



MINISTÉRIO DA CIÊNCIA E TECNOLOGIA

**INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS**

UM ALGORITMO GENÉTICO COM REPRESENTAÇÃO EXPLÍCITA DE  
RELACIONAMENTOS ESPACIAIS PARA MODELAGEM SÓCIO-AMBIENTAL

Adair Santa Catarina

Proposta de Tese de Doutorado em Computação Aplicada, Orientada pelos  
Drs. João Ricardo de Freitas Oliveira e Antônio Miguel Vieira Monteiro

INPE  
São José dos Campos  
2006



## RESUMO

A utilização intensiva de computadores e de algoritmos computacionais no estudo, análise e representação de dados espaciais fez surgir uma nova área de conhecimento: a geoinformática. Um algoritmo computacional utilizado na análise de dados espaciais é o algoritmo genético, utilizado com sucesso num sistema de modelagem de distribuição de espécies, o GARP (*Genetic Algorithm for Rule Set Prediction*). Apesar deste sistema ser bem aceito por biólogos, ecólogos e outros, a dependência espacial, um aspecto intrínseco dos fenômenos espaciais, é nele negligenciada, assim como o conhecimento existente sobre as características da região que interferem na dispersão da espécie em estudo. Estas limitações não são exclusividade do GARP, mas sim decorrentes da ausência de um algoritmo genético que trate adequadamente os relacionamentos espaciais. Tomando em consideração estas limitações, esta proposta busca incorporar num algoritmo genético a representação explícita de relacionamentos espaciais. Para tanto, uma matriz de vizinhança generalizada é inserida na estrutura do algoritmo, objetivando considerar a dependência espacial presente no fenômeno em estudo. Outra inovação é a inserção de *layers* de peso associadas às variáveis manipuladas por estes algoritmos permitindo computar o efeito local das mesmas, indicando sua importância para o problema em estudo. Para verificar a qualidade do algoritmo genético com representação explícita de relacionamentos espaciais um estudo de caso será realizado. Neste estudo de caso o algoritmo proposto será aplicado num sistema para criação de modelos de distribuição de espécies.



## SUMÁRIO

<b>LISTA DE FIGURAS.....</b>	<b>iii</b>
<b>1 INTRODUÇÃO.....</b>	<b>1</b>
1.1 JUSTIFICATIVA.....	3
1.2 HIPÓTESE.....	4
1.3 OBJETIVO GERAL.....	4
1.3.1 Objetivos Específicos.....	4
1.4 ORGANIZAÇÃO DO TEXTO.....	5
<b>2 CONCEITOS GERAIS DE GEOINFORMÁTICA.....</b>	<b>7</b>
2.1 DEPENDÊNCIA ESPACIAL.....	7
2.2 VIZINHANÇA ESPACIAL.....	8
2.3 MATRIZ DE VIZINHANÇA GENERALIZADA.....	9
<b>3 ALGORITMOS GENÉTICOS (AGs).....</b>	<b>11</b>
3.1 INTRODUÇÃO AOS AGs.....	11
3.2 OS OPERADORES GENÉTICOS.....	15
3.2.1 Seleção por Monte Carlo.....	15
3.2.2 Elitismo.....	16
3.2.3 Cruzamento e Mutação.....	17
3.3 PARÂMETROS GENÉTICOS.....	21
3.4 HIBRIDIZAÇÃO.....	22
3.5 <i>SIMULATED ANNEALING</i> (SA).....	22
3.5.1 O Algoritmo SA.....	25
<b>4 MECANISMOS HEURÍSTICOS SEMI-AUTOMÁTICOS EM ANÁLISE ESPACIAL.....</b>	<b>26</b>
4.1 APLICAÇÕES DE AGs NA ANÁLISE DE DADOS GEOGRÁFICOS.....	26
4.1.1 Plano de Uso de Solos.....	27
4.1.2 Construção de Modelos de Interação Espacial.....	28
4.2 MODEL BREEDERS.....	29
4.3 GARP MODELLING SYSTEM – GENETIC ALGORITHM FOR RULE SET PREDICTION.....	31
4.3.1 Regras.....	32
4.3.2 Codificação das Regras.....	34
4.3.3 Mecanismo Evolutivo.....	35
4.3.4 Avaliação da Qualidade dos Modelos Ajustados com o GARP.....	38
<b>5 DETALHAMENTO DA PROPOSTA.....</b>	<b>41</b>

5.1 AS INOVAÇÕES NA ARQUITETURA PROPOSTA.....	41
5.2 OBTENÇÃO DOS MODELOS DE DISTRIBUIÇÃO DE ESPÉCIES.....	43
5.3 CRONOGRAMA.....	46
<b>6 CONSIDERAÇÕES FINAIS .....</b>	<b>47</b>
<b>REFERÊNCIAS BIBLIOGRÁFICAS.....</b>	<b>49</b>

## LISTA DE FIGURAS

Figura 2.1	– Grade regular sobre o estado de Tennessee, EUA .....	8
Figura 2.2	– Exemplos de vizinhança.....	9
Figura 3.1	– Fluxograma que descreve brevemente um algoritmo genético.....	13
Figura 3.2	– Um exemplo de seleção através do método da roleta.....	16
Figura 3.3	– Um exemplo do operador de cruzamento em um ponto.....	19
Figura 3.4	– O algoritmo SA.....	25
Figura 4.1	– Exemplo da codificação empregada no cromossomo .....	30
Figura 4.2	– Representação de um sistema de modelagem de distribuição de espécies...	31
Figura 4.3	– Forma Geral de uma Regra.....	32
Figura 4.4	– Exemplo de Regra Atômica.....	33
Figura 4.5	– Exemplo de Regra BIOCLIM .....	33
Figura 4.6	– Exemplo de Regra de Faixas .....	34
Figura 4.7	– Exemplo de Regra Logística .....	34
Figura 4.8	– Conjunto de regras.....	35
Figura 4.9	– Exemplo de operação cruzamento sobre as regras no GARP.....	35
Figura 4.10	– Exemplo de operação de junção sobre as regras no GARP.....	36
Figura 4.11	– Exemplo de operações de mutação sobre as regras no GARP.....	36
Figura 4.12	– O processo de seleção do GARP .....	37
Figura 4.13	– Matriz de Confusão .....	38
Figura 4.14	– Curva padrão para a relação Omissão/Comissão .....	39
Figura 4.15	– Classificação das regiões quanto à intensidade dos erros de omissão/comissão.....	39
Figura 4.16	– Aplicação do limiar para eliminar modelos com muitos erros de omissão ....	40
Figura 4.17	– Identificação da região com os melhores modelos ajustados .....	40
Figura 5.1	– Sistema de modelagem de distribuição de espécies.....	41
Figura 5.2	– Estrutura proposta para um novo sistema de modelagem de distribuição de espécies .....	42
Figura 5.3	– Linearização do espaço utilizando o esquema de indexação de Morton.....	44
Figura 5.4	– Índice de Morton para uma matriz de ordem 8 .....	44
Figura 5.5	– Regiões operadas pelo cruzamento UNBLOX .....	45



## 1 INTRODUÇÃO

A busca pelo conhecimento presente em conjuntos de dados geográficos, através do uso intensivo de computadores, fez surgir uma nova área de conhecimento: a geoinformática.

O termo *geocomputation* 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 que tratam de análise estatística espacial, visualização de dados geoespaciais, modelos dinâmicos de interação espacial e modelos de dinâmicas espaço-temporais, entre outros. Uma definição sucinta, porém adequada nesta proposta, é dada por OPENSHAW e ABRAHART (2000)<sup>1</sup> como sendo o processo de aplicação da tecnologia computacional para a solução de problemas geográficos.

Na geoinformática um conjunto de algoritmos “inteligentes” tem sido utilizados para explorar conjuntos de dados geográficos. Dentre estes algoritmos de busca “inteligentes” estão os Algoritmos Genéticos (AGs), utilizados na implementação de *software* como os *model breeders* (OPENSHAW e OPENSHAW, 1997; SANTA CATARINA, 2005) e no GARP (STOCKWELL e PETERS, 1993).

*Model breeders* são sistemas semi-automáticos que analisam conjuntos de variáveis dependentes e independentes encontrando relações entre elas. Estas relações são expressas através de uma expressão aritmética, como nos modelos de regressão estatística. O GARP é uma ferramenta amplamente utilizada para modelagem de nichos ecológicos, ou seja, dado um conjunto amostra de uma determinada espécie e um conjunto de *layers* de informações ambientais e climáticas, o algoritmo é capaz de criar um modelo para predizer quais regiões são aptas para o desenvolvimento daquela espécie.

Os AGs são algoritmos heurísticos de busca, que utilizam regras baseadas numa metáfora do processo evolutivo proposto por Charles Darwin, operando sobre um

---

<sup>1</sup> Geocomputation is the process of computational technology application to solve geographical problems (OPENSHAW e ABRAHART, 2000, p.19).

espaço de soluções codificado (HOLLAND, 1975; GOLDBERG, 1989). Estes algoritmos são utilizados na resolução de problemas combinatórios complexos, onde as técnicas determinísticas de resolução não conseguem encontrar a solução em tempo hábil. O uso de AGs possibilita que uma solução aproximadamente ótima destes problemas seja encontrada num tempo viável.

Um conceito chave no entendimento e análise dos fenômenos espaciais é o conceito de dependência espacial. Este conceito é oriundo da primeira lei da geografia conhecida como Lei de Tobler<sup>2</sup>. Em outras palavras, a maior parte das ocorrências naturais, ou sociais, apresentam entre si uma relação que depende da distância.

Este conceito ressalta a importância de se considerar o espaço na vizinhança do fenômeno em estudo. Entretanto a vizinhança, e conseqüentemente a dependência espacial, é negligenciada nos *softwares* anteriormente citados. Esta negligência não é exclusividade destes *softwares*; ela decorre da ausência de um algoritmo genético que incorpore, em seus mecanismos evolutivos, os relacionamentos espaciais.

Assim, a necessidade de tratar adequadamente os relacionamentos espaciais no processo de análise de dados espaciais, particularmente usando mecanismos semi-automáticos baseados em AGs, desperta questionamentos, como:

- É possível incorporar os relacionamentos espaciais nestes mecanismos semi-automáticos de análise de dados geográficos?
- Estes mecanismos de análise de dados geográficos são capazes de operar sobre um modelo generalizado de relacionamentos espaciais?
- Estes mecanismos de análise de dados geográficos são capazes de quantificar o efeito dos relacionamentos espaciais sobre as variáveis envolvidas num fenômeno espacial?
- O conhecimento sobre as influências, favoráveis ou desfavoráveis, das formações naturais ou artificiais pode ser representado nestes mecanismos?

---

<sup>2</sup> "Everything is related to everything else, but near things are more related than distant things" (TOBLER, 1970, p.236).

A primeira questão remete à necessidade de computar os efeitos dos relacionamentos espaciais, de considerar os efeitos que a região na vizinhança de um ponto do espaço exercem sobre o fenômeno em estudo; uma alusão direta à Lei de Tobler.

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

A terceira questão relaciona-se com a quantificação dos efeitos dos relacionamentos espaciais. Esta quantificação tornará possível identificar quais variáveis afetam, de modo mais significativo, o fenômeno em estudo.

A quarta questão trata da inserção do conhecimento de um especialista sobre o fenômeno e a região em estudo. Os elementos constituintes da paisagem, sejam naturais ou artificiais, podem interferir sobre o fenômeno de modo favorável ou desfavorável. Quando esta interferência é significativa este conhecimento deveria ser inserido e considerado no processo.

Esta proposta de tese admite que as respostas das questões anteriormente apresentadas são afirmativas. O desenvolvimento de um AG com representação explícita de relacionamentos espaciais, aplicado na modelagem de distribuição de espécies, possibilitará validar este pressuposto.

