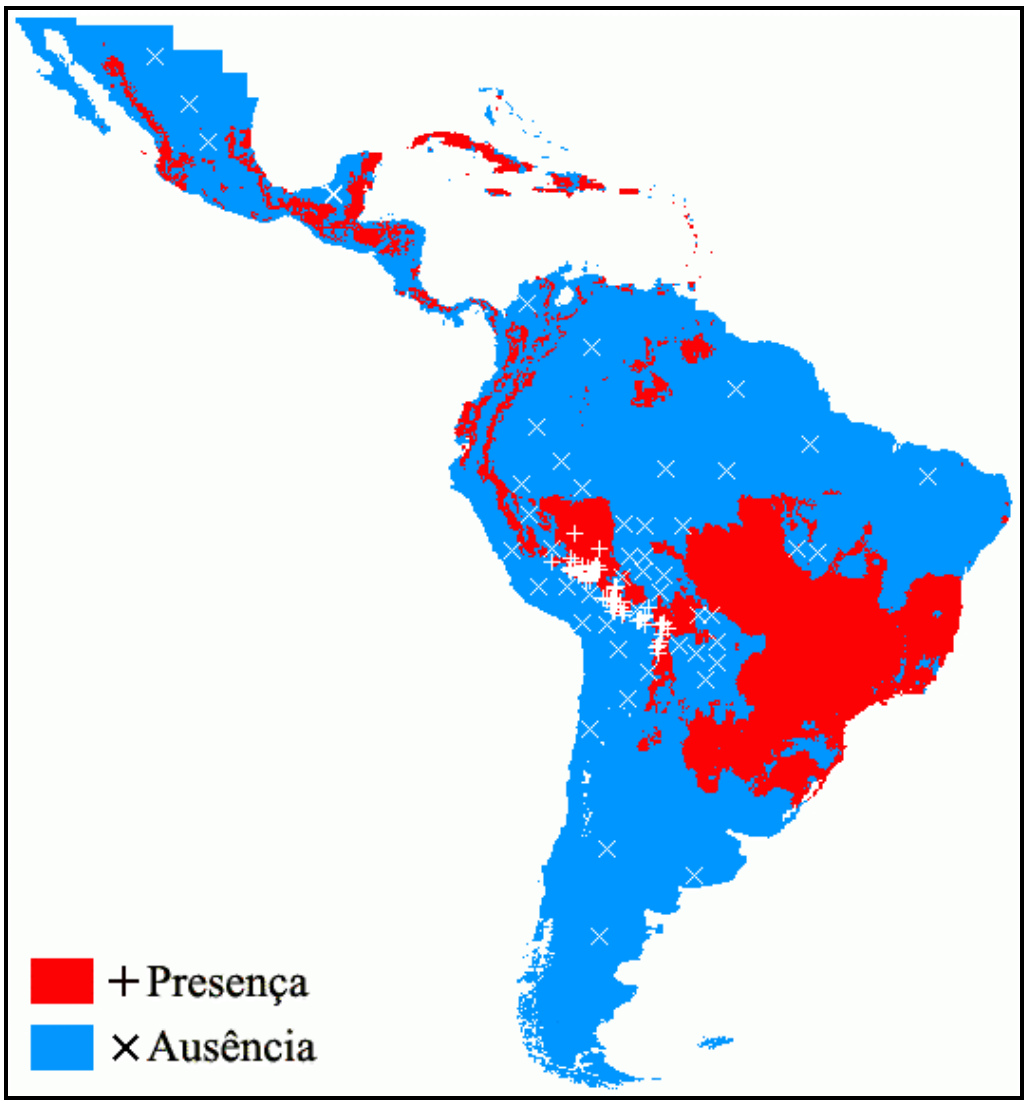

Figura 4.18 – Distribuição potencial da espécie *Thalurania furcata boliviana* (modelo TGSR)

Tabela 4.9 – Índices de presença e ausência segundo os modelos TGSR e TGBS

Classes	Modelo TGSR		Modelo TGBS	
	Número células	%	Número células	%
Presença	16878	26,1%	13125	20,3%
Ausência	47803	73,9%	51556	79,7%
Total	64681	100%	64681	100%

