Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Otimização Multiobjetivo Aplicada ao Planejamento Sistemático de Conservação para Espécies de Plantas do Cerrado Brasileiro

Shana Schlottfeldt Santos

Brasília
2015

Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Otimização Multiobjetivo Aplicada ao Planejamento Sistemático de Conservação para Espécies de Plantas do Cerrado Brasileiro

Shana Schlottfeldt Santos

Tese apresentada como requisito parcial
para conclusão do Doutorado em Informática

Orientadora
Prof.ª Dr.ª Maria Emília M. Telles Walter

Coorientador
Prof. Dr. André Carlos P. de L. F. de Carvalho

Brasília
2015

Universidade de Brasília — UnB
Instituto de Ciências Exatas
Departamento de Ciência da Computação
Doutorado em Informática

Coordenadora: Prof.ª Dr.ª Alba Cristina M. Alves de Melo

Banca examinadora composta por:

Prof.ª Dr.ª Maria Emília M. Telles Walter (Orientadora) — CIC/UnB
Prof. Dr. André Carlos P. de L. F. de Carvalho (Coorientador) — SCC/USP
Prof. Dr. José Alexandre F. Diniz Filho — ECOEVOL/UFG
Prof.ª Dr.ª Mariana Pires de Campos Telles — ECOEVOL/UFG
Prof.ª Dr.ª Célia Ghedini Ralha — CIC/UnB

CIP — Catalogação Internacional na Publicação

Santos, Shana Schlottfeldt.

Otimização Multiobjetivo Aplicada ao Planejamento Sistemático de Conservação para Espécies de Plantas do Cerrado Brasileiro / Shana Schlottfeldt Santos. Brasília : UnB, 2015.
263 p. : il. ; 29,5 cm.

Tese (Doutorado) — Universidade de Brasília, Brasília, 2015.

1. otimização multiobjetivo, 2. planejamento sistemático de conservação, 3. conservação da biodiversidade, 4. variabilidade genética, 5. Ecoinformática

CDU 004.8

Endereço: Universidade de Brasília
Campus Universitário Darcy Ribeiro — Asa Norte
CEP 70910-900
Brasília–DF — Brasil

Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Otimização Multiobjetivo Aplicada ao Planejamento Sistemático de Conservação para Espécies de Plantas do Cerrado Brasileiro

Shana Schlottfeldt Santos

Tese apresentada como requisito parcial
para conclusão do Doutorado em Informática

Prof.ª Dr.ª Maria Emília M. Telles Walter
CIC/UnB

Prof. Dr. André Carlos P. de L. F. de Carvalho
SCC/USP

Prof. Dr. José Alexandre F. Diniz Filho
ECOEVOL/UFG

Prof.ª Dr.ª Mariana Pires de Campos Telles
ECOEVOL/UFG

Prof.ª Dr.ª Célia Ghedini Ralha
CIC/UnB

Prof.ª Dr.ª Alba Cristina M. Alves de Melo
Coordenadora do Programa de Pós-Graduação em Informática

Brasília, 24 de junho de 2015

Agradecimentos

O caminho que conduz à conclusão de um Doutorado pode ser imensamente gratificante e recompensador, mas ao mesmo tempo uma tarefa árdua. Eu não estaria na posição em que me encontro agora sem o suporte de diversas pessoas às quais dirijo meus mais sinceros agradecimentos, em especial:

A Maria Emília Machado Telles Walter, Mia, minha querida orientadora e amiga, por sua excelente orientação, encorajamento e conselhos ao longo dos anos, não só durante o Doutorado, mas desde minha graduação em Ciência da Computação.

A André Carlos Ponce de Leon Ferreira de Carvalho, meu co-orientador, que sempre demonstrou seu apoio encarando comigo, sem hesitação, diversos desafios.

A José Alexandre Felizola Diniz Filho e Mariana Pires de Campos Telles por sua colaboração, sempre pacientes com uma não-especialista em Ecologia fazendo perguntas para as quais ainda não existiam respostas.

A Yago Saez Achaerandio meu orientador de Mestrado na Universidade Carlos III de Madrid, Espanha, pela amizade e por ter me apresentado o fascinante mundo da Otimização.

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), pela oportunidade de aprofundamento teórico e desenvolvimento parcial de minha tese na Universidade de York, Reino Unido, por meio da concessão de uma bolsa de Doutorado Sanduíche, no âmbito do Programa Ciências sem Fronteiras.

A Jon Timmis que tão gentilmente me recebeu na Universidade de York, durante o período no exterior de meu Doutorado Sanduíche. Por me ensinar a apreciar o humor, o clima e o chá inglês. Foi um privilégio e um prazer tê-lo como meu orientador no exterior.

A Chris Mellor, do *Math Skills Centre* da Universidade de York, que tão gentilmente me tutoriou ao longo de diversas reuniões para discutir o tratamento estatístico dos dados desta tese.

Aos colegas da Universidade de York em especial os integrantes do *Intelligent Systems Group*, pela colaboração, pela troca de experiências e pelas enriquecedoras discussões.

A minha família, meus pais e meu querido irmão, por tudo o que fizeram para que eu chegasse onde estou, pelo amor e suporte incondicionais. Escrever uma tese é muito difícil, vocês tornaram essa tarefa muito mais leve.