## 1.1 JUSTIFICATIVA

Esta proposta busca estender mecanismos, como os *Model Breeders* e o GARP, utilizados em geoinformática. A representação explícita de relacionamentos espaciais nos algoritmos de análise espacial, particularmente nos AGs, caracteriza uma inovação relevante pois possibilita considerar os efeitos da dependência espacial nos mesmos.

A representação explícita de relacionamentos espaciais também permite a inserção do conhecimento sobre os elementos naturais e artificiais que compõem a região e que afetam o fenômeno em estudo.

Outro aspecto relevante é a quantificação dos relacionamentos espaciais vinculados às variáveis presentes no fenômeno em estudo. Ela possibilitará avaliar a importância destas variáveis, permitindo descobrir quais delas interferem significativamente no fenômeno. Assim, pode-se simplificar estudos futuros, tornando desnecessária a coleta de informações sobre variáveis não significativas.

## 1.2 HIPÓTESE

As hipóteses sobre as quais esta proposta foi elaborada são:

- a) Os AGs, utilizados em sistemas de modelagem sócio-ambiental, podem incorporar uma representação explícita de relacionamentos espaciais, através da qual pode-se inserir o conhecimento sobre os elementos naturais e artificiais presentes na região em estudo, resultando em modelos mais confiáveis;
- b) A representação explícita de relacionamentos espaciais fornece meios para quantificar os efeitos destes relacionamentos sobre as variáveis do fenômeno em estudo, possibilitando avaliar quais delas afetam significativamente o fenômeno em estudo.

## 1.3 OBJETIVO GERAL

Esta proposta tem como objetivo principal incorporar nos AGs, utilizados na modelagem de fenômenos sócio-ambientais, uma estrutura de representação explícita de relacionamentos espaciais.

### 1.3.1 Objetivos Específicos

Como objetivos específicos pode-se citar:

- Elaborar um esquema de codificação, que seja capaz de representar explicitamente os relacionamentos espaciais, para ser utilizada em AGs aplicados no processo de análise de dados espaciais;

- Definir uma função de avaliação que considere o conhecimento sobre os elementos naturais e artificiais existentes no espaço e que afetam o fenômeno em estudo;
- Realizar um estudo de caso: implementação de um sistema para criação de modelos de distribuição de espécies, usando um AG que incorpore a codificação elaborada e a função de avaliação definida;
- Avaliar a qualidade dos modelos obtidos pelo sistema implementado no estudo de caso.

#### 1.4 A PROPOSTA NO CONTEXTO DAS ATIVIDADES DE PESQUISA DO INPE

No contexto das atividades de pesquisa do Instituto Nacional de Pesquisas Espaciais, esta proposta está inserida na linha de pesquisa em Geoprocessamento do programa de pós-graduação em computação aplicada.

Diversas teses e dissertações foram elaboradas nesta área, tanto no programa de pós-graduação em computação aplicada quanto no programa de pós-graduação em sensoriamento remoto. Como exemplo podemos citar os trabalhos de PEDROSA (2003), ALMEIDA (2003), SILVA (2004), CARNEIRO (2004), AGUIAR (2004) e FEITOSA (2005).

A partir de 2005 este Instituto participa do projeto *Open Modeller*, financiado pela FAPESP, em parceria com o Centro de Referência em Informação Ambiental (CRIA) e a Universidade de São Paulo (USP). Este projeto visa desenvolver um ambiente de software livre para modelagem de nichos ecológicos para biodiversidade.

#### 1.5 ORGANIZAÇÃO DO TEXTO

Os capítulos de 2 a 4 desta proposta são de fundamentação teórica. No capítulo 2 são apresentados conceitos gerais em geoinformática que permeiam este trabalho. No capítulo 3 é apresentada uma classe de algoritmos heurísticos utilizados nos mecanismos semi-automáticos de análise de dados espaciais: os algoritmos

genéticos. No capítulo 4 são apresentados e discutidos alguns mecanismos heurísticos semi-automáticos utilizados na análise de dados espaciais. O capítulo 5 detalha a solução imaginada para o problema proposto como estudo de caso e apresenta o cronograma de atividades a realizar. O capítulo 6 apresenta as considerações finais desta proposta.

## 2 CONCEITOS GERAIS DE GEOINFORMÁTICA

Geoinformática é um campo de pesquisa emergente que propõe o uso de técnicas intensivas em computação, tais como redes neurais, busca heurística e autômatos celulares para análise de dados espaciais (CÂMARA e MONTEIRO, 2001).

A geoinformática inclui, também, a análise de dados espaciais, modelagem dinâmica, visualização e os processos dinâmicos espaço-temporais (LONGLEY, 1998).

OPENSHAW (1999) argumenta que geoinformática não é apenas o uso de técnicas computacionais para resolver problemas espaciais, mas um caminho completamente novo para se fazer ciência num contexto geográfico.

O entendimento de alguns conceitos, como dependência espacial, correlação espacial, vizinhança generalizada faz-se necessário para aqueles que trabalham com análise espacial. Nas seções subseqüentes estes importantes conceitos são apresentados.

### 2.1 DEPENDÊNCIA ESPACIAL

A dependência espacial é um conceito chave no entendimento e análise dos fenômenos espaciais. Oriundo da primeira lei da Geografia, enunciada por Waldo Tobler: "*todas as coisas são parecidas, mas coisas mais próximas se parecem mais que coisas mais distantes*". Ou seja, a maior parte das ocorrências naturais, ou sociais, apresentam entre si uma relação que depende da distância (DRUCK *et al.*, 2004).

A dependência espacial é expressa e quantificada pela autocorrelação espacial. A autocorrelação é derivada do conceito de correlação, um conceito da estatística que mede o grau de associação entre variáveis aleatórias. A autocorrelação mede, então, o grau de associação de uma variável com ela mesma, quando amostrada em diferentes locais no espaço.

As medidas de autocorrelação espacial baseiam-se numa mesma idéia: verificar

como varia a dependência espacial, a partir da comparação entre os valores de uma amostra e de seus vizinhos (DRUCK *et al.*, 2004).

## 2.2 VIZINHANÇA ESPACIAL

O uso de grades regulares para representar o espaço bi-dimensional em geoinformática é corriqueiro. Alguns trabalhos que se utilizam desta representação são: HARRIS e CARTWRIGHT (1993), STOCKWELL e PETERS (1993), RUDOLPH e SPRAVE (1995), BENNETT *et al.* (1996), BENNETT *et al.* (1999), XIAO *et al.* (2000), XIAO *et al.* (2002) e GOUD (2003), entre outros.

As motivações para o uso deste modelo de dados são influenciadas pela natureza “pixelizada” 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 2.1, onde cada célula representa uma área de 64 km<sup>2</sup> (8 km x 8km).

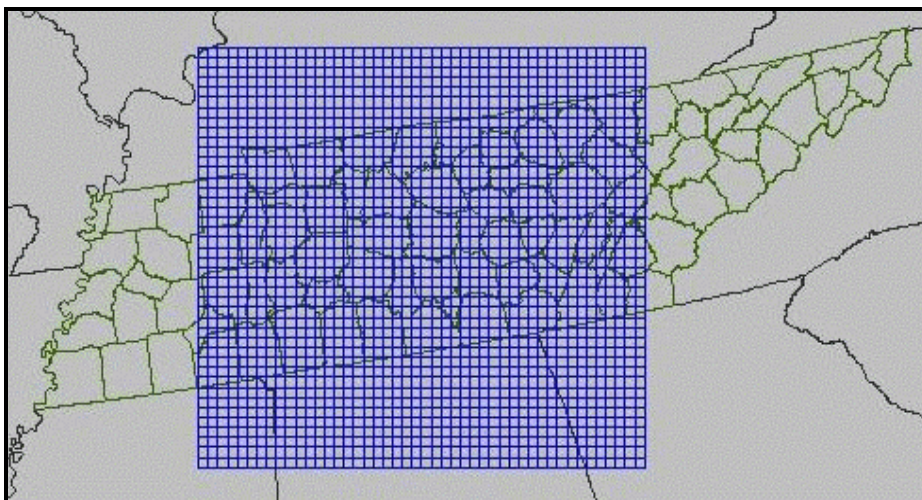


Figura 2.1 – Grade regular sobre o estado de Tennessee, EUA  
Fonte: Center for Environmental Modeling for Policy Development (2006)  
<http://www.cep.unc.edu/empd/projects/mims/spatial/installing.html>

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 2.2.

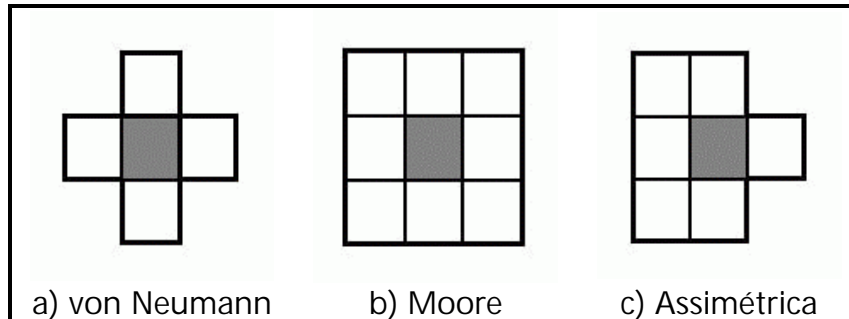


Figura 2.2 – Exemplos de vizinhança  
Fonte: Adaptado de PEDROSA (2003)

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.

No caso da distribuição das espécies de peixes ao longo de um rio, as espécies à jusante de uma cachoeira não conseguem superar esta barreira natural. No caso do 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.

Assim, uma representação mais adequada, para estudar estes fenômenos, seria aquela capaz de representar as relações de vizinhança entre as células. Em modelos celulares uma 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 matriz de vizinhança generalizada, vista em maiores detalhes na seção seguinte.

### 2.3 MATRIZ DE VIZINHANÇA GENERALIZADA

A matriz de vizinhança generalizada, ou matriz de proximidade 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 e PEDROSA, 2003).

Uma matriz de vizinhança generalizada é composta por um conjunto de objetos espaciais  $O$ , um grafo  $G$  e uma matriz de proximidade  $V$ :

- Os objetos espaciais ( $O$ ) são representados por células regulares ou polígonos, de acordo com a representação espacial utilizada.
- 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.
- 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:

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

### 3 ALGORITMOS GENÉTICOS (AGs)

Nesta seção será apresentada uma breve revisão bibliográfica sobre os AGs. Estes algoritmos são utilizados na construção de alguns mecanismos heurísticos semi-automáticos de análise de dados espaciais, como os *model breeders* e o GARP.

Inicialmente a revisão conta com o histórico e com a apresentação da estrutura básica destes algoritmos. Posteriormente serão apresentados os principais operadores genéticos e a influência dos parâmetros genéticos sobre o desempenho dos mesmos. Ao final da seção será apresentada uma tendência observada com as heurísticas utilizadas em otimização: a hibridização. No caso a hibridização dos AGs com *Simulated Annealing*.