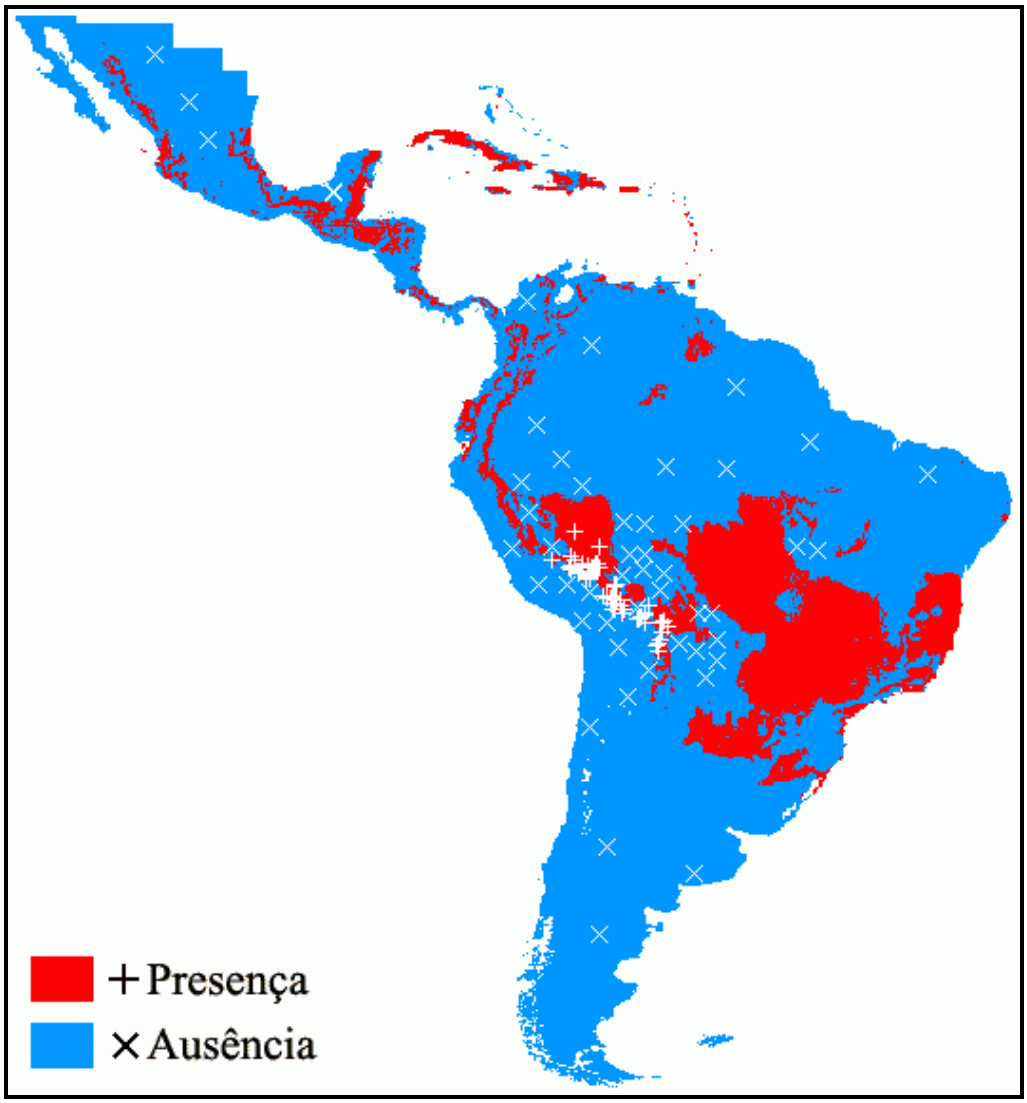

Figura 4.19 – Distribuição potencial da espécie *Thalurania furcata boliviana* (modelo TGBS)

Os dois modelos ajustados através do SAHGA SDM, modelo T1 e modelo T2, apresentaram bons resultados. A acurácia de ambos foi igual ou superior a 80% com AUC superiores a 0,88.

A acurácia dos modelos TGSR (90,8%) e TGBS (89,2%) são superiores à acurácia dos modelos T1 (82,2%) e T2 (80%). As taxas de erro de omissão dos modelos TGSR (9,2%) e TGBS (10,8%) são inferiores às taxas dos modelos T1 (16%) e T2 (24%). Entretanto, na comparação do índice AUC, os modelos T1 e T2 apresentaram valores ligeiramente superiores.

Para a espécie *Thalurania furcata boliviana* ficou evidente que os relacionamentos espaciais, entre os pontos amostrais, exercem influência na previsão da distribuição potencial da espécie. O mapa da distribuição potencial para o modelo T2 (Figura 4.16) é visivelmente diferente do mapa de distribuição potencial para o modelo T1 (Figura 4.15).

Desta vez os modelos ajustados com o SAHGA SDM predizem áreas de presença próximas aos modelos TGSR e TGBS; o modelo T1 prediz 22,9%, o modelo T2 prediz 19,5%, o modelo TGSR prediz 26,1% e o modelo TGBS prediz 20,3% de área de presença. Este é um bom resultado mostrando que, para a espécie *Thalurania furcata boliviana*, os modelos T1 e T2 apresentam taxas de erros do tipo FP (comissão) próximas às taxas produzidas pelos modelos TBSR e TGBS.

4.5 Conclusões

O sistema SAHGA SDM, para os modelos de distribuição potencial ajustados para a espécie *Strix varia* apresentaram qualidade superior aos modelos ajustados para a espécie *Thalurania furcata boliviana*, conforme observado na Tabela 4.10. Utilizou-se, nos modelos S1 e T1, uma GPM ignorando os relacionamentos espaciais entre as amostras. Nos modelos S2 e T2 empregou-se uma GPM com relacionamentos espaciais que variam em função da distância entre os pontos amostrais.

As diferenças observadas entre as métricas dos modelos S1 e S2 e as métricas dos modelos T1 e T2 corroboram a conclusão obtida com a análise dos resultados fornecidos pelo SAHGA MB; os relacionamentos espaciais, quando considerados, também interferem nos modelos ajustados pelo SAHGA SDM.

Tabela 4.10 – Métricas para avaliação dos modelos ajustados pelo SAHGA SDM

Métrica	*Strix varia*		*Thalurania furcata boliviana*	
	Modelo S1	**Modelo S2**	**Modelo T1**	**Modelo T2**
Acurácia	91%	91%	82,2%	80%
Erro de Omissão	6%	6%	16%	24%
CCM	0,821	0,821	0,640	0,606
AUC	0,957	0,964	0,886	0,892
Área de presença	28,5%	27,3%	22,9%	19,5%
Área de ausência	71,5%	72,7%	77,1%	80,5%

Para avaliar a qualidade dos modelos ajustados com o SAHGA SDM, ajustou-se modelos de distribuição potencial para as espécies *Strix varia* e *Thalurania furcata boliviana* utilizando os algoritmos GARP *Single Run* e GARP *Best Subset,* ambos implementados no software openModeller Desktop v1.0.6. As métricas para avaliação destes modelos são apresentadas na Tabela 4.11. Os modelos SGSR e TGSR foram ajustados pelo algoritmo GARP *Single Run* e os modelos SGBS e TGBS foram ajustados pelo algoritmo GARP *Best Subset*.

Tabela 4.11 – Métricas para avaliação dos modelos ajustados pelo openModeller Desktop

Métrica	*Strix varia*		*Thalurania furcata boliviana*	
	Modelo SGSR	**Modelo SGBS**	**Modelo TGSR**	**Modelo TGBS**
Acurácia	86%	98%	90,8%	89,2%
Erro de Omissão	14%	2%	9,2%	10,8%
AUC	0,85	0,91	0,85	0,88
Área de presença	15,5%	18,6%	26,1%	20,3%
Área de ausência	84,5%	81,4%	73,9%	79,7%

Comparando a Tabela 4.10 com a Tabela 4.11 percebe-se que os modelos ajustados pelo SAHGA SDM, para a espécie *Strix varia*, apresentam métricas tão boas ou melhores que as métricas apresentadas pelos modelos ajustados pelos algoritmos GARP *Single Run* e GARP *Best Subset*. Para a espécie *Thalurania furcata boliviana* o mesmo não se observou; os modelos ajustados pelo SAHGA SDM apresentaram menor acurácia e maior taxa de erros de omissão, mas ainda apresentam maiores valores de AUC.

A qualidade inferior do modelo ajustado para a espécie *Thalurania furcata boliviana*, indica que o modelo aditivo linear, empregado no SAHGA SDM, foi inadequado para prever a ocorrência desta espécie em função das variáveis ambientais escolhidas.