Resumo

Nesta tese, propôs-se a aplicação de conceitos de Otimização Multiobjetivo (MOO) e de Computação Bioinspirada a problemas de Planejamento Sistemático de Conservação (SCP). Foram estudados três problemas específicos. No primeiro, buscou-se o menor conjunto de populações locais a serem conservadas para representar a diversidade genética de uma espécie vegetal do Cerrado. O método proposto foi capaz de identificar uma maior diversidade de soluções com a quantidade mínima de populações ao mesmo tempo em que refinou os resultados, indicando as combinações com maior diversidade intraespecífica e maior possibilidade de persistência ao longo do tempo. No segundo problema, buscou-se: (i) selecionar um conjunto de amostras geneticamente complementares a uma coleção de germoplasma de plantas já existente; (ii) definir uma *core collection* para uma coleção de germoplasma. Com a utilização de MOO foi possível identificar os indivíduos exatos que deveriam ser selecionados para complementar o germoplasma. Ademais, definiu-se um protocolo para tratar um grande volume de amostras a fim de estabelecer uma *core collection*. A abordagem proposta pode ser usada para construir *core collections* com máxima riqueza alélica, bem como ser estendido a casos de conservação *in situ*. Por fim, no terceiro problema, SCP foi associado à estimativa da ocorrência de espécies projetada para o futuro com base em simulações climáticas objetivando definir prioridades de conservação. O método proposto identificou locais com: (i) alta prioridade para conservação; (ii) risco significativo de investimento; e, (iii) que poderiam tornar-se atrativos no futuro. Foi proposto, também, um algoritmo multiobjetivo baseado em Sistemas Imunológicos Artificiais, o *Multi-Objective Artificial Immune System (MAIS)*. MOO permitiu trabalhar com instâncias de problemas com mais de duas dimensões, possibilitando maior confiabilidade na indicação do portfolio de soluções, aumentando, assim, o poder de decisão do método computacional e a qualidade da informação fornecida aos tomadores de decisão. O presente trabalho é pioneiro no país ao resolver problemas de SCP usando técnicas avançadas de otimização, colaborando para a implantação da área de Ecoinformática no Brasil.

Palavras-chave: otimização multiobjetivo, planejamento sistemático de conservação, conservação da biodiversidade, variabilidade genética, Ecoinformática, Cerrado.

Abstract

This thesis proposes a more sophisticated, yet general, solution to the systematic conservation planning problem (SCP) based on multi-objective optimization (MOO) and bio-inspired computing. We worked with three problems using data from plants of the Brazilian Cerrado biome. In the first problem, we looked for the smallest set of local populations of a plant species aiming its conservation. The method was able to find a larger portfolio of solutions and to refine the results as well, indicating solutions with more intra-specific diversity and higher probability of persistence throughout time. In the second problem, we aimed: (i) to select a set of individuals genetically complementary to an existing plant germplasm collection; and, (ii) to define a core collection for a germplasm collection. We were able to identify within a population of several individuals, the exact accessions/samples that should be chosen in order to preserve the species diversity. Moreover, we defined a method (a protocol) to deal with large amounts of accessions in the context of MOO. The proposed approach can be used to help constructing collections with maximal allelic richness and can also be extended to the *in situ* conservation. Finally, in the third problem, we applied MOO to SCP associated to climate forecasting, in a dynamic spatial prioritization analysis for biodiversity conservation. Our method was able to identify sites: (i) of high priority for conservation; (ii) with significant risk of investment; and, (iii) that may become attractive in the future. We also proposed a constrained multi-objective artificial immune system algorithm (MAIS). The MOO approach to SCP increases reliability by including additional objectives, which while increasing the complexity, significantly augments the amount and quality of information used to provide users with an improved decision support system. This thesis is pioneer in solving the SCP problem using advanced optimization techniques contributing to the insertion and consolidation of the new area of ecoinformatics in Brazil.

Keywords: multi-objective optimization, systematic conservation planning, biodiversity conservation, genetic variability, ecoinformatics, Brazilian Cerrado.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Algoritmos

Capítulo 1

Introdução

Segundo o Glossário de Termos para a *Terceira Reunião das Partes do Protocolo de Cartagena sobre Biossegurança e 8ª Conferência das Partes da Convenção sobre Diversidade Biológica* (MOP3/COP8) [66]:

> ***Biodiversidade**: é a variedade de vida na Terra. Constituída pelas variedades inter-espécies (sic), entre espécies e de ecossistemas. Também se refere às relações entre os seres vivos e o seu meio ambiente. Conjunto de plantas, animais, microrganismos e ecossistemas que sobrevivem na natureza – estimado em mais de 10 milhões de espécies. A biodiversidade inclui serviços ambientais responsáveis pela manutenção da vida na Terra, pela interação entre os seres vivos e pela oferta dos bens e serviços que sustentam as sociedades humanas e suas economias. Esses bens e serviços incluem alimentos, medicamentos, água e ar limpos, e outros recursos naturais que suportam a variedade de atividades humanas e indústrias.*

A conservação efetiva da biodiversidade é essencial para a sobrevivência humana e para a manutenção dos ecossistemas e tem se mostrado ao longo dos anos uma preocupação nos meios acadêmicos, mas com repercussão cada vez maior na política, na economia e mesmo na sociedade organizada [4, 28, 116, 219, 223, 230]. Entretanto, a biodiversidade não tem ainda a mesma visibilidade já alcançada por questões como a energia e as mudanças climáticas [186].

Estimativas recentes afirmam que o valor econômico e os benefícios advindos da conservação da biodiversidade de ecossistemas naturais podem variar entre 10 a 100 vezes o custo de sua manutenção [186].

O crescimento continuado das populações humanas e do consumo *per capita* tiveram como resultado uma exploração insustentável da biodiversidade em todo mundo, intensificada por alterações climáticas, pela acidificação dos oceanos, pela invasão de espécies alienígenas, bem como por outros impactos ambientais antrópicos, em especial a degradação, fragmentação e destruição de *habitats* [186].

Apesar do considerável volume de pesquisas examinando ameaças à diversidade biológica, poucos estudos estão focados na documentação de tais ameaças ou no projeto e implementação de intervenções [223].