#### 3.1 INTRODUÇÃO AOS AGs

Um AG é um tipo de algoritmo de busca que se utiliza do paradigma genético/evolucionário (HOLLAND, 1975). Os AGs foram criados com o intuito de imitar alguns dos processos observados na evolução natural das espécies. Os mecanismos que realizam esta evolução ainda não estão completamente compreendidos, mas algumas de suas características já são bem compreendidas e aceitas. A evolução acontece nos cromossomos, que são os elementos orgânicos responsáveis pela codificação genética dos seres vivos (DAVIS, 1996). As características e fenômenos específicos desta codificação ainda são objetos de muitas pesquisas. Segundo DAVIS (1996), as principais características gerais da teoria evolucionária que já são amplamente aceitas são:

- a) a seleção natural é um processo que atua sobre os cromossomos e, portanto, sobre os seres vivos que eles codificam;
- b) a seleção natural é o elo entre cromossomos e a performance das suas estruturas decodificadas. O processo de seleção natural faz com que os cromossomos que codificam estruturas bem sucedidas se reproduzam mais vezes e com maior probabilidade que as estruturas mal sucedidas;

- c) o processo de reprodução é o ponto onde a evolução acontece. Mutações podem provocar mudanças nos cromossomos dos filhos, fazendo com que eles sejam diferentes dos padrões genéticos dos seus pais, e processos de recombinação podem criar diferentes cromossomos para os filhos, pela combinação dos cromossomos dos pais;
- d) a evolução biológica não tem memória. Tudo o que se sabe sobre como produzir indivíduos bem adaptados ao seu meio ambiente está contido no seu genoma - o conjunto de cromossomos carregados pelos indivíduos da população atual - e na estrutura dos cromossomos decodificados.

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 *strings* de 0 e 1, a qual ele 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 os cromossomos que estavam mais adaptados e, com base nisto, aumentar as 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 AGs.

De maneira geral, um AG pode ser brevemente descrito através do fluxograma apresentado na Figura 3.1.

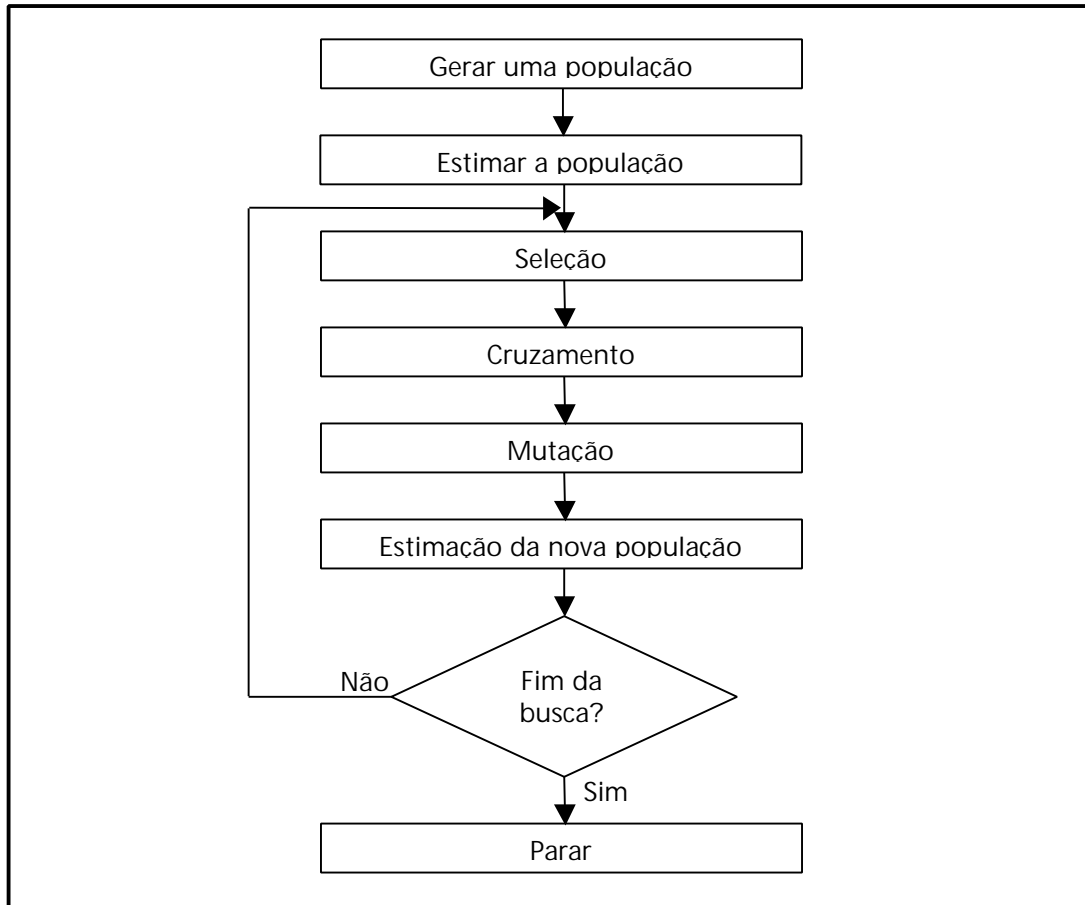


Figura 3.1 – Fluxograma que descreve brevemente um algoritmo genético.  
Fonte: Adaptado de CORTES (1999)

A técnica usada para codificar as soluções varia de problema para problema e de AG para AG. A codificação clássica usada no trabalho de Holland, e até hoje a mais usada, consistia em usar *strings* de bits, mas 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 bits na *string*. (GALVÃO e VALENÇA, 1999)

A consequência imediata do aumento da *string*, 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, LOZANO e VERDEGAY, 1996)

Uma das formas não clássicas de codificação mais utilizada é a codificação real. Esta forma de codificação consiste em representar, num gene ou cromossomo, uma variável numérica contínua 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 LUCASIU 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.

O elemento de ligação entre o AG e o problema a ser resolvido é a função de avaliação. A função de avaliação, chamada de função *fitness*, toma como entrada um cromossomo e retorna um número, ou lista de números, que representam a medida de performance do cromossomo com relação ao problema a ser resolvido. Esta função desempenha no AG o mesmo papel desempenhado pelo meio ambiente na teoria da evolução natural das espécies.

Segundo GOLDBARG e LUNA (2000), os AGs possuem as seguintes características gerais:

- a) Operam em um conjunto de pontos, denominado como população, e não a partir de pontos isolados;
- b) Trabalham com um conjunto de parâmetros codificados e não com os próprios parâmetros;
- c) Necessitam como informação somente o valor de uma função objetivo,

denominada função de adaptabilidade ou *fitness*;

d) Usam transições probabilísticas e não regras determinísticas.

### 3.2 OS OPERADORES GENÉTICOS

HOLLAND (1975) define três técnicas para criar filhos diferentes dos pais: cruzamento, mutação e inversão. Estes três elementos estão intimamente relacionados no modelo básico de um algoritmo genético; os três fazem a evolução da população acontecer.

A finalidade da seleção em um algoritmo 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 da função *fitness* mais elevado.

A mais conhecida e utilizada forma de fazer a seleção é a roleta, ou algoritmo Monte Carlo (DAVIS, 1996). Na seqüência apresentaremos o funcionamento da seleção através do mencionado algoritmo.

#### 3.2.1 Seleção por Monte Carlo

Na seleção através do algoritmo Monte Carlo, também conhecida como seleção por roleta, cada indivíduo da população é representado numa 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. Um exemplo de aplicação do método da roleta é apresentado na Figura 3.2.

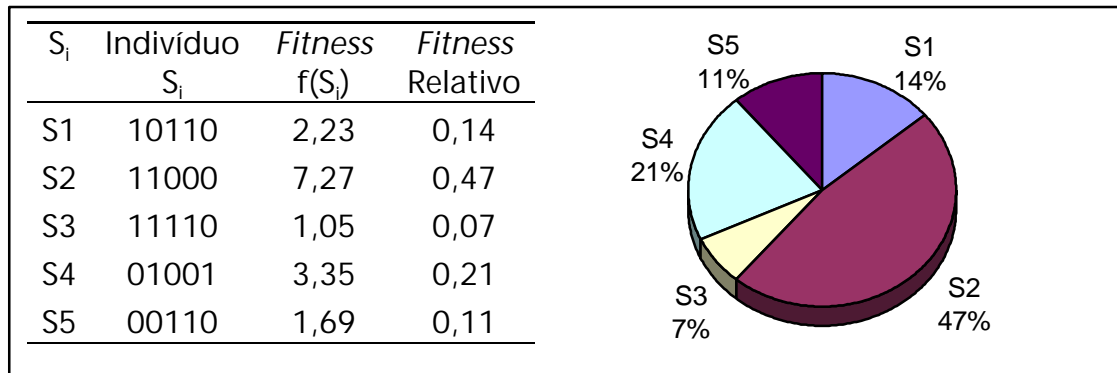


Figura 3.2 – Um exemplo de seleção através do método da roleta

### 3.2.2 Elitismo

Para melhorar a convergência dos AGs foi desenvolvida uma técnica chamada elitismo. O elitismo é a técnica mais utilizada para melhorar a convergência destes algoritmos. Ele foi primeiramente introduzido por Kenneth De Jong, em 1975, e é uma adição aos métodos de seleção que força os AGs a reter um certo número de "melhores" indivíduos em cada geração (YEPES, 2000). 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 reproduzir-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.

### 3.2.3 Cruzamento e Mutação

O objetivo final de ambos 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, e o operador de mutação é responsável pela introdução de pequenas mudanças aleatórias nos cromossomos dos filhos. Vários tipos de operadores de cruzamento foram desenvolvidos por vários pesquisadores, alguns adequados a um tipo específico de codificação dos cromossomos, outros com intenção de serem mais genéricos. Mencionaremos aqui apenas os operadores mais comumente utilizados.

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 *string* do cromossomo é avaliado para saber se este bit deverá sofrer uma mutação; caso este bit deva sofrer mutação, o seu valor é simplesmente trocado por um valor determinado aleatoriamente entre os valores que podem ser assumidos pelo cromossomo.

A Tabela 3.1 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<sup>os</sup> Aleatórios” indicam probabilidades muito baixas e, portanto, serão os genes que sofrerão mutação. Os dígitos em negrito na coluna “Cromossomo novo” são os genes alterados.

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).

Tabela 3.1 – Exemplos de mutação de bit

Cromossomo Anterior	N <sup>os</sup> Aleatórios				Novo bit	Cromossomo novo
0011	0,653	<b>0,001</b>	0,287	0,373	1	0111
1001	0,721	0,432	0,043	0,840	-	1001
1110	<b>0,002</b>	0,076	0,934	0,471	0	<b>0110</b>

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  $N(\mu, \sigma^2)$ , onde  $\mu$  é igual ao valor de gene a ser substituído e a variância é definida pelo pesquisador. GALVÃO e VALENÇA (1999) citam que o valor da variância 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 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.

O operador de cruzamento em um ponto é a técnica de cruzamento mais simples e a mais utilizada. Esta técnica consiste em dividir os cromossomos selecionados num ponto de sua cadeia, ponto este escolhido aleatoriamente. Após isso, copia-se para os novos cromossomos uma parte de cada um dos cromossomos selecionados - cromossomos pais, formando assim os novos cromossomos - 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 número de alelos permita o número desejado de combinações diferentes. A Figura 3.3 apresenta um exemplo do operador de cruzamento em um ponto.

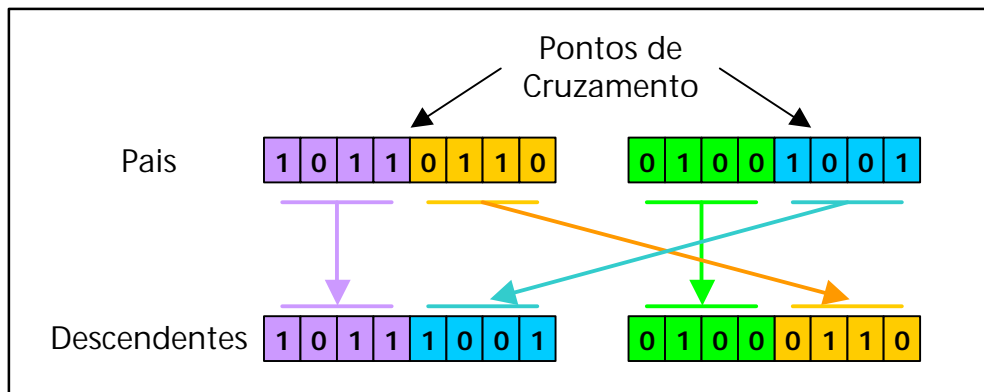


Figura 3.3 – Um exemplo do operador de cruzamento em um ponto.  
 Fonte: YEPES (2000)

Outra técnica de cruzamento, um pouco menos utilizada que a de cruzamento em um ponto, é o cruzamento em múltiplos pontos. Esta técnica 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 (DAVIS, 1996), média geométrica, *BLX-a* (ESHELMAN e SHAFFER, 1993), aritmético e heurístico (MICHALEWICZ, 1996).