Os modelos ajustados pelo SAHGA SDM para a espécie *Strix varia* prevêem áreas de presença maiores que os modelos ajustados pelos algoritmos GARP *Single Run* e GARP *Best Subset*, indicando tendência a cometer mais erros do tipo FP (comissão). Entretanto este indicativo, para ser tomado como verdade, deve ser validado por especialista sobre a espécie, pois erros de comissão nem sempre são erros verdadeiros; a espécie ocorre naquele local e não foi observada ou fatores topológicos ou biológicos impedem a fixação da espécie.

5 CONCLUSÕES

O uso de mecanismos automáticos e semi-automáticos visa reduzir o esforço humano empregado na busca por informações contidas em conjunto de dados geoespaciais. A ciência da computação, mais especificamente a inteligência computacional, tem contribuído com soluções para alcançar este objetivo; sistemas especialistas, redes neurais, lógica nebulosa e computação evolutiva são técnicas de inteligência computacional aplicadas na análise de dados geoespaciais.

Dentre os algoritmos utilizados em computação evolutiva destacam-se os algoritmos genéticos, utilizados em sistemas de análise de dados geoespaciais como os *Model Breeders* e o GARP. Porém, estes sistemas negligenciam a dependência espacial, um conceito fundamental para analisar e compreender os fenômenos geográficos. Esta negligência ocorre não apenas nestes software; ela deve-se à ausência de um algoritmo genético que incorpore, em seus mecanismos evolutivos, os relacionamentos espaciais.

O algoritmo SAHGA – *Spatially Aware Hybrid Genetic Algorithm* – algoritmo heurístico híbrido com representação explícita de relacionamentos espaciais foi desenvolvido para acomodar a representação de associações ou relacionamentos espaciais. Como estratégia, o algoritmo SAHGA utiliza o conceito de Matriz de Proximidade Generalizada, GPM.

Para demonstrar o SAHGA dois sistemas foram desenvolvidos: o SAHGA MB e o SAHGA SDM. O primeiro, um sistema semi-automático para analisar dados geoespaciais encontrando modelos que relacionam variáveis dependentes e independentes. O segundo, um sistema utilizado na geração de modelos de distribuição potencial de espécies.

Visando experimentar os conceitos introduzidos pelo SAHGA realizou-se estudos de caso com os sistemas implementados. O SAHGA MB foi empregado na análise de um conjunto de dados sócio-econômicos e o SAHGA SDM na modelagem da distribuição potencial das espécies *Strix varia* e *Thalurania furcata boliviana*.

O sistema SAHGA MB mostrou-se capaz de encontrar modelos que relacionam variáveis dependentes e independentes. No estudo de caso realizado, ele ajustou modelos que permitiram estimar adequadamente o valor padronizado da variável "número de filhos nascidos vivos", em função do conjunto de variáveis independentes utilizado.

Os modelos de distribuição das espécies *Strix varia* e *Thalurania furcata boliviana*, ajustados pelo sistema SAHGA SDM, apresentaram boas métricas de avaliação. Para a espécie *Strix varia*, por exemplo, as métricas calculadas foram tão boas ou melhores que as métricas apresentadas pelos modelos ajustados com os algoritmos GARP *Single Run* e GARP *Best Subset*.

Para averiguar se os relacionamentos espaciais incorporados aos mecanismos evolutivos influenciam os modelos ajustados, os sistemas foram testados sem e com relacionamentos espaciais. Tanto no SAHGA MB quanto no SAHGA SDM os resultados obtidos, sem e com relacionamentos espaciais, foram distintos. Esta diferença mostra que os relacionamentos espaciais influenciam os modelos ajustados.

No teste realizado com o SAHGA MB empregou-se uma GPM que simulava associações no espaço relativo; as associações entre estados, definidas arbitrariamente, não dependiam da distância; também não eram válidas para todos os estados da União. Nos testes efetuados com o SAHGA SDM, empregou-se uma GPM que representava associações no espaço absoluto e

que variavam em função da distância. Estes dois exemplos mostram como um especialista pode representar, no algoritmo SAHGA, o conhecimento prévio sobre os elementos que compõem o cenário e que, no seu julgamento, afetam o fenômeno espacial em estudo.

Uma característica do algoritmo SAHGA que se deve ressaltar é sua adaptabilidade. Neste trabalho empregou-se o algoritmo SAHGA no ajuste de modelos lineares multivariados, tanto no SAHGA MB quanto no SAHGA SDM. Entretanto, pode-se utilizá-lo no ajuste de modelos mais complexos que envolvam parâmetros não lineares, por exemplo. Para tanto basta incluir na estrutura cromossômica do SAHGA um número maior de coeficientes e escrever uma função de aptidão considerando termos não lineares.

5.1 Trabalhos Futuros

Ao concluir este trabalho percebeu-se que o algoritmo SAHGA pode ser melhorado e expandido para novas aplicações, caracterizando novos desafios tecnológicos e científicos.

Como desafios tecnológicos tem-se:

a) Incluir rotinas para geração automática da GPM, utilizando relacionamentos espaciais pré-definidos como toca, perto de, intercepta etc.;

b) Disponibilizar estruturas cromossômicas e função de aptidão pré-definidas, possibilitando ao usuário apenas escolher o tipo de modelo a ser ajustado através do algoritmo; e

c) Incluir rotinas de pré-análise de dados para, por exemplo, eliminar

variáveis com alto índice de correlação.

Como desafios científicos tem-se:

a) Pesquisar e implementar estratégias para geração automática do pontos de pseudo-ausência; e

b) Expandir as funcionalidades do SAHGA para trabalhar com GPM dinâmicas. O conhecimento representado numa GPM pode alterar-se no tempo em função de sua própria influência, alterações ambientais e legais ou mudanças em políticas públicas, por exemplo. Desta forma o algoritmo SAHGA poderia ser empregado na análise de dados espaço-temporais.

REFERÊNCIAS BIBLIOGRÁFICAS

AGUIAR, A. P. D. *et al.* Modeling spatial relations by generalized proximity matrices. In: Brazilian Symposium on Geoinformatics, 5, 2003. Campos do Jordão - SP. **Anais eletrônicos...** São José dos Campos: INPE, Nov. 2003. Disponível em: <http://www.geoinfo.info/geoinfo2003/papers/geoinfo2003-11.pdf>. Acesso em: 04/07/2006.