A *Convenção sobre a Diversidade Biológica* (CDB) de 1992, um dos principais resultados da *Conferência das Nações Unidas para o Meio Ambiente e o Desenvolvimento* (CNUMAD) realizada no Rio de Janeiro, em junho de 1992 – mais conhecida como "Rio 92" – não conseguiu cumprir sua ambiciosa meta: desacelerar significativamente a perda de biodiversidade até 2010 [65].

Como resultado da CNUMAD realizada na cidade do Rio de Janeiro em 2012, 20 anos após a Rio 92 e por isso conhecida como "Rio+20", foi lançado, em 15 de junho de 2012 o "Plano Estratégico para Biodiversidade 2011-2020" (que teve como princípios as *Metas de Aichi para a Biodiversidade* [51], enunciadas como produto dos trabalhos da CDB que se reuniu em 2011, na cidade de Nagoya, Japão). Tal Plano Estratégico busca instituir ações concretas para deter a perda de biodiversidade no planeta, a fim de garantir que até 2020 os ecossistemas continuem resistentes, fornecendo serviços essenciais, garantindo a variedade do planeta e contribuindo para o bem-estar humano e a erradicação da pobreza [52].

O investimento nacional em conservação é pouco documentado, mas está aumentando e se diversificando. O investimento internacional em biodiversidade tem sido ampliado lentamente e estima-se que seu crescimento tenha sido da ordem de 38% em termos reais entre 1992 (quando a CDB ganhou força) e 2006, mas ainda assim, a soma envolvida ainda é extremamente pequena quando comparada aos valores gastos com subsídios de efeitos danosos ao meio-ambiente [186].

Abordagens de conservação bem-sucedidas precisam ser reforçadas e adequadamente financiadas. Além disso, tem-se sinalizado quanto à necessidade de reconhecer a biodiversidade como um bem público global, que deve ser integrado às políticas e estruturas de decisão para a produção de recursos e consumo, a fim de que sejam promovidas mudanças institucionais e sociais que permitam uma execução mais eficaz da política de conservação da biodiversidade.

Não obstante alguns êxitos de conservação (em especial em escala local) e o aumento do interesse público e governamental em investir em iniciativas sustentáveis, a perda da biodiversidade não tem diminuído. Para se ter uma idéia, apesar da extinção de espécies ser o resultado mais visível da redução da biodiversidade, estima-se que subpopulações são extintas três ordens de magnitude mais rápido do que espécies [186].

Desta forma, a crise de biodiversidade vem conduzindo cientistas em direção ao desenvolvimento de estratégias para efetivamente atingir metas de conservação [28, 171]. O princípio subjacente a tais estratégias encontra-se no chamado *Planejamento Sistemático de Conservação* (*Systematic Conservation Planning–SCP*), que envolve uma série de pas-

sos a serem seguidos a fim de se determinar o melhor custo-benefício de investimentos em ações de conservação.

Sob o ponto de vista computacional, no cerne do SCP está o Problema da Cobertura de Conjuntos, um problema clássico em Complexidade de Algoritmos que é NP-difícil [53].

De maneira simplificada, no âmbito do SCP, a forma usual de enunciar o problema é: selecionar um conjunto de locais (sítios), dentre vários disponíveis, minimizando o custo geral de conservação enquanto se maximiza a representação dos caracteres em estudo (i.e., Problema da Área Mínima ou da Cobertura Mínima de Conjuntos) [30]. É evidente que há dois objetivos conflitantes, os quais se busca otimizar, o que torna o problema SCP um candidato natural para a *Otimização Multiobjetivo* (*Multi-Objective Optimization–MOO*).

Diversos outros objetivos, tais como interesses sociopolíticos, podem ser incorporados ao problema SCP, o que adiciona mais dimensões ao problema que já é multiobjetivo em sua origem.

Como o SCP, muitos problemas do mundo real envolvem otimização simultânea de objetivos conflitantes. Nesse caso, não há uma solução ótima única, mas um conjunto de soluções que devem ser consideradas equivalentes na ausência de informação relativa à relevância de cada objetivo em relação aos demais [107]. A solução do problema dependerá da noção de equilíbrio utilizada para resolver os conflitos que surgem da consideração simultânea dos vários objetivos [103]. Tais soluções são ótimas no sentido que nenhuma solução no espaço de busca é melhor que outra quando *todos* os objetivos são considerados. São os chamados *Ótimos de Pareto* [246].

Em que pese a característica multiobjetivo do SCP, não raras vezes, os modelos de otimização restringem-se ao tratamento do problema como se monobjetivo fosse, por meio da atribuição de pesos diferenciados aos diferentes objetivos do problema a fim de agregá-los em uma única função objetivo ou de avaliação (mais conhecida como função *fitness*).

Há uma série de razões para a utilização da otimização multiobjetivo, dentre elas o fato de ser encontrado um conjunto de soluções em vez de somente uma (como seria o resultado encontrado pela aplicação de técnicas monobjetivo), mas também pela grande flexibilidade no tipo de dados e mesmo de restrições que podem ser integradas, enquanto mantém-se a tratabilidade do problema [248].

Um mesmo problema multiobjetivo pode ter infinitos Ótimos de Pareto não comparáveis entre si, tornando evidente a necessidade de critérios adicionais para se chegar a uma solução final. Tais critérios adicionais são fornecidos por um decisor, que procura selecionar uma solução que proporcione uma relação de compromisso entre os objetivos do problema [103].

Algoritmos Bioinspirados parecem ser particularmente apropriados para a tarefa de encontrar Ótimos de Pareto, pois muitos deles são capazes de processar um conjunto de

soluções em paralelo. Alguns pesquisadores sugerem que a pesquisa multiobjetivo e a otimização podem ser áreas onde Algoritmos Bioinspirados tenham melhor desempenho que outras estratégias [107, 226, 231, 248].