Os cruzamentos média e média geométrica consistem em gerar um novo cromossomo usando a média simples e a média geométrica de dois cromossomos pais, respectivamente.

O cruzamento *BLX-a* consiste em gerar um novo cromossomo a partir da seguinte expressão:

$$c = p_1 + \mathbf{b}(p_2 - p_1) \quad (\text{eq. 3.1})$$

onde  $c$  é o novo cromossomo gerado,  $p_1$  e  $p_2$  são os cromossomos pais e  $\mathbf{b} \hat{I} U(-\mathbf{a}, 1 + \mathbf{a})$ .  $\mathbf{a}$  é um pequeno valor que estende os limites para a definição de  $c$ . Caso o cromossomo seja formado por múltiplos genes a eq. 3.1 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 expressão:

$$\begin{aligned}c_1 &= \mathbf{b}p_1 + (1 - \mathbf{b})p_2 \\c_2 &= (1 - \mathbf{b})p_1 + \mathbf{b}p_2\end{aligned}\tag{eq. 3.2}$$

onde  $\mathbf{b} \hat{I} U(0, 1)$ .

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. Então é produzido um cromossomo  $c$  da seguinte forma:

$$c = p_1 + r(p_1 - p_2), \text{ onde } f(p_1) > f(p_2)\tag{eq. 3.3}$$

onde  $r \hat{I} U(0, 1)$ .

Se compararmos os dois esquemas de reprodução, veremos que no esquema de reprodução sexuada é necessário haver mais de um tipo de indivíduo. Estes indivíduos devem ter diferenças significativas em alguns aspectos e devem desprender uma boa parcela de seu tempo e energia para encontrar um parceiro certo para a reprodução. Isto representa um custo a mais para o indivíduo/ algoritmo. Porém, como o esquema de reprodução sexuada parece ter vencido esta guerra, pode-se concluir que este talvez seja um preço pequeno a pagar, comparado aos benefícios que ele traz consigo.

Um benefício proporcionado pela reprodução sexuada é a combinação rápida 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.

### 3.3 PARÂMETROS 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 AGs. 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 maiores 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 muito 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 muito lento.

### 3.4 HIBRIDIZAÇÃO

A técnica de hibridização resulta na integração de uma boa maneira convencional de resolver um problema aos conceitos usuais de AGs. O resultado costuma ser melhor que o obtido com qualquer uma das duas técnicas isoladamente (DAVIS, 1996). A hibridização agrega a representação usual de dados no domínio original, bem como as técnicas de otimização usual já existentes. Isto permite a incorporação de heurísticas otimizadoras ao conjunto de operadores genéticos (recombinação 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.

Um exemplo de hibridização possível é quando o problema exige codificação com base em números reais e não em 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.

Uma heurística utilizada hibridamente com os AGs é o *Simulated Annealing*. A seção a seguir apresenta maiores detalhes desta heurística.

### 3.5 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 subsequente 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 função:

$$P_{c_k}(\text{aceitar } j) = \begin{cases} 1 & \text{se } g_j \leq g_i \\ \exp\left(\frac{-(g_j - g_i)}{c_k}\right) & \text{se } g_j > g_i \end{cases} \quad (\text{eq. 3.4})$$

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 eq. 3.4, se uma solução candidata  $j$  é melhor que a solução corrente  $i$ , ou seja ( $g_j \leq g_i$ ), 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 frequência, 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.

Entende-se por esquema de vizinhança o mecanismo apropriado, dependente do problema que está sendo tratado, 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.

### 3.5.1 O Algoritmo SA

Uma característica muito interessante do algoritmo SA é a simplicidade de sua implementação computacional, conforme mostrado na Figura 3.4.

```
Ler  $\alpha$  e NR; //Constante 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; // $\Delta E = g(S') - g(S)$ ;
    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 *  $\alpha$ ;
Fim enquanto;
```

Figura 3.4 – 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  $T = T * a$  corresponde ao processo de diminuição da temperatura, normalmente o parâmetro  $a$  é uma constante menor que um.

## 4 MECANISMOS HEURÍSTICOS SEMI-AUTOMÁTICOS EM ANÁLISE ESPACIAL

A aplicação de algoritmos computacionais na análise de dados geográficos possibilitou o desenvolvimento de uma nova área de conhecimento: a geoinformática.

Dentre os algoritmos computacionais utilizados estão aqueles desenvolvidos na área de inteligência artificial como a programação lógica, as redes neurais e os AGs. Estes algoritmos possibilitaram o desenvolvimento de mecanismos heurísticos semi-automáticos para análise de dados espaciais. A pesquisa bibliográfica realizada permitiu a identificação de duas classes de aplicação destes mecanismos heurísticos, apresentadas nesta seção.

Posteriormente dois destes mecanismos heurísticos semi-automáticos serão apresentados num nível maior de detalhes, permitindo uma melhor visualização da relação entre inteligência artificial e a análise de dados espaciais.

### 4.1 APLICAÇÕES DE AGs NA ANÁLISE DE DADOS GEOGRÁFICOS

Duas classes de aplicações de AGs na análise de dados geográficos serão aqui apresentadas: concepção de plano de uso de solos e construção de modelos de interação espacial e/ou espaço-temporal.

Na primeira classe de aplicações estão os problemas de concepção de planos de uso de solos. Problemas que envolvem a distribuição de parcelas de solo às atividades de modo a reduzir riscos ambientais maximizando lucros, por exemplo. Trabalhos nesta classe de problemas foram desenvolvidos por BENNETT, ARMSTRONG e WADE (1996), RONALD e KIRKBY (1998), MATTHEWS *et al.* (1999), MATTHEWS *et al.* (2000) e BJORNSSON e STRANGE (2000).

Na segunda classe de aplicações está a construção de modelos de interação espacial e/ou espaço-temporal. Os trabalhos de STOCKWELL e PETERS (1993), OPENSHAW e OPENSHAW (1997), FISCHER e LEUNG (1998), XIAO, BENNETT e ARMSTRONG (2000) estudam mecanismos baseados em AGs para criar modelos de

interação espacial a partir de vários tipos de dados.

#### 4.1.1 Plano de Uso de Solos

Problemas ambientais são decorrentes de decisões não coordenadas sobre o uso de solos. Quando os tomadores de decisão não trabalham em conjunto, impactos ambientais significativos podem ocorrer.

BENNET, ARMSTRONG e WADE (1996) mostraram como um algoritmo genético pode ser utilizado para construir um elo de ligação entre critérios de decisão e o espaço geográfico, evoluindo-os mutuamente até atingir soluções aceitáveis para problemas ambientais complexos. Agentes inteligentes foram utilizados para auxiliar os tomadores de decisão a considerar vários critérios, aprender com os sucessos e com os erros das soluções geradas. Este conhecimento pode ser usado para auxiliar na avaliação de soluções alternativas e gerar soluções melhores para problemas ambientais complexos.

RONALD e KYRKBY (1998) desenvolveram um AG capaz de delimitar fronteiras aproximadamente ótimas em mapas otimizando um conjunto de critérios relacionados com fronteiras. Distritos censitários foram agrupados em 9 áreas procurando mantê-las com o menor tamanho possível e ainda com um número similar de habitantes, visando um processo seletivo equânime de jogadores de futebol. Também apresentaram um novo operador genético, chamado "Operador de transferência de grupo de mutação mínima", que aliado à codificação utilizada proporciona um conjunto com fronteiras contínuas.

MATTHEWS *et al.* (1999) e MATTHEWS *et al.* (2000) utilizaram AGs para planejar o uso de solos. Consideraram este problema como um problema de alocação espacial. No primeiro trabalho otimizou-se o uso dos solos com base num único critério, o econômico. A evolução apresentada no segundo trabalho diz respeito a multi-otimização; neste trabalho além de se considerar o critério econômico considerou-se também questões como continuidade das áreas.

BJORNSSON e STRANGE (2000) aplicam um algoritmo genético para resolver um

problema ambiental que ocorre no oeste da Dinamarca: alagar lotes de terras para atividades agrícolas com o menor custo sem com isso causar desvios significativos no teor  $Fe^{+2}$  na água, o que alteraria as condições do habitat do salmão. Os autores informam resultados adequados, ou seja, conseguiu-se uma forma mais econômica de alocar as regiões alagáveis sem com isso prejudicar o habitat dos peixes. Informam também que o algoritmo implementado é computacionalmente caro.

#### 4.1.2 Construção de Modelos de Interação Espacial

STOCKWELL e PETERS (1993) construíram um sistema automático para modelar o comportamento de uma espécie biológica a partir de bases de dados. Este sistema, chamado GARP, tem em seu núcleo um Algoritmo Genético que busca um conjunto ótimo de regras que relaciona uma variável dependente, como a presença de uma determinada espécie vegetal numa região, com um conjunto de variáveis independentes. Este sistema tem sido utilizado para modelar nichos ecológicos de espécies animais e vegetais; alguns exemplos de aplicação do GARP são o monitoramento de espécies ameaçadas de extinção e predição da expansão de espécies invasoras.

OPENSHAW e OPENSHAW (1997) demonstram como utilizar um algoritmo genético para descobrir relações entre variáveis. Estas variáveis podem representar fenômenos naturais que variam no espaço. Perceber como estas variáveis estão relacionadas é um problema complexo pois matematicamente existem muitas formas de combinar as diferentes variáveis. O algoritmo genético neste caso é utilizado para encontrar quais os coeficientes, operações aritméticas e transcendentais adequados para relacionar as variáveis, ou seja, o modelo matemático que relaciona causas e efeitos. Este algoritmo genético é então chamado de criador de modelos (*model breeder*).

FISCHER e LEUNG (1998) apresentam um algoritmo genético com capacidade para maximizar funções com alto grau de complexidade. Este algoritmo é utilizado para ajustar a topologia de uma rede neural treinando-a e aumentando sua velocidade de convergência. Esta rede é, então, utilizada para modelar interação de dados

espaciais.

XIAO, BENNETT e ARMSTRONG (2000) utilizaram um estudo de caso em que modelaram a dispersão e a fixação de sementes de uma espécie de cipreste do pântano (*bald cypress*), caracterizando um processo complexo de otimização espaço-temporal.