ARAUJO, H. A. **Algoritmo simulated annealing:** uma nova abordagem. 2001. 95 p. Dissertação (Mestrado em Engenheria da Produção) - Universidade Federal de Santa Catarina, Florianópolis.

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

AUKEMA, B. H. *et al.* Landscape level analysis of mountain pine beetle in British Columbia, Canada: spatiotemporal development and spatial synchrony within the present outbreak. **Ecography**, v. 29, n. 3, p. 427-441, Abr. 2006.

AUSTIN, M. P.; MEYERS, J. A. Current approaches to modelling the environmental niche of eucalypts: implication for management of forest biodiversity. **Forest Ecology and Management**, v. 85, n. 1-3, p. 95-106, Set. 1996.

BAÇÃO, F. L. **Geospatial data mining**. 2006. Disponível em: <http://edugi.uji.es/Bacao/Geospatial%20Data%20Mining.pdf>. Acesso em: 01/08/2006.

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

BITTENCOURT, G. **Algoritmos genéticos**. 1998. Disponível em: <http://www.das.ufsc.br/gia/softcomp/node17.html>. Acesso em: 15/07/2007.

BRAGA, A. C. S. **Curvas ROC:** aspectos funcionais e aplicações. 2000. 243 p. Tese (Doutorado em Engenharia de Produção e Sistemas) - Universidade do Minho, Braga.

CÂMARA, G.; MONTEIRO, A. M. V. Geocomputation techniques for spatial analysis: are they relevant to health data? **Cadernos Saúde Pública**, v. 17, n. 5, p. 1059-1081, Set./Out. 2001.

CÂMARA, G. *et al.* Análise espacial e geoprocessamento. In: DRUCK, S. *et al.* (Eds). **Análise espacial de dados geográficos**. Brasília: EMBRAPA, 2004. p. 21-52.

CASTRO, J. P. **Um algoritmo evolucionário para geração de planos de rotas**. 1999. 91 p. Dissertação (Mestrado em Engenharia da Produção) - Universidade Federal de Santa Catarina, Florianópolis.

CENTER FOR ENVIRONMENTAL MODELING FOR POLICY DEVELOPMENT (CEMPD). **8-km grid over Nashville, Tennessee**. 2008. Disponível em: <http://www.ie.unc.edu/cempd/projects/mims/spatial/nash08km.jpg>. Acesso em: 25/08/2008.

CENTRO DE REFERÊNCIA EM INFORMAÇÃO AMBIENTAL (CRIA); ESCOLA POLITÉCNICA DA USP (POLI-USP); INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS (INPE). **openModeller**. 2008. Disponível em: <http://openmodeller.sourceforge.net/>. Acesso em: 12/07/2008.

COUCLELIS, H. From cellular automata to urban models: new principles for model development and implementation. **Environment and planning B**, v. 24, n. 2, p. 165-174, Mar. 1997.

_____. Geocomputation in context. In: LONGLEY, P. A. *et al.* (Eds). **Geocomputation:** a primer. New York: John Wiley & Sons, 1998. p. 17-29.

DAVIS, L. Adapting operator probabilities in genetic algorithms. In: International conference on genetic algorithms, 3, 1989. Fairfax. **Proceedings...** San Francisco: Morgan Kaufmann, Jun. 1989. p. 61-69.

_____ (Ed.) **Handbook of genetic algorithms**. New York: Van Nostrand Reinhold, 385 p., 1991.

DE BONA, A. A.; ALGERI, T. Algoritmo de otimização combinatorial: uma proposta híbrida utilizando algoritmos simulated annealing e genético em ambiente multiprocessado. In: Encontro Paranaense de Computação, 1, 2005. Cascavel - PR. **Anais...** Cascavel: Curso de Informática/UNIOESTE, set. 2005.

DE JONG, K. A. **An analysis of the behavior of a class of genetic adaptive systems**. 1975. 256 p. Tese (Ph. D. in Computer and Communication Sciences) - University of Michigan, Ann Arbor.

DE JONG, K. A.; SPEARS, W. M. An analysis of the interacting roles of population size and crossover in genetic algorithms. In: SCHWEFEL, H. P. e MÄNNER, R. (Eds). **Proceedings from the 1st workshop on parallel problem solving from nature**. Berlin: Springer, v.496, 1991. p. 38-47. (Lecture notes in computer science).

DELEO, J. M. Receiver operating characteristic laboratory (ROCLAB): software for developing decision strategies that account for uncertainty. In: International symposium on uncertainty modelling and analysis, 2, 1993. Maryland. **Proceedings...** Los Alamitos: IEEE Computer Society Press, 1993. p. 318–325.

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, Abr. 2004.

ESHELMAN, L. J.; SCHAFFER, J. D. Real-coded genetic algorithms and interval schemata. In: WHITLEY, L. D. (Ed.) **Foundations of genetic algorithms - 2**. San Francisco: Morgan Kauffman, 1993. p. 187-202.

FAN, B. A hybrid spatial data clustering method for site selection: The data driven approach of GIS mining. **Expert Systems With Applications**, v. 36, n. 2P2, p. 3923-3936, Mar. 2009.

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

GOLDBERG, D. E. **Genetic algorithms in search, optimization & machine learning**. Reading: Addison-Wesley, 1989. 432 p.

GOUD, R. N. K. **GA optimization technique's in interpolation for dynamic GIS**. 2003. Disponível em: <http://www.gisdevelopment.net/technology/rs/mi03046.htm>. Acesso em: 01/05/2006.

GREFENSTETTE, J. J. Optimization of control parameters for genetic algorithms. **IEEE transactions on systems, man and cybernetics**, v. 16, n. 1, p. 122-128, Jan./Fev. 1986.

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

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

GWEE, B. H.; CHANG, J. S. A hybrid genetic hill-climbing algorithm for four-coloring map problems. In: ABRAHAM, A. *et al.* (Eds). **Design and application of hybrid intelligent systems**. Amsterdam: IOS Press, v.104, 2003. p. 252-261. (Frontiers in Artificial Intelligence and Applications).

HARTSHORNE, R. **Propósitos e natureza da geografia**. 2. ed. São Paulo: HUCITEC/EDUSP, 1978.

HERRERA, F.; LOZANO, M.; VERDEGAY, J. L. Tackling Real-Coded Genetic Algorithms: Operators and Tools for Behavioural Analysis. **Artificial intelligence review**, v. 12, n. 4, p. 265-319, Ago. 1998.