Como uma linha em pesquisa em Algoritmos Bioinspirados, tem-se destacado a investigação de *Algoritmos Evolutivos Multiobjetivos* (*Multiobjective Evolutionary Algorithms–MOEA*) [48] e de *Sistemas Imunológicos Artificiais* (*Artificial Immune Systems–AIS*) [73, 74, 82, 83].

O Sistema Imunológico é capaz de identificar a presença de agentes externos ao organismo, desencadeando uma série de processos que levam à neutralização e eliminação dos invasores de maneira adequada [218]. Baseadas em princípios imunológicos, novas técnicas computacionais têm sido desenvolvidas, objetivando não apenas uma melhor compreensão do sistema, mas a resolução de problemas.

Apesar da crescente pesquisa na área, até recentemente, foram poucos os esforços empreendidos em estender a aplicação de princípios de AIS no contexto da MOO, e as aplicações observadas são de âmbito genérico. Além disso, são poucos e ainda recentes os algoritmos que contam com um mecanismo explícito para tratar restrições baseadas em conceitos multiobjetivos [17, 114, 141, 160, 175, 233].

Resultados obtidos em estudo cotejando algoritmo AIS multiobjetivo comparado a algoritmos que representam o estado da arte em MOO [210] demonstraram que AIS multiobjetivo obteve soluções melhores não só em termos quantitativos (de otimização absoluta, i.e., os resultados encontram-se mais próximo à origem dos eixos), mas também em termos qualitativos na medida em que tais soluções se encontravam distribuídas de maneira mais regular no espaço de soluções. Em favor dos AIS, cumpre também destacar que Cutello, Nicosia e Pavone [61] mostraram que tal técnica – ainda que na forma monobjetivo – apresentou melhores resultados para o problema NP-difícil da coloração de vértices (que pode ser reduzido ao problema da Cobertura Mínima de Conjuntos) quando comparado a algoritmos que representam o estado da arte em otimização.

Voltando ao problema SCP, quando se fala em unidades de conservação no Brasil, historicamente, não se tem seguido, na maioria dos casos, uma lógica fundamentada em aspectos técnico-científicos, principalmente em função de questões de conflitos pelo território que terminam por suplantar quase toda e qualquer justificativa técnica [102]. Apesar de tais questões sociopolíticas serem consideradas inerentes ao processo de criação, para que um sistema de unidades de conservação efetivamente conserve a biodiversidade, o estabelecimento de uma rede ideal de áreas protegidas deveria ser orientado por princípios da biologia da conservação [127].

Há regiões significativas que são ambientalmente sensíveis e que estão bastante vulneráveis, em especial onde ocorrem solos, substratos geológicos e relevos suscetíveis à erosão

hídrica, à desertificação antrópica, à contaminação, à compactação, à perda de fertilidade, à perda de biotas e de processos ecológicos [190].

No que diz respeito, em especial, aos biomas do Cerrado e do Pantanal, notadamente nas últimas quatro décadas, ambos vêm sofrendo forte pressão, o que já resultou em áreas degradadas na ordem de 50% e 12%, respectivamente [190].

O Cerrado abrange, hoje, cerca de 2.036.448 km^2, área equivalente a quase um quarto (23,92%) do território nacional [125], dos quais 57% estão no Centro-Oeste [126]. Hospeda rico patrimônio de recursos naturais e exibe uma das maiores riquezas biológicas mundiais. É um bioma rico e globalmente significativo por sua extensão, diversidade ecológica, estoques de carbono e função hidrológica no continente sul-americano [33]. No Cerrado, são conhecidas aproximadamente 6 mil espécies de árvores, 800 espécies de aves, além de 780 das 3 mil espécies de peixes da América do Sul, entretanto, extensa parte da biodiversidade do Cerrado permanece desconhecida, tornando-o uma das 25 áreas do mundo consideradas críticas (*hotspots*) para a conservação [167, 190].

Apesar disso, trata-se de um bioma profundamente ameaçado pelo avanço da fronteira agrícola. O estágio de conservação de áreas de Cerrado é pouco expressivo, enquanto o avanço da fronteira agrícola se dá de forma rápida e desordenada. Dados de 2004, apontavam que cerca de 55% da área original do Cerrado, o que correspondia a aproximadamente 880.000 km^2 – superior à área desmatada da amazônia brasileira –, já havia sido desmatada ou transformada pela ação humana [152]. Em especial, dados mostram que a taxa de desmatamento no Cerrado, em que pese ter sofrido um leve declínio entre 2004–2009, voltou a crescer acentuadamente, em especial a partir de 2010, tendo alcançado em 2012 uma taxa de 7.652 km^2/ano [217]. Mais preocupantes são os resultados de análises que sugerem que o novo Código Florestal[1] [24] permitirá desmatamento adicional do Cerrado [217].

Já a planície do Pantanal abrange 151.199 km^2, sendo caracterizada por inundações sazonais e influência florística da Amazônia, do Chaco, da Floresta Atlântica e do Cerrado. Em que pese não apresentar endemismos (grupos taxonômicos que se desenvolveram numa região restrita), a riqueza de espécies animais por quilômetro quadrado tende a ser maior no Pantanal do que em florestas como a Amazônia e a Mata Atlântica. A importância do Pantanal para a conservação da biodiversidade global evidencia-se pelas denominações por meio das quais é conhecido [190]: Reserva da Biosfera (nomenclatura que lhe foi atribuída pela Unesco) e Área Úmida de Importância Internacional (desde 1993, segundo critérios da convenção de Ramsar [224]).