## 4.2 MODEL BREEDERS

*Model Breeders* são ferramentas para modelagem automática OPENSHAW (1997). Estas ferramentas são capazes de encontrar um modelo matemático que relaciona variáveis independentes a uma variável dependente, seguindo a expressão geral:

$$y = f(x_1, x_2, \dots, x_n) \quad (\text{eq. 4.1})$$

Na ferramenta proposta por Openshaw, 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 (2005). O diferencial desta implementação diz respeito à forma de codificação utilizada nos cromossomos genéticos. Em sua proposta Openshaw utilizou-se da codificação binária, com as informações codificadas em cadeias de 0's e 1's. Nesta implementação utilizou-se uma codificação híbrida envolvendo cadeias binárias e números reais em base decimal. Assim, um polinômio geral:

$$y = c_1 \cdot x_i^{Exp_1} op_1 c_2 \cdot x_j^{Exp_2} op_2 \cdots c_n \cdot x_n^{Exp_n} \quad (\text{eq. 4.2})$$

onde:

- $y$ : variável independente;
  - $c_i$ : coeficiente de cada termo do polinômio;
  - $x_{ind}$ : variáveis independentes;
  - $Exp_i$ : expoentes das variáveis independentes;
  - $op_i$ : operadores que relacionam os termos do polinômio (+, -, x, /)
- foi assim representado:

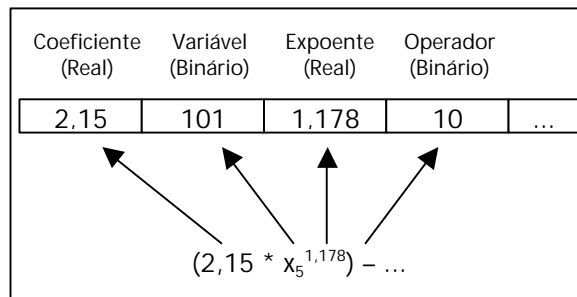


Figura 4.1 – Exemplo da codificação empregada no cromossomo

A função de avaliação utilizada baseou-se na soma dos quadrados dos desvios e é calculada pela expressão:

$$Fitness_k = \frac{Min(SQT_1, SQT_2, \dots, SQT_{Tp})}{SQT_k} \quad (\text{eq. 4.3})$$

onde:

- $Fitness_k$ : grau de aptidão da  $k$ -ésima solução, com  $k = 1..Tp$ ;
- $Tp$ : tamanho da população avaliada;
- $SQT$ : somatório dos quadrados dos desvios total:

$$SQT = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \quad (\text{eq. 4.4})$$

- $Y_i$ : valor assumido pela variável dependente na amostra  $i$ ;
- $\hat{Y}_i$ : valor estimado para a variável dependente na amostra  $i$ ;
- $n$ : número total de amostras coletadas.

Segundo o autor, a escolha da função (eq. 4.3) foi motivada por sua simplicidade e pela capacidade de medir adequadamente o ajuste do modelo encontrado. Os testes realizados conduziram a resultados considerados bons, permitindo concluir que a ferramenta era adequada para realizar uma análise exploratória dos dados.

### 4.3 GARP MODELLING SYSTEM

A modelagem matemática aliada às ferramentas computacionais gera a possibilidade da previsão de ocorrência de espécies através da geração de superfícies temáticas, indicando presença ou ausência, com os chamados modelos de distribuição de espécies (GUISAN e THUILLER, 2005). A Figura 4.2 é uma representação de um sistema de modelagem de distribuição de espécies.

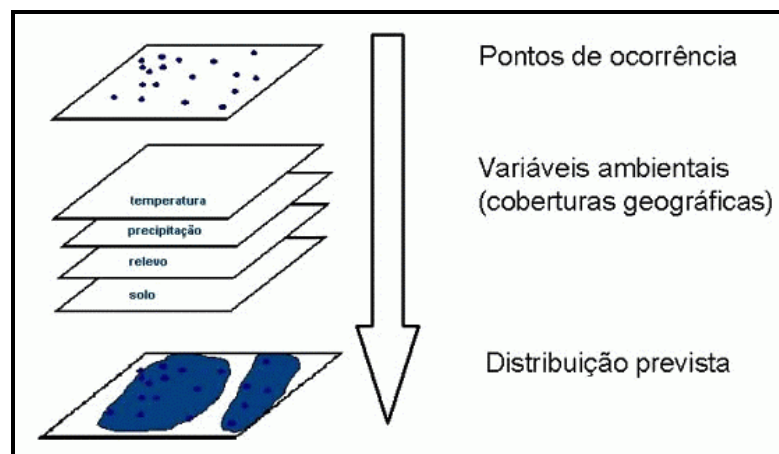


Figura 4.2 – Representação de um sistema de modelagem de distribuição de espécies

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).

O desenvolvimento de mapas de distribuição de espécies impõe uma carga enorme de trabalho num especialista. Cada questão obriga ao especialista a acessar uma base de dados, usando um pacote estatístico de modelagem, preparar e imprimir

mapas num GIS. A automação desta tarefa contribui para uma maior disponibilidade de dados diminuindo o tempo de resposta, o custo e ainda liberando especialistas para o desenvolvimento de tarefas mais desafiadoras (STOCKWELL e PETERS, 1999).

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).

#### 4.3.1 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, 1993).

A forma geral de uma regra é visualizada na Figura 4.3.

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

Figura 4.3 – 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 4.4 mostra um exemplo deste tipo de regra.

<i>Se TANN = 23 e GEO = 4 então PRESENT</i>
---

Figura 4.4 – 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 4.5 mostra um exemplo deste tipo de regra.

<i>Se TANN = (23, 29] e TMNCM = (10, 16] e          TMXWM = (35, 38] e TSPAN = (19, 72] e          TCLQ = (21, 23] e TWQM = (23, 30] e          ⋮          RCLQ = (1, 16] e RWMQ = (272, 532] então SP = PRESENT</i>
--

Figura 4.5 – 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 4.6.

*Se  $GEO = (6, 244]$  e  $TMNEL = (228, 1480]$  então  $SP = ABSENT$*

Figura 4.6 – 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 por:

$$p = \frac{1}{e^{-y} + 1} \quad (\text{eq. 4.5})$$

onde  $y = -(c_0 + c_1 \cdot d_1 + c_2 \cdot d_2 + \dots + c_n \cdot d_n)$  é uma equação obtida por análise de regressão linear múltipla.

Um exemplo deste tipo de regra é apresentado na Figura 4.7.

*Se  $0,1 - GEO * 0,1 + TMNEL * 0,3$  então  $SP = ABSENT$*

Figura 4.7 – Exemplo de Regra Logística

#### 4.3.2 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

4.8 foi codificado nos cromossomos apresentados na Tabela 4.1.

$r_1$ : Se  $TMIN = (5, 10]$  e  $TMED = (10, 22]$  e  $ELEV = (1, 2]$  então *PRESENT*  
 $r_2$ : Se  $TMIN = (0, 15]$  e  $TMED = (0, 50]$  e  $ELEV = (0, 20]$  então *ABSENT*  
 $r_3$ : Se  $TMIN * 0,80 + TMED * -0,2 + ELEV * 0,45$  então *ABSENT*

Figura 4.8 – Conjunto de regras  
 Fonte: SIQUEIRA (2005)

Tabela 4.1 – Cromossomos que codificam o conjunto de regras da Figura 4.8

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

#### 4.3.3 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 gerando filhos que carregam características dos cromossomos pais.

A Figura 4.9 mostra um exemplo de aplicação do operador cruzamento no GARP.

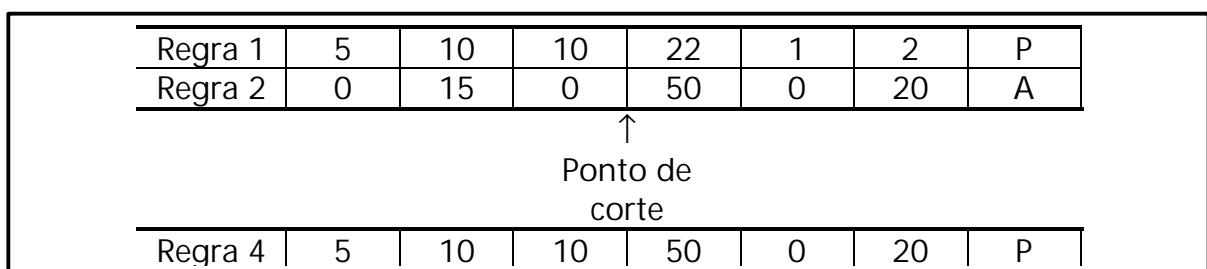


Figura 4.9 – Exemplo de operação cruzamento sobre as regras no GARP

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 4.10 mostra um exemplo do operador de junção.

Regra 1	5	10	10	22	1	2	P
Regra 2	0	15	0	50	0	20	A
			↑			↑	
		Início do corte			Fim do corte		
Regra 6	5	10	0	50	0	2	P

Figura 4.10 – 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 4.11 mostra um exemplo para cada uma destas mutações.

Regra 1	5	10	10	22	1	2	P
				↑			
Regra 7	5	10	10	50	1	2	P
Regra 2	0	15	0	50	0	20	A
						↑	
Regra 8	5	10	10	50	1	21	P

Figura 4.11 – 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*<sup>3</sup> ou regras por demais especializadas. A significância é medida através de um teste  $\chi^2$  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.

O processo de avaliação consiste em testar cada uma das regras utilizando um conjunto de dados de teste, previamente selecionado. O valor obtido pela eq. 4.6 será o valor de aptidão da regra.

<sup>3</sup> *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.

$$Sig = \frac{pXYs - no \cdot \frac{pYs}{n}}{\sqrt{no \cdot pYs \cdot \left(1 - \frac{pYs}{n}\right) / n}} \quad (\text{eq. 4.6})$$

onde:

- *Sig*: valor de aptidão da regra (significância);
- *pXYs*: número de pontos amostrados que a regra prevê corretamente;
- *no*: número de pontos amostrados avaliados pela regra;
- *pYs*: número de pontos amostrados com a mesma conclusão que a regra;
- *n*: 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 4.12 ilustra o processo de seleção, onde  $f(r)$  é equivalente ao valor obtido  $Sig$  pela eq. 4.6.

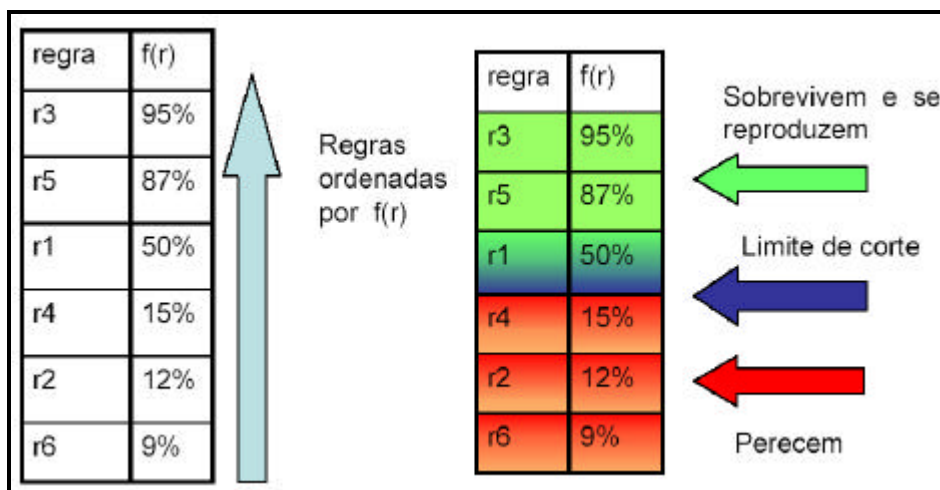


Figura 4.12 – O processo de seleção do GARP  
Fonte: SIQUEIRA (2005)

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.

#### 4.3.4 Avaliação da Qualidade dos Modelos Ajustados com o GARP

Existem, basicamente, duas estratégias para se avaliar a qualidade dos modelos ajustados. A primeira delas é coletar novas amostras e testa-las no modelo. A segunda é dividir as amostras iniciais em dois conjuntos: um conjunto para ajustar o modelo e outro para verificar o modelo.

A partir de um conjunto de amostras distinto daquele utilizado para ajustar o modelo é possível construir uma matriz de confusão. Esta matriz é utilizada para quantificar a qualidade do modelo ajustado. A Figura 4.13 mostra o formato da matriz de confusão.

Os valores A e D são predições corretas. B e C são considerados erros de predição. B é o erro por comissão que gera um falso positivo; C é o erro por omissão que gera um falso negativo.

	Presente	Ausente
Predição – Presente	A	B
Predição – Ausente	C	D

Figura 4.13 – Matriz de Confusão

Os erros do tipo B (comissão) não são considerados erros graves podendo ser causados por diversos fatores:

- a) A área é adequada à espécie mas não foi amostrada; a espécie pode ser encontrada na área;
- b) A área é adequada à espécie, mas fatores topológicos e/ou biológicos impedem que a espécie ocupe a área;
- c) A área é mesmo inadequada – caso de erro verdadeiro.