HIRZEL, A.; GUISAN, A. Which is the optimal sampling strategy for habitat suitability modelling. **Ecological Modelling**, v. 157, n. 2-3, p. 331-341, Nov. 2002.

HÖGLUND, J.; SHOREY, L. Genetic divergence in the superspecies Manacus. **Biological journal of the Linnean society**, v. 81, n. 3, p. 439-447, Nov. 2004.

HOLLAND, J. H. **Adaptation in natural and artificial systems**. Cambridge: MIT Press, 1992. 228 p.

HUTCHINSON, G. E. Concluding remarks - Cold Spring Harbor symposium. **Quantitative Biology**, v. 22, p. 415-427, 1957.

ICHIHARA, J. A. **Um método de solução heurístico para a programação de edifícios dotados de múltiplos pavimentos-tipo**. 1998. 173 p. Tese (Doutorado em Engenharia da Produção) - Universidade Federal de Santa Catarina, Florianópolis.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA (IBGE). **Metodologia do censo demográfico 2000**. Rio de Janeiro: IBGE, 2003. 568 p. Disponível em: <http://www.ibge.gov.br/home/estatistica/populacao/censo2000/metodologia/metodologiacenso2000.pdf>. Acesso em: 09/11/2007.

_____. **Censo demográfico e contagem de população**. 2007. Disponível em: <http://www.sidra.ibge.gov.br/cd/default.asp>. Acesso em: 05/07/2008.

IWASHITA, F. **Sensibilidade de modelos de distribuição de espécies a erros de posicionamento de dados de coleta**. 2007. 75 p. Dissertação (Mestrado em Sensoriamento Remoto) - Instituto Nacional de Pesquisas Espaciais, São José dos Campos.

JAYAWARDENA, A. W.; MUTTIL, N.; LEE, J. H. W. Comparative analysis of data-driven and GIS-Based Conceptual Rainfall-Runoff model. **Journal of Hydrologic Engineering**, v. 11, n. 1, p. 1-11, Jan./Fev. 2006.

KANSAS UNIVERSITY. **DesktopGarp**. 2007. Disponível em: <http://www.nhm.ku.edu/desktopgarp/>. Acesso em: 03/05/2008.

KIRKPATRICK, S.; GELATT, C. D.; VECCHI, M. P. Optimization by simulated annealing. **Science**, v. 220, n. 4598, p. 671-680, Maio 1983.

LACERDA, E. G. M.; CARVALHO, A. C. P. L. F. Introdução aos algoritmos genéticos. In: GALVÃO, C. O. e VALENÇA, M. J. S. (Eds). **Sistemas Inteligentes:** aplicações a recursos hídricos e sistemas ambientais. Porto Alegre: Editora da UFRGS/ABRH, 1999. p. 99-148.

LEE, Z. J. *et al.* Genetic algorithm with ant colony optimization (GA-ACO) for multiple sequence alignment. **Applied Soft Computing**, v. 8, n. 1, p. 55-78, Jan. 2008.

LONGLEY, P. A. *et al.* (Eds). **Geocomputation:** a primer. New York: John Wiley & Sons, 290 p., 1998.

LUCASIUS, C. B.; KATEMAN, G. Application of genetic algorithms in chemometrics. In: International conference on genetic algorithms, 3, 1989. Fairfax. **Proceedings...** San Francisco: Morgan Kaufmann, 1989. p. 170-176.

MAHFOUD, S. W.; GOLDBERG, D. E. Parallel recombinative simulated annealing: A genetic algorithm. **Parallel Computing**, v. 21, n. 1, p. 1-28, Jan. 1995.

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

METROPOLIS, N. *et al.* Equation of state calculations by fast computing machines. **The Journal of Chemical Physics**, v. 21, n. 6, p. 1087, Jun. 1953.

MEYER, E. M. Ecological niche modelling: inter-model variation, best-subset models selection. In: Workshop on Biodiversity Data Modelling, 2005. Cidade do México. **Anais eletrônicos...** Copenhague: GBIF, Abr. 2005. Disponível em: <http://www.gbif.org/prog/ocb/modeling_workshop/bangalore/presentations/ENMIV>. Acesso em: 21/10/2007.

MICHALEWICZ, Z. **Genetic algorithms + data structures = evolution programs**. 3. ed. Berlin: Springer-Verlag, 1996. 387 p.

MIDGLEY, G. F. *et al.* Assessing the vulnerability of species richness to anthropogenic climate change in a biodiversity hotspot. **Global Ecology & Biogeography**, v. 11, n. 6, p. 445-451, Dez. 2002.

MITCHELL, M. **An introduction to genetic algorithms**. Cambridge: The MIT Press, 1998. 221 p.

MOJTAHEDI, A. **Barred owl (*Strix varia*)**. Jan. 2005. Imagem digital, color., 2000 x 2000 pixels. Disponível em: <http://upload.wikimedia.org/wikipedia/commons/3/3f/Strix-varia-005-crop.jpg>.Acesso em: 10/10/2008.

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

O'SULLIVAN, D. Toward micro-scale spatial modeling of gentrification. **Journal of geographical systems**, v. 4, n. 3, p. 251-274, Out. 2002.

OLIVEIRA, J. R. F. **O uso de algoritmos genéticos na decomposição morfológica de operadores invariantes em translação aplicados a imagens digitais**. 2000. 110 p. Tese (Doutorado em Computação Aplicada) - Instituto Nacional de Pesquisas Espaciais, São José dos Campos.

OPENSHAW, S. **A geocomputational research agenda for a new millennium**. 1999. Disponível em: <http://www.geog.leeds.ac.uk/presentations/99-2/index.htm>. Acesso em: 11/06/2007.

OPENSHAW, S.; ABRAHART, R. J. Geocomputation. In: International Conference on Geocomputation, 1, 1996. Leeds. **Proceedings...** Leeds: University of Leeds, 1996. p. 665-666.

_____ (Eds). **GeoComputation**. London: CRC Press, 413 p., 2000.

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

PAYNE, K.; STOCKWELL, D. R. B. **GARP modelling system user's guide and technical reference**. 2001. Disponível em: <http://landshape.org/enm/garp-modelling-system-users-guide-and-technical-reference/>. Acesso em: 03/07/2007.

PEDROSA, B. M. **Ambiente computacional para modelagem dinâmica**. 2003. 71 p. Tese (Doutorado em Computação Aplicada) - Instituto Nacional de Pesquisas Espaciais, São José dos Campos.

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, Jul. 2001.