[1] Lei nº 12.651, de 25 de maio de 2012, que dispõe sobre a proteção da vegetação nativa, estabelecendo normas gerais sobre a proteção da vegetação, áreas de Preservação Permanente e as áreas de Reserva Legal; a exploração florestal, o suprimento de matéria-prima florestal, o controle da origem dos produtos florestais e o controle e prevenção dos incêndios florestais, e prevê instrumentos econômicos e financeiros para o alcance de seus objetivos.

Consideradas como principais regiões de atuação das *Instituições de Educação Superior (IES)* da Região Centro-Oeste, o Cerrado e o Pantanal têm sido alvo de inúmeros projetos multidisciplinares desenvolvidos por universidades e instituições de pesquisa da região. Entretanto, persistem lacunas consideráveis na produção de conhecimento voltado ao manejo adequado e à utilização dos recursos naturais visando o desenvolvimento racional e sustentável para as futuras gerações [22, 153, 193].

As espécies alimentícias nativas do Cerrado ou do Pantanal pertencem a diversos gêneros e famílias de interesse para subsistência, comercialização e industrialização [22]. Muitas são comercializadas com grande aceitação popular em feiras. Sem embargo, tais espécies estão perdendo espaço em consequência da forte ação antrópica – nela incluída a expansão da produção agrícola da Região Centro-Oeste, com uma estrutura de agronegócios altamente competitiva em níveis nacional e internacional –, antes mesmo que suas características e potencialidades sejam conhecidas com profundidade [173, 190].

Neste contexto, assume notável relevância a implementação da *Rede Centro-Oeste de Pós-Graduação, Pesquisa e Inovação (Rede PRO-CENTRO-OESTE)*, que visa acelerar o processo de geração de conhecimentos, tecnologias, inovações, produtos e serviços que viabilizem um salto qualitativo e competitivo na agregação de valor aos recursos naturais, potencializando seu aproveitamento, bem como sua conservação [190].

A Rede PRO-CENTRO-OESTE, foi instituída por meio da Portaria Interministerial nº 1.038, de 10 de dezembro de 2009, assinada pelo Ministério da Ciência e Tecnologia (MCT) e pelo Ministério da Educação (MEC). Congrega instituições de ensino e pesquisa dos estados de Goiás, Mato Grosso, Mato Grosso do Sul e do Distrito Federal (Figura 1.1), suas respectivas Secretarias de Estado de Ciência e Tecnologia e Fundações de Amparo à Pesquisa (FAPs), visando à formação de recursos humanos e à produção de conhecimento científico, tecnológico e de inovação que contribuam para o desenvolvimento sustentável da Região Centro-Oeste.

A Rede PRO-CENTRO-OESTE foi instituída para trabalhar em duas frentes:

1. produzir conhecimento, com vistas à conservação e ao uso sustentável dos recursos naturais do Cerrado e do Pantanal; e,
2. fortalecer e consolidar a formação de recursos humanos para o desenvolvimento sustentável da Região Centro-Oeste.

A Rede PRO-CENTRO-OESTE apresenta as seguintes diretrizes:

1. apoiar a pesquisa em biotecnologia e biodiversidade na Região Centro-Oeste;
2. consolidar e integrar grupos de pesquisa da região;
3. formar, atrair e fixar doutores na região;

Figura 1.1: Estados que formam a Rede PRO-CENTRO-OESTE [190].

4. fortalecer e contribuir para a consolidação de programas de pós-graduação;
5. instituir um programa de doutorado de caráter multi-institucional;
6. contribuir para o desenvolvimento sustentável da região.

Como primeira ação da Rede, foi lançado o Edital MCT/CNPq/FNDCT/FAPs/MEC/CAPES/PRO-CENTRO-OESTE Nº 031/2010, contemplando três linhas de pesquisa:

1. Linha 1: Ciência, Tecnologia e Inovação para Sustentabilidade da Região Centro-Oeste.
2. Linha 2: Bioeconomia e Conservação dos Recursos Naturais.
3. Linha 3: Desenvolvimento de Produtos, Processos e Serviços Biotecnológicos.

A *Rede 9–Genética Geográfica e Planejamento Regional para Conservação de Recursos Naturais no Cerrado (GENPAC)*, no contexto da qual este trabalho se insere, é uma das 16 redes de pesquisa que integram a Rede PRO-CENTRO-OESTE.

A rede GENPAC tem como objetivo integrar grupos de pesquisa de diferentes instituições da região Centro-Oeste do Brasil que estudam macroecologia, filogeografia, genética de populações e genética molecular em espécies do Cerrado, buscando caracterizar padrões de variabilidade genética em tais espécies, a fim de compreender os processos ecológicos e biogeográficos de origem e manutenção da biodiversidade e, a partir desse conhecimento, desenvolver estratégias que conciliem desenvolvimento econômico e conservação dos recursos naturais [113].

Em suas análises genético-moleculares, a GENPAC tem como foco algumas *espécies-alvo*, e inova em unir análises moleculares sofisticadas a padrões macroecológicos, com

vistas à elaboração de estratégias mais eficientes de conservação e manejo das espécies-alvo (em especial plantas frutíferas de importância econômica) e da biodiversidade como um todo [113].

SCP tem sido aplicado em nível de espécies [27, 138, 180], ou mesmo em níveis hierarquicamente mais altos (tipos de vegetação [117]; *taxa* [133]; comunidades [154]; *habitats* [158]; ecossistemas [27]; ecorregiões [150] ou outra classificação espacial; processos ecológicos [27]). Quando se fala em nível mais baixo, a análise, em geral, não ultrapassa o nível de algumas características fenotípicas, e em especial, está ligada a análises filogenéticas [57, 123, 193].