Os erros do tipo C (omissão) são considerados erros graves. Ou seja, um local onde se sabe que espécie é encontrada e está sendo predito como ausente.

Algoritmos estocásticos como o GARP produzem vários modelos com os mesmos dados de entrada. De posse destes vários modelos é possível calcular seus erros e

plotá-los num espaço omissão/comissão. Para uma espécie com um grande número de ocorrências a curva padrão desses erros é apresentada na Figura 4.14.

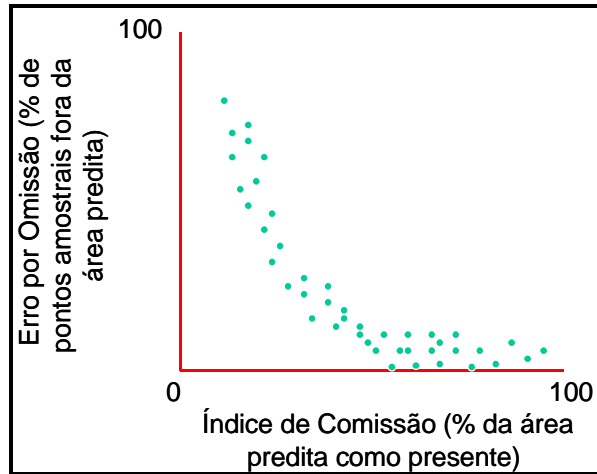


Figura 4.14 – Curva padrão para a relação Omissão/Comissão

A Figura 4.15 identifica no gráfico as regiões com relação a intensidade dos erros de omissão/comissão.

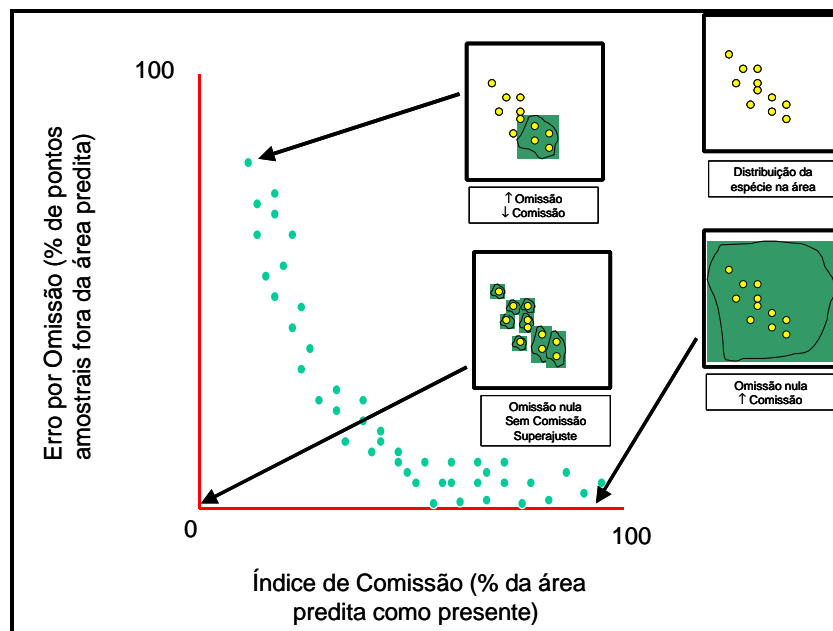


Figura 4.15 – Classificação das regiões quanto à intensidade dos erros de omissão/comissão

As regiões com altos erros da classe C (omissão) corresponde a modelos inadequados. A aplicação de um limiar para eliminar esses modelos deve ser então

aplicado. A Figura 4.16 mostra como seria a aplicação deste limiar.

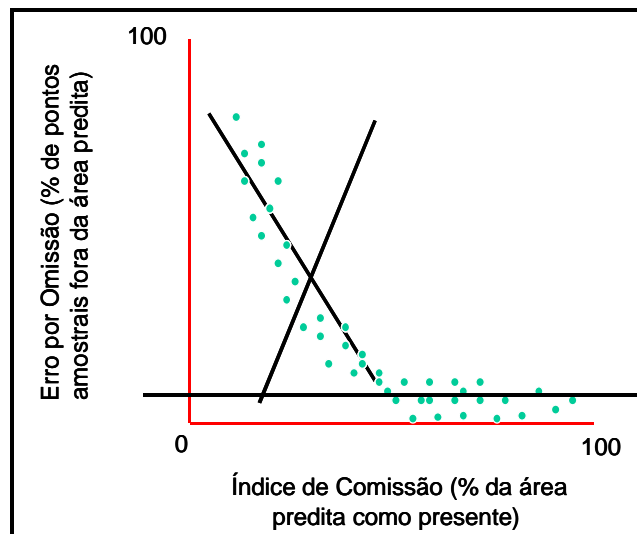


Figura 4.16 – Aplicação do limiar para eliminar modelos com muitos erros de omissão

Após a aplicação do limiar ainda restarão os modelos que apresentam superajuste e superpredição. Estes também não são considerados bons modelos. Assim, modelos na região da mediana entre estas classes são considerados os melhores modelos ajustados. Uma representação gráfica desta situação é apresentada na Figura 4.17.

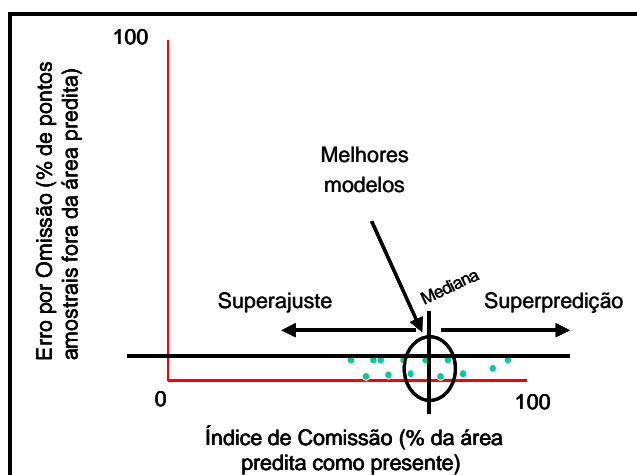


Figura 4.17 – Identificação da região com os melhores modelos ajustados

## 5 DETALHAMENTO DA PROPOSTA

Para atingir os objetivos apresentados nesta proposta, corroborando com as hipóteses levantadas, um estudo de caso será realizado. Neste estudo de caso um sistema para criação de modelos de distribuição de espécies será implementado.

Os sistemas de modelagem de distribuição de espécies, representados pela Figura 5.1, constroem regras de predição considerando apenas os pontos de ocorrência e as variáveis ambientais, negligenciando os efeitos da dependência espacial.

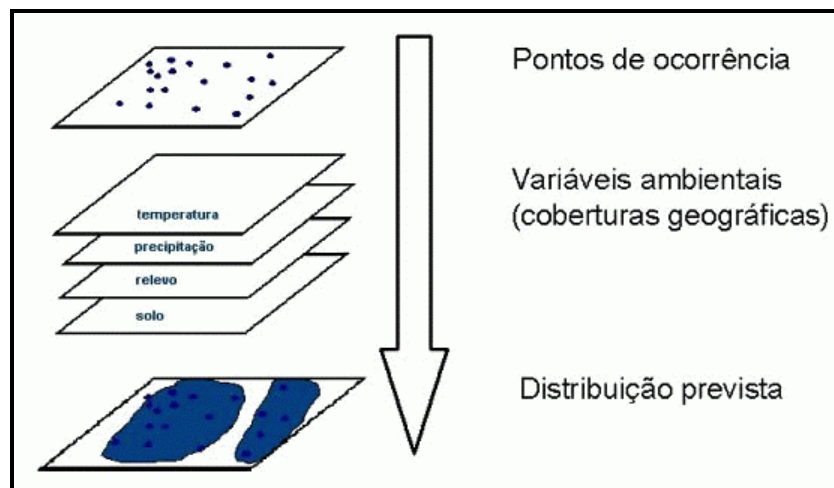


Figura 5.1 – Sistema de modelagem de distribuição de espécies

Para incorporar os efeitos desta componente, concebeu-se uma nova estrutura para um sistema de modelagem de distribuição de espécies. A Figura 5.2 apresenta esta nova estrutura.

### 5.1 AS INOVAÇÕES NA ESTRUTURA PROPOSTA

As inovações na estrutura proposta são a inserção de uma matriz de vizinhança generalizada e a associação de uma *layer* de pesos para cada uma das variáveis ambientais.

A presença da matriz de vizinhança generalizada cumpre um objetivo duplo: incorporar as relações espaciais no processo de modelagem e inserir o

conhecimento existente sobre elementos naturais, ou artificiais, presentes na região em estudo e que, sabidamente, interferem no fenômeno a ser modelado.

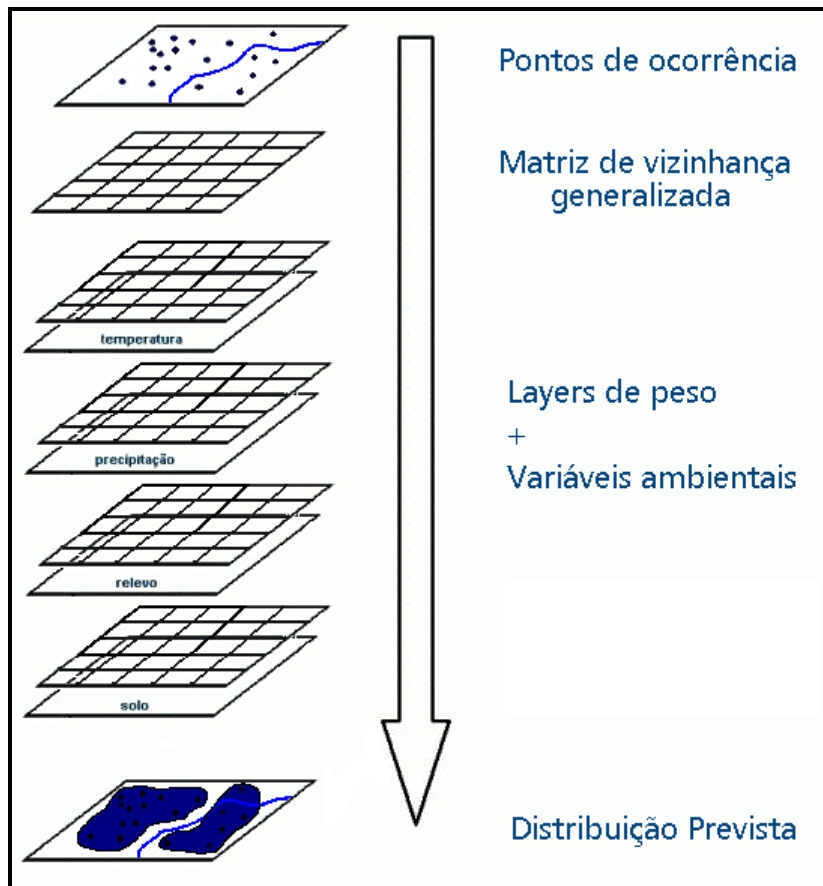


Figura 5.2 – Estrutura proposta para um novo sistema de modelagem de distribuição de espécies

A inserção do conhecimento existente sobre os elementos naturais está na manipulação dos pesos ( $W_{ij}$ ) que constituem a matriz de proximidade ( $V$ ) da matriz de vizinhança generalizada (maiores detalhes em 2.3).

A *layer* de pesos associada a cada uma das variáveis ambientais cumpre o objetivo de quantificar o efeito dos relacionamentos espaciais sobre as variáveis envolvidas. Os valores assumidos pela *layer* de pesos estará no intervalo [0; 1]. O cômputo da influência de cada variável ambiental não será pontual, mas sim de acordo com uma vizinhança local materializada numa submatriz de ordem 3. Os valores assumidos pela *layer* de pesos permitem avaliar a importância da variável ambiental associada bem como a localização espacial desta importância.