RANGEL, T. F. L. V. B.; DINIZ FILHO, J. A. F.; BINI, L. M. Towards an integrated computational tool for spatial analysis in macroecology and biogeography. **Global ecology and biogeography**, v. 15, n. 4, p. 321-327, Maio 2006.

RAXWORTHY, C. J. *et al.* Predicting distributions of known and unknown reptile species in Madagascar. **Nature**, v. 426, p. 837-841, Dez. 2003.

REES, P.; TURTON, I. Geocomputation: solving geographical problems with new computing power. **Environment and plannig A**, v. 30, n. 10, p. 1835-1838, Out. 1998.

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

SANTA CATARINA, A.; BACH, S. L. Estudo do efeito dos parâmetros genéticos sobre a solução otimizada e sobre o tempo de convergência em algoritmos genéticos com codificações binária e real. **Acta Scientiarum. Technology**, v. 25, n. 2, p. 147-152, Jul./Dez. 2003.

SANTA CATARINA, A.; OLIVEIRA, J. R. F.; MONTEIRO, A. M. V. Model Breeder: um algoritmo genético para criação de modelos. In: Workshop dos cursos de computação aplicada do INPE, 5, 2005. São José dos Campos. **Anais eletrônicos...** São José dos Campos: INPE, Out. 2005. Disponível em: <http://hermes2.dpi.inpe.br:1905/col/dpi.inpe.br/hermes2@1905/2005/10.03.20.04/doc/Model%20Breeder%20-%20Worcap2005.pdf>. Acesso em: 01/08/2006.

SANTANA, F. S. *et al.* A reference business process for ecological niche modelling. **Ecological Informatics**, v. 3, n. 1, p. 75-86, Jan. 2008.

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

SIQUEIRA, M. F. **Uso de modelagem de nicho fundamental na avaliação do padrão de distribuição geográfica de espécies vegetais**. 2005. 107 p. Tese (Doutorado em Ciências de Engenharia Ambiental) - Universidade de São Paulo, São Carlos.

SOLOMATINE, D. P. Data-driven modelling: paradigm, methods, experiences. In: International conference on hydroinformatics, 5, 2002. Cardiff. **Proceedings...** Londres: IWA Publishing, Jul. 2002. p. 757-763.

SPENS, J.; ENGLUND, G.; LUNDQVIST, H. Network connectivity and dispersal barriers: using geographical information system (GIS) tools to predict landscape scale distribution of a key predator (*Esox lucius*) among lakes. **Journal of Applied Ecology**, v. 44, n. 6, p. 1127-1137, Set. 2007.

STOCKWELL, D. R. B. Genetic algorithms II. In: FIELDING, A. H. (Ed.) **Machine learning methods for ecological applications**. Boston: Kluwer Academic Publishers, 1999. p. 123-144.

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

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

TOBIAS, J. A.; KOCH, P.; MERKOD, C. **Colibríes de la cuenca de Madre de Dios, Peru**. 2008. Disponível em: <http://www.zoo.ox.ac.uk/egi/Hummingbirds.pdf>. Acesso em: 12/10/2008.

TOBLER, W. R. A computer model simulation of urban growth in the Detroit region. **Economic Geography**, v. 46, n. 2, p. 234-240, 1970.

VOLTERRA, V. Fluctuations in the abundance of a species considered mathematically. **Nature**, v. 118, n. 2972, p. 558-560, Out. 1926.

WISSEL, C. Aims and limits of ecological modelling exemplified by island theory. **Ecological Modelling**, v. 63, n. 1-4, p. 1-12, Set. 1992.

XIAO, N.; BENNETT, D. A.; ARMSTRONG, M. P. Using evolutionary algorithms to generate alternatives for multiobjective site-search problems. **Environment and plannig A**, v. 34, n. 4, p. 639-656, Abr. 2002.

ANEXO A – ALGORITMOS BIOCLIM E GARP

A.1 Algoritmo BIOCLIM

O algoritmo BIOCLIM implementa o conceito de envelope bioclimático (Nix, 1986). O algoritmo calcula a média e o desvio-padrão para cada variável ambiental associada aos pontos de presença da espécie, assumindo uma distribuição normal. Cada variável tem seu próprio envelope representado pelo intervalo $[m - c \cdot s, m + c \cdot s]$, onde *m* é a média, *c* é um parâmetro que representa o ponto de corte e *s* é o desvio padrão. A implementação do algoritmo BIOCLIM, disponível no openModeller Desktop v1.0.6 (CRIA *et al.*, 2008) utiliza *c* = 0,674 como valor padrão. Além do envelope, cada variável ambiental possui também os limites superior e inferior correspondentes aos valores mínimo e máximo associados ao conjunto de pontos de ocorrência da espécie.

Assim, no algoritmo BIOCLIM, qualquer ponto do espaço pode ser classificado como:

a) Adequado: todos os valores para as variáveis ambientais, associadas ao ponto avaliado, encontram-se dentro dos limites do envelope calculado;

b) Marginal: ao menos uma das variáveis ambientais, associadas ao ponto avaliado, possui valor fora do envelope calculado, mas ainda dentro dos limites mínimo e máximo para aquela variável;

c) Inadequado: ao menos uma das variáveis ambientais, associadas ao ponto avaliado, possui valor fora dos limites mínimo e máximo da variável.

A.2 Algoritmo GARP

O GARP (*Genetic Algorithm for Rule Set Prediction*) é um conjunto de módulos desenvolvido para criar modelos de distribuição de espécies a partir de dados raster ambientais e biológicos. Estes módulos executam um conjunto diversificado de funções analíticas automaticamente, possibilitando a produção rápida e não-supervisionada de distribuições de animais ou plantas (Payne e Stockwell, 2001).

GARP é um algoritmo genético que cria modelos de nichos ecológicos para espécies. Os modelos descrevem condições ambientais sobre as quais as espécies podem desenvolver-se. Como dados de entrada, o GARP usa um conjunto de pontos amostrais onde a espécie ocorre e um conjunto de layers geográficos que representam os parâmetros ambientais que podem delimitar a sobrevivência da espécie.

A robustez dos AGs é uma característica bem conhecida. O GARP possui uma característica que acentua a capacidade dos AGs de gerar e testar uma ampla faixa de soluções candidatas – a capacidade de gerar e testar diversos tipos de modelos (regras) como modelos categóricos, por faixas e logísticos (Stockwell e Peters, 1999).