Entretanto, é possível aplicar os mesmos métodos de SCP numa granularidade menor (maior detalhamento), por exemplo, utilizando informação alélica como unidade básica de análise, no contexto do novo campo de conservação genética inaugurado por Diniz-Filho et al. [92, 94], permitindo, com isso, a definição de estratégias em um nível intra-específico.

Em 1999, Prendergast apontava que, no que dizia respeito ao SCP, um refinamento para a abordagem da riqueza das espécies poderia ser obtido pela quantificação genética da diversidade [178], associando uma medida de quão diferentes indivíduos de uma mesma espécie seriam entre si. Entretanto, fazia a ressalva que tal abordagem requereria um nível de recursos e técnica que a tornaria impraticável à época. Porém, o que antes era um limitante impeditivo, hoje já não o é.

Técnicas moleculares permitiram distinguir a Estrutura Genética Espacial (*Spatial Genetic Structure–SGS*) – que corresponde à distribuição não-aleatória de genótipos dentro de uma população como o resultado do fluxo gênico, da dispersão e da seleção – dentro e entre populações [50]. Assim, por meio de tais métodos, a estrutura genética tradicionalmente avaliada somente em grandes escalas espaciais, entre populações, passou também a ser feita em escala local [215].

A SGS é afetada por processos ecológicos e características da história de vida de um organismo. Num contexto de conservação biológica mais explícito, fragmentação e perturbações de *habitats* podem mudar a SGS devido às alterações que podem provocar em processos ecológicos [50]. O monitoramento de parâmetros genéticos e demográficos fornece indícios acerca do tamanho efetivo de populações e fluxo de genes que assumem papel relevante em programas de gestão e conservação de recursos [50].

É possível usar a informação gerada para estabelecer melhores planos de gestão das populações naturais ou mesmo delinear estratégias ótimas de amostragem para coleções de germoplasmas.

No cerne deste problema encontra-se a ideia geral de priorização de conservação intra-específica, amplamente discutida no início dos anos 1990 no contexto da definição de *Unidades Significativa de Evolução* (*Evolutionary Significant Units–ESU*) e *Unidades de*

Gestão (*Management Units–MU*). Embora a definição de ESUs e MUs forneça unidades intra-específicas (como subespécies ou variedades locais), que poderiam ser usadas como objetos de conservação, a definição não é capaz de lidar com a variação genética contínua em nível de espécie [93]. Além disso, não fornece uma maneira de estabelecer quais unidades ou componentes de variação intra-específica devem ser priorizadas no contexto do SCP.

1.1 Motivação

Diniz-Filho et al. [92] propuseram, recentemente, uma abordagem para complementaridade[2] que poderia ser utilizada para otimizar a conservação da variabilidade genética, expressa como variação alélica (presença-ausência de alelos) derivada de marcadores de *loci* de microssatélites, e resolveram o problema utilizando *simulated annealing*. Essa abordagem, desenvolvida no contexto da cobertura mínima de conjuntos, foi aplicada à conservação *in situ*[3] e *ex situ*[4] de *Dipteryx alata* (baru, uma árvore do Cerrado). Na abordagem *in situ*, a idéia foi selecionar o menor número de populações locais que deveriam ser preservadas para criar uma rede ou portfólio de populações locais que representassem todos os alelos. Na abordagem *ex situ*, buscou-se encontrar o menor número de populações locais que deveriam ser amostradas para complementar um banco de germoplasma existente.

Essas tentativas usaram apenas informação relativa à presença-ausência de alelos nas populações locais, o que não é tão informativo quanto a utilização direta das frequências dos alelos, que reflete de um modo bastante adequado os processos ecológicos e evolutivos que conduzem a diversidade genética em populações locais e pode estar mais relacionada à persistência da população. Tal medida seria equivalente à utilização de características mais complexas relacionadas às espécies (e.g., abundância e adequabilidade do ambiente) quando da aplicação de SCP em níveis hierárquicos superiores, melhorando potencialmente a persistência a longo prazo nas áreas de conservação.

Numa aplicação padrão de SCP usando os *softwares* C-Plan ou MARXAN, baseados em *simulated annealing* (v. [92]), as restrições são geralmente expressas como um peso, obtido por uma função complexa resultante da combinação de diversos atributos conflitantes (e.g., formas potenciais de uso da terra). No entanto, vários problemas do mundo real envolvem otimização simultânea de múltiplos objetivos conflitantes, que devem ser

[2]Para definição de *complementaridade*, *vide* Seção 2.1.

[3]na qual busca-se a conservação de ecossitemas e *habitats* naturais bem como a manutenção e recuperação de populações viáveis de espécies em seu meio natural e, no caso de espécies domesticadas ou cultivadas, nos arredores onde as mesmas desenvolveram suas propriedades distintivas [120].

[4]conservação de germoplasma fora de seu *habitat* natural.

analisados como dimensões independentes e não combinados em uma única função ponderada.

Devido à simplicidade computacional das soluções para SCP anteriormente desenvolvidas em nível de espécies (e.g., [92]), propõe-se, aqui, uma solução mais complexa e geral para SCP, baseada em MOO, o que permite tratar simultaneamente mais do que dois objetivos pela inclusão de objetivos adicionais, acrescentando, com isso, mais complexidade ao problema, mas ao mesmo tempo garantindo maior flexibilidade e informação aos *tomadores de decisão* (*decision-makers–DM*).

Associada à proposta de utilização da abordagem MOO, nesta tese propõe-se também um algoritmo AIS multiobjetivo para lidar com SCP, o *Multi-Objective Artificial Immune System Algorithm (MAIS)*. Tal proposta foi motivada por resultados obtidos em estudo cotejando este tipo de algoritmo comparado a algoritmos que representam o estado da arte em MOO para resolver o problema de cobertura de antenas [210], tal estudo demonstrou que AIS obteve soluções melhores não só em termos quantitativos (de otimização absoluta, i.e., os resultados estavam mais próximo à origem dos eixos), mas também em termos qualitativos (na medida em que tais soluções encontravam-se distribuídas de maneira mais regular no espaço de soluções).