## 5.2 OBTENÇÃO DOS MODELOS DE DISTRIBUIÇÃO DE ESPÉCIES

Nesta proposta visualiza-se a construção de um AG com representação explícita de relacionamentos espaciais para modelagem de sistemas sócio-ambientais. Este AG será aplicado no desenvolvimento de um novo sistema para criação de modelos de distribuição de espécies. Estuda-se, ainda, a possibilidade de hibridizar o AG proposto com *simulated annealing*.

O mecanismo heurístico de busca será utilizado para encontrar os valores das *layers* de pesos, associadas às variáveis ambientais, que otimizam a capacidade preditiva do modelo de distribuição de espécies.

Para que o AG opere sobre os espaço de busca faz-se necessário codificar as informações manipuladas, no caso as *layers* de peso. Por se tratar de uma estrutura em grade, elas serão codificadas em matrizes bidimensionais reais, permitindo que os pesos assumam qualquer valor no intervalo [0; 1].

Para que um AG possa operar sobre matrizes existem duas representações possíveis: matriz linearizada e matriz bi-dimensional. A opção por uma delas afeta, particularmente, a operação de cruzamento efetuada no AG.

Ao linearizar uma matriz deve-se buscar um método que impeça a destruição dos esquemas presentes nos cromossomos. Um método de linearização utilizado com sucesso é a aplicação do esquema de indexação de Morton (BENNETT, WADE e ARMSTRONG, 1999).

Neste esquema o índice de cada célula é calculado como função da posição da linha e da coluna. O índice de Morton é construído intercalando-se os *bits* associados com a representação binária da linha e coluna que indexam cada célula (Figura 5.3). O resultado da aplicação do índice de Morton para uma matriz de ordem superior pode ser visualizado na Figura 5.4.

A linearização de uma matriz bidimensional facilita a implementação do operador de cruzamento, pois o espaço passa a ser representado num vetor unidimensional, que é a representação convencionalmente utilizada em AGs. Porém, a identificação das células adjacente a uma célula em particular é dificultada.

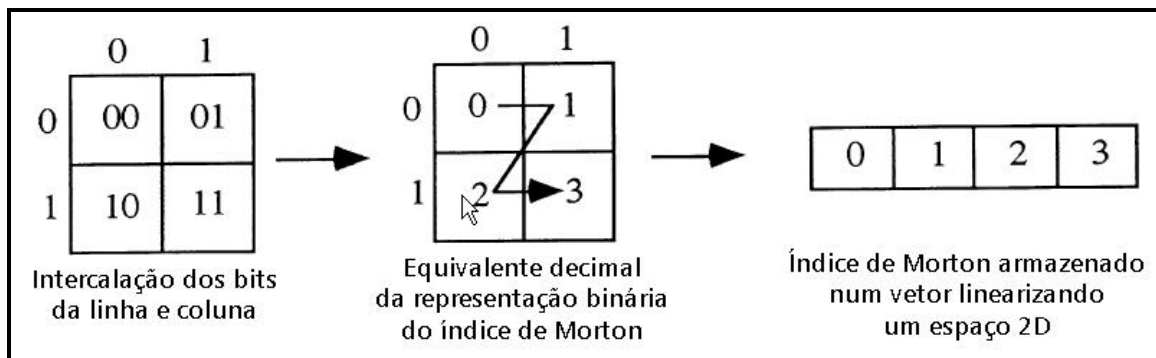


Figura 5.3 – Linearização do espaço utilizando o esquema de indexação de Morton

Fonte: Adaptado de BENNETT, WADE e ARMSTRONG (1999)

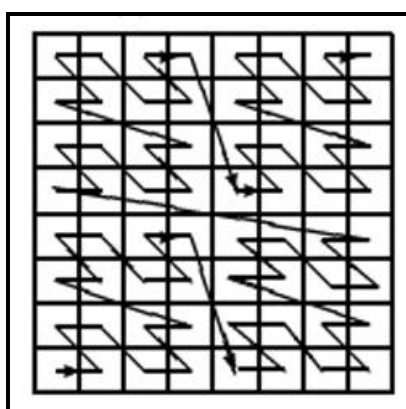


Figura 5.4 – Índice de Morton para uma matriz de ordem 8

A representação de uma matriz sob sua forma natural, bidimensional, facilita a identificação das células adjacentes, porém torna mais complexa a implementação do operador de cruzamento. Um operador de cruzamento, chamado UNBLOX, mostrou-se adequado para este tipo de representação. Este operador de cruzamento permite amostrar equilibradamente todas as áreas da matriz. (CARTWRIGHT e HARRIS, 1993)

A Tabela 5.1 exibe as regras que definem as regiões da matriz operadas pelo cruzamento UNBLOX. A Figura 5.5 exibe uma representação gráfica das regiões operadas pelo cruzamento UNBLOX.

A operação de mutação para matrizes não apresenta grande diferença com relação à operação de mutação sobre vetores. A lógica é a mesma; escolhe-se aleatoriamente uma célula da matriz e então substitui-se o valor ali armazenado por outro.

Tabela 5.1 – Regras que definem as regiões operadas pelo cruzamento UNBLOX

Caso	Pontos de cruzamento	Retângulo(s) (definidos pelos vértices abaixo) que são trocados entre cromossomos*
A	$x_1 < x_2$ $y_1 < y_2$	$(x_1, y_1)$ e $(x_2, y_2)$
B	$x_1 < x_2$ $y_1 > y_2$	$(x_1, y_1)$ e $(x_2, r)$ $(x_1, 1)$ e $(x_2, y_2)$
C	$x_1 > x_2$ $y_1 < y_2$	$(1, y_1)$ e $(x_2, y_2)$ $(x_1, y_1)$ e $(s, y_2)$
D	$x_1 > x_2$ $y_1 > y_2$	$(1, 1)$ e $(x_2, y_2)$ $(1, y_1)$ e $(x_2, r)$ $(x_1, 1)$ e $(s, y_2)$ $(x_1, y_1)$ e $(r, s)$

\* $(r \times s)$  é a ordem das matrizes manipuladas pelo operador

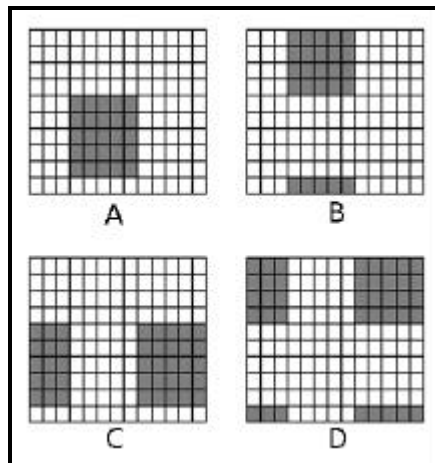


Figura 5.5 – Regiões operadas pelo cruzamento UNBLOX

Fonte: Adaptado de CARTWRIGHT e HARRIS (1993)

A seleção será realizada através do método da roleta, favorecendo aos indivíduos que apresentarem maior aptidão. A aptidão será medida através da função de avaliação, definida pela eq. 5.1.

$$Y = \sum \left( W_{ij} \cdot \left( \sum_{k=1}^n P_{xyk} \cdot V_k \right) \right) \quad (\text{eq 5.1})$$

onde:

- $W_{ij}$ : pesos da matriz de vizinhança generalizada;
- $i, j$ : índices da matriz de vizinhança generalizada.  $i = 1..r, j = 1..s$
- $P_{xyk}$ : pesos na *layer* associada à variável  $k$ ;
- $x, y$ : índices da submatriz de pesos na vizinhança  $i, j$  da variável  $k$ ;
- $V_k$ : variáveis ambientais;
- $n$ : número total de variáveis ambientais.
- $Y$ : Índice de aptidão bruto.

Y é dito índice de aptidão bruto pois deverá sofrer transformação, visando converter seu valor para um intervalo [0, 100] indicando a eficiência do indivíduo avaliado. Quanto maior este valor, maior a aptidão do indivíduo. Uma transformação que pode ser aplicada nesta conversão é aquela utilizada em regressão logística, apresentada na eq. 4.5.

Parâmetros do AG como tamanho da população, critério de parada, taxas de mutação e cruzamento serão definidos empiricamente, através da realização de testes com o sistema implementado.

A avaliação da qualidade dos modelos ajustados será efetuada através da construção da matriz de confusão.

Considerando que os AGs são algoritmos heurísticos estocásticos, ou seja, algoritmos capazes de fornecerem diversas soluções para um mesmo problema, há a possibilidade de implementar o algoritmo de seleção do melhor subconjunto solução, conforme descrito no item 4.3.4.

### 5.3 CRONOGRAMA

<b>Atividade</b>	<b>Jun Jul</b>	<b>Ago Set</b>	<b>Out Nov</b>	<b>Dez Jan</b>	<b>Fev Mar</b>	<b>Abr Mai</b>	<b>Jun Jul</b>	<b>Ago</b>
Definição do mecanismo de avaliação								
Artigo GeoInfo 2006								
Artigo Revista/Journal								
Implementação do sistema								
Avaliação do sistema								
Artigos								
Redação da Tese								
Defesa da Tese								

## 6 CONSIDERAÇÕES FINAIS

O uso de mecanismos semi-automáticos em geoinformática busca reduzir o esforço despendido na busca do conhecimento presente em conjuntos de dados. A estatística, a pesquisa operacional e o processamento de imagens, por exemplo, são áreas de conhecimento aplicadas neste processo. Outra área de conhecimento que tem sido aplicada no processo é a inteligência computacional; através da aplicação de sistemas especialistas, redes neurais, AGs e outros.

O GARP, sistema para construção de modelos de distribuição de espécies, demonstra que os AGs podem ser empregados com sucesso no processo. Entretanto, o GARP não considera um princípio elementar nos fenômenos espaciais: a dependência espacial. Essa negligência decorre da ausência de um AG que incorpore, em seus mecanismos evolutivos, os relacionamentos espaciais.

Outra limitação do GARP é ignorar o conhecimento existente sobre a região estudada; conhecimento como a presença de elementos naturais ou artificiais que interferem na distribuição da espécie.

Considerando estas limitações propôs-se uma nova estrutura para um AG. Um AG com representação explícita de relacionamentos espaciais que será utilizado na construção de um sistema de modelagem de distribuição de espécies. Neste novo AG as limitações anteriormente descritas são superadas, através da inserção de uma matriz de vizinhança generalizada e de *layers* de pesos associadas às variáveis ambientais e climáticas.

Uma informação relevante, a ser evidenciada, é a possibilidade de se utilizar este AG em outros problemas. Os AGs trabalham sobre soluções codificadas e usam uma função de avaliação para classificar estas soluções encontradas heurísticamente; a substituição da estrutura de codificação e da função de avaliação podem fazer com que outros problemas possam ser solucionados com a arquitetura proposta.

Ao término deste trabalho espera-se que o AG proposto seja utilizado, com sucesso, na construção de um sistema para criação de modelos de distribuição de espécies, cujos resultados validem as hipóteses descritas nesta proposta.