A.2.1 Regras

Um algoritmo genético pode ser visto como uma máquina de aprendizado. O algoritmo genético GARP é responsável por criar um conjunto de regras. Cada regra é um modelo em si mesma; um condicional se-então utilizado para fazer inferência sobre os valores de uma variável de interesse. O conjunto de regras desenvolvido pelo GARP é mais precisamente descrito como um modelo

inferencial do que como um modelo matemático. Modelos inferenciais diferem de modelos matemáticos no ponto em que os primeiros estão mais relacionados com a lógica do que com matemática e o processo básico é a inferência lógica ao invés de cálculos (Stockwell e Peters, 1999). A forma geral de uma regra é visualizada na Figura A.1.

Se A então B, e A é verdadeiro, então ocorre B.

Figura A.1 – Forma Geral de uma Regra

A precisão da regra é determinada a partir de cálculos probabilísticos simples. Um conjunto de dados pode estar identificado com a condição de uma regra (por exemplo o conjunto de dados onde a precipitação está entre 600 mm e 700 mm). A probabilidade de ocorrência das espécies pode ser calculada a partir do número de células no qual a espécie ocorre dividido pelo número total de células. Quatro tipos de regras estão presentes no GARP: regras atômicas, regras BIOCLIM, regras de faixas e regras logísticas.

- Regras Atômicas

É o tipo mais simples de regra utilizada pelo GARP. Uma regra atômica usa somente um valor para cada variável na condição da regra. A Figura A.2 mostra um exemplo deste tipo de regra.

Se Temp = 23 e Elevação = 250 então Presente

Figura A.2 – Exemplo de Regra Atômica

- Regras BIOCLIM

Uma regra BIOCLIM está baseada no modelo utilizado no programa BIOCLIM (Nix, 1986). O programa BIOCLIM produz um modelo envelopando os valores

ambientais para os quais determinada espécie ocorre; este envelope é definido estatisticamente, tipicamente considerando a faixa do percentil 95. Isto é, o envelope ambiental definido na regra envolve 95% dos pontos de dados onde determinada espécie ocorre. Um ponto analisado é predito como presente se estiver contido dentro do envelope e ausente em caso contrário. A Figura A.3 mostra um exemplo deste tipo de regra.

Se TempMedia = (23, 29] e TempMin = (10, 16] e
TempMax = (35, 38] e Elevação = (140, 650] e
:
Umidade = (20, 60] e Precipitação = (1400, 1750] então Presente

Figura A.3 – Exemplo de Regra BIOCLIM

Regras BIOCLIM não estão restritas apenas às variáveis climáticas; qualquer variável pode ser usada. Estas regras podem predizer tanto a presença como ausência de uma espécie, mas nunca ambas. A negação de uma regra BIOCLIM pode ser usada para predizer a presença ou ausência de uma espécie.

- Regras de Faixas

É uma generalização das regras BIOCLIM. Numa regra de faixa várias variáveis podem ser consideradas irrelevantes. Um exemplo deste tipo de regra é apresentado na Figura A.4.

Se TempMedia = (23, 27] e Elevação = (228, 1480] então Ausente

Figura A.4 – Exemplo de Regra de Faixas

- Regras Logísticas

Regras logísticas são uma adaptação dos modelos de regressão logísticos.

Uma regressão logística segue uma equação onde a saída é transformada numa probabilidade. Por exemplo, a regressão logística tem como saída uma probabilidade *p* indicando se uma regra deve ser aplicada. *p* é calculada através da Equação A.1.

$$p = \frac{1}{1 + e^{-y}} \tag{A.1}$$

onde $y = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \cdots + \beta_n \cdot x_n)$ é uma equação obtida por análise de regressão linear múltipla. Um exemplo deste tipo de regra é apresentado na Figura A.5.

*Se 0,1 – Elevação * 0,1 + TempMedia * 0,3 então Ausente*

Figura A.5 – Exemplo de Regra Logística

A.2.2 Codificação das Regras

Para que as regras sejam manipuladas pelo GARP é necessário que sejam codificadas numa estrutura manipulável computacionalmente. Esta estrutura, nos AGs, recebe o nome de cromossomo. O conjunto de regras apresentados na Figura A.6 foi codificado nos cromossomos apresentados na Tabela A.1.

Regra 1*: Se TMIN = (5, 10] e TMED = (10, 22] e ELEV = (1, 2] então Presente*
Regra 2: Se TMIN = (0, 15] e TMED = (0, 50] e ELEV = (0, 20] então Ausente
*Regra 3: Se TMIN * 0,80 + TMED * -0,2 + ELEV * 0,45 então Ausente*

Figura A.6 – Conjunto de regras

Tabela A.1 – Cromossomos que codificam o conjunto de regras da Figura A.6

Regra	TMIN	TMIN	TMED	TMED	ELEV	ELEV	P/A
1	5	10	10	22	1	2	P
2	0	15	0	50	0	20	A
3	0,8	---	-0,2	---	0,45	---	A

A.2.3 Mecanismo Evolutivo

O mecanismo evolutivo do GARP utiliza-se de 4 operadores genéticos: a seleção, o cruzamento, a junção e a mutação.

O mais importante operador de recombinação é o cruzamento. Este operador combina partes de dois cromossomos pais (Regra 1 e Regra 2) gerando filhos (Regra 4 e Regra 5) que carregam características dos cromossomos pais. A Figura A.7 mostra um exemplo de aplicação do operador cruzamento no GARP.

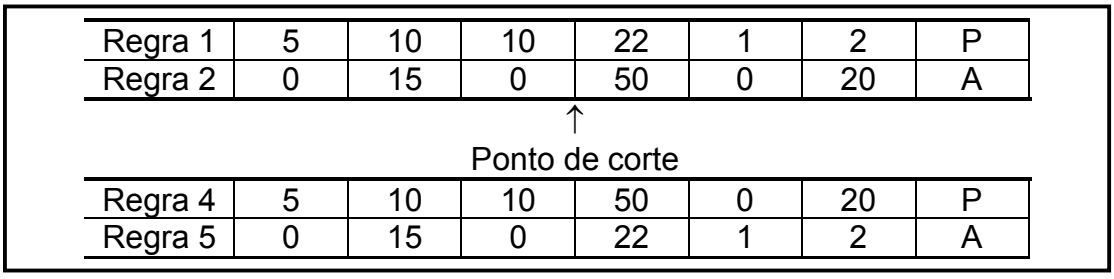

Figura A.7 – Exemplo de operação cruzamento sobre as regras no GARP

O operador de junção é ligeiramente distinto do operador de cruzamento; somente um filho é gerado a partir da combinação de dois pais. A Figura A.8 mostra um exemplo do operador de junção.