1.2 Problema

Inicialmente foram estabelecidos dois problemas principais a serem trabalhados (Problemas 1 e 2). Diante dos resultados positivos obtidos ainda nos experimentos iniciais, estendeu-se os experimentos para contemplar um terceiro problema.

Problema 1: Seleção de Populações

Encontrar soluções ótimas indicando o menor conjunto de populações locais de uma dada espécie que devem ser conservadas para representar a diversidade genética da espécie, objetivando a conservação *in situ* da espécie.

Problema 2: Seleção de Indivíduos

Em vez de tratar as populações como uma unidade, o problema deveria ser resolvido em nível de indivíduos e seria formulado da seguinte maneira: maximizar o número de alelos, minimizando o número de indivíduos necessários para representar todos os alelos, ao mesmo tempo que se maximiza a heterozigose do conjunto de indivíduos final.

Problema 3: Priorização Espacial de Conservação Associada a Mudanças Climáticas

Por meio da comparação entre resultados obtidos pela aplicação de SCP tanto a dados de distribuição de plantas do Cerrado no presente quanto a dados resultantes de simulações que procuram prever a provável distribuição das espécies no futuro, buscou-se identificar regiões: 1) com alta prioridade de conservação; 2) com alto risco de investimento; e, 3) que poderiam tornar-se opções atrativas no futuro.

1.3 Objetivos

Principal

Propõe-se aplicar conceitos de MOO ao problema SCP, o que permite trabalhar com uma instância do problema SCP com mais de duas dimensões, possibilitando maior flexibilidade pela inclusão de objetivos adicionais, bem como acrescentando mais complexidade e aumentando, assim, o poder de decisão do método computacional.

Específicos

Mais especificamente, propõe-se:

- investigar um algoritmo inspirado em AIS para encontrar o menor conjunto de populações locais (Problema 1) ou indivíduos (Problema 2) de uma espécie do Cerrado brasileiro que devem ser conservados para representar a diversidade genética da espécie, utilizando informação alélica proveniente de análise molecular em nível populacional como unidade básica de investigação;
- estender a utilização do método proposto, incorporando análise dinâmica de biodiversidade para prover os DM com informação acerca da projeção das decisões atuais de conservação face a cenários futuros de mudança climática, possibilitando rever tais decisões no presente com base em uma decisão informada (Problema 3);
- comparar os resultados com outros algoritmos encontrados na literatura, a fim de verificar a efetividade do método para os problemas propostos.

1.4 Descrição dos Capítulos

Esta tese está estruturada em nove capítulos, organizados da seguinte forma:

Capítulo 2 [Planejamento Sistemático de Conservação]: apresenta conceitos básicos relativos ao SCP, formula o problema SCP e expõe as abordagens até então utilizadas para tratá-lo.

Capítulo 3 [Otimização Multiobjetivo]: introduz noções básicas usadas em Otimização Multiobjetivo.

Capítulo 4 [Inspiração Biológica: EA e AIS]: aponta conceitos básicos de Algoritmos Evolutivos (EA) e Sistemas Imunológicos Artificiais (AIS). Em especial, quanto aos EA, aborda os algoritmos NSGA-II e SPEA2, usados como *baseline* de comparação do algoritmo proposto nesta tese.

Capítulo 5 [MAIS: Multi-Objective Artificial Immune System Algorithm]: apresenta, em detalhes, o algoritmo proposto nesta tese.

Capítulo 6 [Estudo de Caso]: discorre acerca dos elementos necessários à análise e discussão do experimento final desenvolvido no âmbito desta tese. Apresenta os dados utilizados, material e métodos.

Capítulo 7 [Calibragem de Parâmetros e Avaliação de Desempenho do MAIS]: apresenta e discute os resultados encontrados no experimento final, realizado com dados de *D. alata* e utilizando MAIS.

Capítulo 8 [Trabalhos Publicados]: traz a coletâneas dos trabalhos publicados e aceitos para publicação, resultantes da pesquisa desenvolvida no âmbito desta tese.

Capítulo 9 [Conclusões e Trabalhos Futuros]: trata das contribuições e sugere oportunidade de trabalhos futuros.

Capítulo 2

Planejamento Sistemático de Conservação

O *Planejamento Sistemático de Conservação (Systematic Conservation Planning–SCP)* busca estabelecer, de maneira clara, quais objetos de conservação (espécies, ecossistemas, processos ecológicos) são significativos, quais são as metas de conservação almejadas, e qual a área mínima necessária para que estes objetos persistam ao longo do tempo. Outrossim, procura: (1) proteger uma amostra representativa da biodiversidade regional de maneira inteligente; (2) identificar áreas prioritárias para conservação, considerando os eventuais conflitos no uso da terra; e, (3) dados os limitados recursos disponíveis para conservação, apontar uma solução de consenso que corresponda à melhor relação custo/benefício. Em suma, busca proteger o máximo de objetos de conservação com o mínimo de investimento/área protegida [196].

O grande objetivo do SCP é transformar conhecimento técnico e científico em ações de planejamento e, estas, em medidas concretas de manejo, proteção, conservação e uso sustentável do patrimônio natural.

Neste capítulo, é feita uma introdução ao SCP apresentando suas características gerais (Seção 2.1), bem como a formulação do Problema da Mínima Representação que está no âmago do SCP (Seção 2.2). As principais abordagens até então utilizadas para tratar SCP, os algoritmos desenvolvidos e as ferramentas de suporte à decisão incorporando tais algoritmos são descritos na Seção 2.3.