## REFERÊNCIAS BIBLIOGRÁFICAS

1. AGUIAR, A. P. D.; CÂMARA, G.; MONTEIRO, A. M. V.; SOUZA, R. C. M. Modelling spatial relations by generalized proximity matrices. In: SIMPÓSIO BRASILEIRO DE GEOINFORMÁTICA, 5., 2003, Campos do Jordão. **Proceedings...** Campos do Jordão, 2003. Disponível em: <<http://www.geoinfo.info/portuguese/geoinfo2003/papers/geoinfo2003-11.pdf>>. Acesso em: 01/05/2006.
2. AGUIAR, A. P. D. **Modelagem de mudanças de uso e cobertura do solo na Amazônia em múltiplas escalas**. 2004. 60 f. Proposta de tese (Doutorado em Sensoriamento Remoto) – Instituto Nacional de Pesquisas Espaciais, São José dos Campos.
3. ALMEIDA, C. M. **Modelagem da dinâmica espacial como uma ferramenta auxiliar ao planejamento: simulação de mudanças de uso da terra em áreas urbanas para as cidades de Bauru e Piracicaba (SP), Brasil**. 2003. 321 f. Tese (Doutorado em Sensoriamento Remoto) – Instituto Nacional de Pesquisas Espaciais, São José dos Campos.
4. ARAUJO, H. A. **Algoritmo simulated annealing: uma nova abordagem**. Florianópolis, 2001. 95f. Dissertação (Mestrado em Engenharia da Produção) – Universidade Federal de Santa Catarina, Florianópolis.
5. BENNETT, D. A.; ARMSTRONG, M. P.; WADE, G. A. Agent mediated consensus-building for environmental problems: a genetic algorithm approach. In: INTERNATIONAL CONFERENCE ON ENVIRONMENTAL MODELING AND GEOGRAPHIC INFORMATION SYSTEMS, 3., 1996, Santa Barbara. **Proceedings...** Santa Barbara, 1996. Disponível em: <[http://www.ncgia.ucsb.edu/conf/SANTA\\_FE\\_CD-ROM/sf\\_papers/bennett\\_david/my\\_paper.html](http://www.ncgia.ucsb.edu/conf/SANTA_FE_CD-ROM/sf_papers/bennett_david/my_paper.html)>. Acesso em: 01/05/2006.
6. BENNETT, D. A.; WADE, G. A.; ARMSTRONG, M. P. Exploring the solution space of semi-structured geographical problems using genetic algorithms. **Transactions in GIS**, Oxford, v. 3, n. 1, p.51-71, Jan. 1999.
7. BJORNSSON, C.; STRANGE, N. **Heuristic allocation of wetlands in GIS**. 2000. Disponível em: <<http://gis.esri.com/library/userconf/proc00/professional/papers/PAP238/p238.htm>>. Acesso em: 01/05/2006.
8. 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.
9. CARNEIRO, T. G. S. **Uma arquitetura para modelagem ambiental empírica e baseada nas teorias dos autômatos celulares, híbridos e**

- situados**. 2004. 52 f. Proposta de tese (Doutorado em Computação Aplicada) – Instituto Nacional de Pesquisas Espaciais, São José dos Campos.
10. CARTWRIGHT, H. M.; HARRIS, S. P. The application of the genetic algorithm to two-dimensional strings: the source apportionment problem. In: INTERNATIONAL CONFERENCE ON GENETIC ALGORITHMS, 5., 1993, Urbana-Champaign. **Proceedings...** San Mateo : Morgan Kaufmann, 1993. p. 631.
  11. CASTRO, J. P. **Um algoritmo evolucionário para geração de planos de rotas**. 1999. 91 f. Dissertação (Mestrado em Engenharia da Produção) – Universidade Federal de Santa Catarina, Florianópolis.
  12. DAVIS, L. Adapting operator probabilities in Genetic Algorithms. In: INTERNATIONAL CONFERENCE ON GENETIC ALGORITHMS, 3., 1989, Fairfax. **Proceedings...** San Mateo : Morgan Kaufmann, 1989. p. 61-69.
  13. DAVIS, L. **Handbook of Genetic Algorithms**. Reissue edition. Stamford : International Thomson Publishing, 1996.
  14. DRUCK, S.; CARVALHO, M. S.; CÂMARA, G.; MONTEIRO, A.V. M. Análise espacial e geoprocessamento. In: DRUCK, S. et al. **Análise Espacial de Dados Geográficos**. Brasília : EMBRAPA, 2004.
  15. ESHELMAN, L. J.; SHAFFER, D. J. Real-coded genetic algorithms and interval-schemata. In: WHITLEY, D. L. **Foundations of genetic algorithms-2**. San Mateo : Morgan Kaufman, 1993, p.187-202.
  16. FEITOSA, F. F. **Índices espaciais para mensurar a segregação residencial: o caso de São José dos Campos (SP)**. 2005. 151 f. Dissertação (Mestrado em Sensoriamento Remoto) – Instituto Nacional de Pesquisas Espaciais, São José dos Campos.
  17. FISCHER, M. M.; LEUNG, Y. A genetic-algorithms based evolutionary computational neural network for modelling spatial interaction data. In: CONGRESS OF EUROPEAN REGIONAL SCIENCE ASSOCIATION, 38., 1998, Vienna. **Proceedings...** Vienna : ERSA Papers, 1998. Disponível em: <<http://www.ersa.org/ersaconfs/ersa98/papers/478.pdf>>. Acesso em: 01/05/2006.
  18. GALVÃO, C. O.; VALENÇA, M. J. S. **Sistemas inteligentes: aplicações a recursos hídricos e sistemas ambientais**. Porto Alegre: Ed. Universidade/UFRGS/ABRH, 1999.
  19. GOLDBARG, M. C.; LUNA, H. P. L. **Otimização combinatória e programação linear: modelos e algoritmos**. Rio de Janeiro : Campus, 2000.
  20. GOLDBERG, D. E. **Genetic algorithms in search, optimization & machine learning**. Reading : Addison-Wesley, 1989.
  21. 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.
22. GUISAN, A.; THUILLER, W. Predicting species distribution: offering more than simple habitat models. **Ecology Letters**, v. 8, n. 9, p. 993-1009, Set. 2005.
23. 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.
24. HOLLAND, J. H. **Adaptation in natural and artificial systems**. Ann Arbor: University of Michigan Press, 1975.
25. KIRKPATRICK, S.; GELLAT, D. C.; VECCHI, M. P. Optimization by simulated annealing. **Science**. v. 220, n. 4598. p. 671-680, 1983.
26. LONGLEY, P. A. Foundations. In: LONGLEY, P. A. *et al.* **Geocomputation: A Primer**. New York : John Wiley & Sons, 1998.
27. LUCASIU, C. B.; KATEMAN, G. Applications of genetic algorithms in chemometrics. In: INTERNATIONAL CONFERENCE ON GENETIC ALGORITHMS, 3., 1989, Fairfax. **Proceedings...** San Mateo : Morgan Kaufmann, 1989. p. 170-176.
28. MATTHEWS, K. B.; CRAW, S.; ELDER, S.; SIBBALD, A. R. Evaluating multi-objective land use planning tools using soft systems methods. In: WORKSHOP OF THE UK PLANNING AND SCHEDULING SPECIAL INTEREST GROUP, 19., 2000, Milton Keynes. **Proceedings...** Milton Keynes : The Open University, 2000. p. 109-120.
29. MATTHEWS, K. B.; CRAW, S.; MACKENZIE, S. E.; SIBBALD, A. R. Applying genetic algorithms to land use planning. In: WORKSHOP OF THE UK PLANNING AND SCHEDULING SPECIAL INTEREST GROUP, 18., 1999, Salford. **Proceedings...** Salford : University of Salford, 1999. p. 109-115.
30. METROPOLIS, W.; ROSENBLUTH, A.; ROSENBLUTH, M.; TELLER, A.; TELLER, E. Equation of state calculations by fast computing machines. **Journal of chemical physics**. v. 21, n. 6, p. 1087-1092, Jun. 1953.
31. MICHALEWICZ, Z. **Genetic algorithms + data structures = evolution programs**. 3.ed. Berlin : Springer-Verlag, 1996.
32. NIX, H. A. A biogeographic analysis of Australian elapid snakes. In: R. Longmore. **Atlas of elapid snakes of Australia**. Canberra : Government Publishing Service, 1986, p.4-15.
33. OPENSHAW, S.; ABRAHART, R. J. Geocomputation. In: INTERNATIONAL CONFERENCE ON GEOCOMPUTATION, 1., 1996, Leeds. **Proceedings...** Leeds

: University of Leeds, 1996. p. 665-666.

34. OPENSHAW, S.; ABRAHART, R. J. **Geocomputation**. London : Taylor & Francis, 2000.
35. OPENSHAW, S.; OPENSHAW, C. **Artificial intelligence in geography**. West Sussex : John Wiley & Sons, 1997.
36. 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: 01/05/2006.
37. O'SULLIVAN, D. Toward micro-scale spatial modeling of gentrification. **Journal of Geographical Systems**, Berlin, v. 4, n. 3, p. 251-274, Out. 2002.
38. PAYNE, K; STOCKWELL, D. R. B. **GARP Modelling System User's Guide and Technical Reference**. 2001. Disponível em: <<http://biodi.sdsc.edu/Doc/GARP/Manual/manual.html>>. Acesso em: 01/05/2006.
39. PEDROSA, B. M. **Ambiente computacional para modelagem dinâmica**. 2003. 71 f. Tese (Doutorado em Computação Aplicada) – Instituto Nacional de Pesquisas Espaciais, São José dos Campos.
40. RONALD, S.; KIRKBY, S. Genetic algorithms for geographical boundary assignment. In: THE 1998 IEEE INTERNATIONAL CONFERENCE ON EVOLUTIONARY COMPUTATION, 1998, Anchorage. **Proceedings...** Piscataway : IEEE Press, 1998. p. 235-240.
41. RUDOLPH, G.; SPRAVE, J. A cellular genetic algorithm with self-adjusting acceptance threshold. In: IEE/IEEE INTERNATIONAL CONFERENCE ON GENETIC ALGORITHMS IN ENGINEERING SYSTEMS: INNOVATIONS AND APPLICATIONS, 1., 1995, Sheffield. **Proceedings...** London : IEE Press, 1995. p. 365-372.
42. SANTA CATARINA, A.; OLIVEIRA, J. R. F.; MONTEIRO, A. M. V. **Model Breeder: um algoritmo genético para criação de modelos**. 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: 01/05/2006.
43. SILVA, M. A. S. **Mapas auto-organizáveis na análise exploratória de dados geoespaciais multivariados**. 2004. 115 f. Dissertação (Mestrado em Computação Aplicada) – Instituto Nacional de Pesquisas Espaciais, São José dos Campos.
44. STOCKWELL, D.; PETERS, D. **Spatial predictions using Genetic Algorithm for Rule-set Production**. 1993. Disponível em: <[biodi.sdsc.edu/Symbiotik/Model/GARP/Doc/tutorial.html](http://biodi.sdsc.edu/Symbiotik/Model/GARP/Doc/tutorial.html)>. Acesso em: 01/05/2006.

45. STOCKWELL, D.; 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.
46. TOBLER, W. R. A computer model simulation of urban growth in the Detroit region. **Economic Geography**. v. 46, n. 2, p. 234-240, 1970.
47. XIAO, N. BENNETT, D. A.; ARMSTRONG, M. P. Solving spatio-temporal optimization problems with genetic algorithms: a case study of a bald cypress seed dispersal and establishment model. In: INTERNATIONAL CONFERENCE ON INTEGRATING GIS AND ENVIRONMENTAL MODELING (GIS/EM4): PROBLEMS, PROSPECTS AND RESEARCH NEEDS, 4., 2000, Banff, Canadá. **Proceedings...** Banff, 2000. Disponível em: <<http://www.colorado.edu/research/cires/banff/pubpapers/97/>>. Acesso em: 01/05/2006.
48. XIAO, N.; BENNETT, D. A.; ARMSTRONG, M. P. Using evolutionary algorithms to generate alternatives for multiobjective site-search problems. **Environment and Planning A**, London, v. 34, n. 4, p. 639-656, Abr. 2002.
49. YEPES, I. **Uma incursão aos algoritmos genéticos**. 2000. Disponível em: <<http://www.geocities.com/igoryepes/>>. Acesso em: 01/05/2006.