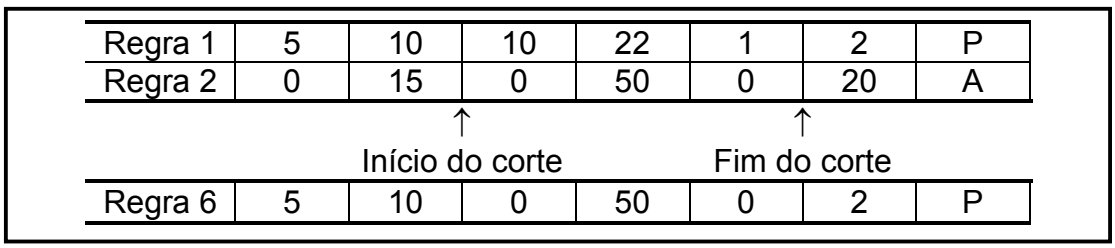

Figura A.8 – Exemplo de operação de junção sobre as regras no GARP

Há dois operadores de mutação disponível no GARP. O primeiro deles, chamado de mutação randômica consiste em substituir um valor qualquer do cromossomo por outro no intervalo entre 1 e 254. O segundo deles, chamada

de mutação incremental, consiste em adicionar uma unidade ao valor selecionado no cromossomo. A Figura A.9 mostra um exemplo para cada uma destas mutações.

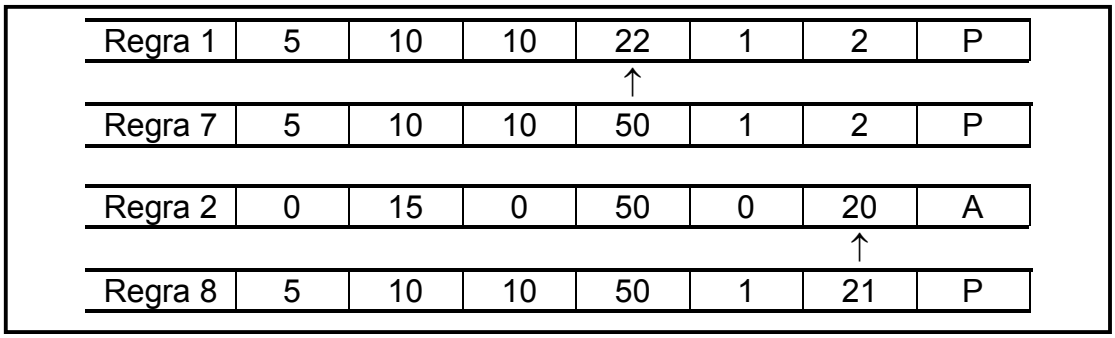

Figura A.9 – Exemplo de operações de mutação sobre as regras no GARP

Os objetivos do GARP são dois: maximizar a significância e a precisão das regras sem criar o problema de *overfitting* ou regras por demais especializadas. A significância é medida através de um teste χ^2 sobre a diferença entre as probabilidades preditas a priori e a posteriori pela regra. Maximizar a significância e a precisão preditiva é uma inovação nos sistemas analíticos; muitos modelos maximizam apenas a significância.

Overfitting é um problema que está sempre presente na modelagem. Um modelo que apresenta *overfit* pode ser excelente sobre os dados para o qual foi ajustado, mas ter uma capacidade preditiva muito pobre.

O processo de avaliação consiste em testar cada uma das regras utilizando um conjunto de dados de teste, previamente selecionado. O valor obtido pela Equação A.2 será o valor de aptidão da regra.

$$Sig = \frac{pXYs - no \cdot \frac{pYs}{n}}{\sqrt{no \cdot pYs \cdot \left(1 - \frac{pYs}{n}\right) \Big/ n}} \qquad \text{(A.2)}$$

onde *Sig* é o valor de aptidão da regra (significância), *pXYs* é o número de pontos amostrados que a regra prevê corretamente, *no* é o número de pontos amostrados avaliados pela regra, *pYs* é o número de pontos amostrados com a mesma conclusão que a regra e *n* é o número total de pontos amostrados.

Ordenando-se as regras pelo índice de aptidão inicia-se o processo de seleção. Um limite de corte é estabelecido e, os indivíduos abaixo deste limite, são descartados. Os indivíduos restantes, os mais aptos, passam novamente pelo processo evolutivo, até que o critério de parada seja atingido. A Figura A.10 ilustra o processo de seleção, onde *f(r)* é equivalente à aptidão da regra, obtida pela Equação A.2.

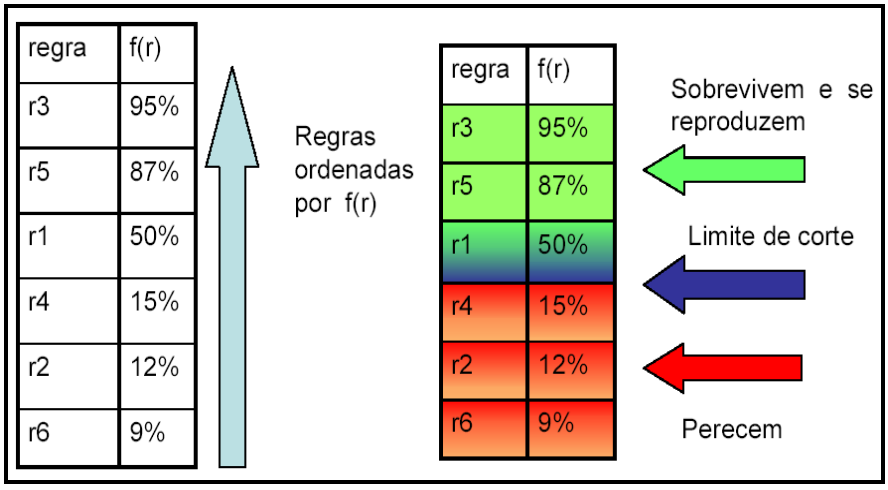

Figura A.10 – O processo de seleção do GARP

Há dois critérios de parada no GARP. O primeiro deles é um número pré-estabelecido de gerações. O segundo consiste em contar o número de melhores regras que são geradas no processo. Caso este número seja inferior a um limiar pré-estabelecido o processo evolutivo pára.

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.