2.1 Conceitos Básicos

É uma prática selecionar áreas para preservação de recursos naturais que, cumprindo definições de áreas de proteção integral, são chamadas de *reservas* [97].

O papel básico das reservas consiste em aumentar a eficácia da conservação *in situ* da biodiversidade [97], separando elementos de biodiversidade dos processos que ameaçam sua existência na natureza, dentro das limitações impostas pelo grande e rápido crescimento da ocupação humana ao redor do mundo e sua concomitante necessidade por espaço, materiais e eliminação de resíduos [15, 155].

Projetar uma reserva não é um processo de acumular a máxima extensão de terra possível, mas fazê-lo de maneira eficiente, dadas as restrições existentes [177]. Além disso, áreas protegidas devem ser geridas mais como uma rede coerente de preservação do que como ilhas isoladas de algum *habitat* [186].

Paradigmas de conservação, práticas e políticas têm mudado ao longo dos anos e têm tido sucesso variável. Nas últimas décadas, abordagens tradicionais, como a criação de parques nacionais, evoluíram para contemplar a conscientização a respeito dos diversos benefícios fornecidos pelas áreas protegidas, a importância de iniciativas de conservação local e interesses na gestão de áreas protegidas, bem como a necessidade de contemplar os custos de oportunidade da conservação em relação a pessoas do meio rural em condição de vulnerabilidade social [186].

A criação de reservas muitas vezes foi feita de maneira *ad hoc* [183]. Em alguns casos, territórios foram convertidos em áreas de proteção não devido a sua biodiversidade, mas por seu valor recreacional, à beleza de sua paisagem ou mesmo porque a região era muito remota ou improdutiva para ter importância econômica, assim, não raras vezes, a seleção de áreas protegidas tendeu para paisagens economicamente marginais que levaram a severa sub-representação de espécies, *habitats* e ecossistemas [223]. Além disso, as reservas existentes ocasionalmente são muito pequenas para suportar populações viáveis de uma gama ampla de espécies. Assim, apesar dos esforços de conservação a perda de diversidade continua acelerando [230, 238].

Desta forma, não faltam casos de unidades de conservação constituídas em áreas inadequadas, adotando-se critérios que são antes políticos e econômicos do que imperativamente científicos [18, 118, 155, 177].

Um alvo ou objeto de conservação corresponde a todo e qualquer elemento que se deseje conservar em determinada região, incluindo espécies (ou qualquer outro *taxon*), populações, comunidades, ecossistemas, processos ecológicos, etc., desde que seja possível mapeá-lo [127].

A escolha dos objetos depende [127]:

1. do objetivo do trabalho (conservação e/ou recuperação e/ou indicação de áreas sensíveis);

2. da área escolhida (extensão, localização); e

3. da disponibilidade de dados (bases cartográficas de qualidade e de escala compatível, existência e disponibilidade de dados biológicos).

Idealmente, o objeto deve ter a distribuição espacial bem conhecida, ser representativo da biodiversidade local e/ou ser vulnerável pela raridade, endemicidade e perda de *habitat*, entre outros.

Em geral, são escolhidas para conservação espécies chamadas *guarda-chuva* [147] – avaliadas como indicadoras de qualidade ambiental, raras e/ou com algum grau de ameaça de extinção –, pois são consideradas como bons substitutos (*surrogates*) da biodiversidade, isto é, o território onde ocorrem geralmente apresenta riqueza de recursos e endemismo, muitas vezes considerados prioritários em ações de conservação [150, 236], mas isso não garante que conservarão adequadamente a *biota* (conjunto de seres vivos de um ecossistema) regional [9, 177].

O SCP como vem sendo tratado ao longo dos anos, busca definir, de maneira clara, quais objetos de conservação são relevantes e quais são as áreas mínimas (sítios) necessárias para que estes objetos persistam.

Em outras palavras, SCP busca estabelecer uma lista sequencialmente priorizada de sítios baseado em seu conteúdo de biodiversidade. Tal lista pode, então, ser usada para eleger redes de reservas que serão selecionadas para representar a biodiversidade de uma área de maneira tão completa quanto possível [135]. As áreas escolhidas por tal abordagem formam o que se chama de *Rede de Áreas de Conservação* (*Conservation Area Network–CAN*).

O conceito de SCP evoluiu ao longo do tempo, para contemplar outros aspectos da complexidade envolvida na seleção de áreas prioritárias, incluindo, por exemplo, a previsão de existência de restrições e conflitos.

Em quase todas as circunstâncias, a conservação da biodiversidade não é o único uso de uma região. Outros usos potenciais incluem a ocupação humana, seja para habitação, recreação, produção – inclusive agrícola, pscicultura, extração mineral, etc. Interesses humanos são particularmente importantes quando ligados ao bem-estar econômico de grupos desprivilegiados, como geralmente é o caso das áreas biologicamente mais importantes do mundo [202].

A abordagem SCP, procura, assim, proteger uma amostra representativa da biodiversidade regional e de maneira racional identificar áreas prioritárias para conservação, incorporando, quando possível, outros valores naturais e metas político-econômico-sociais por meio de *Análise Multicriterios* (*Multi-Criteria Analysis–MCA*) [155, 202]. Considera, assim, as possíveis restrições existentes, tais como conflitos no uso da terra e limitações econômico-financeiras, apontando uma solução de consenso que represente a melhor relação custo-benefício.

Uma das maiores tarefas da conservação é selecionar uma CAN que represente completamente a biodiversidade de uma área, mas em cenários do mundo real, esta representação completa nem sempre é possível, pois nem sempre todos os sítios de interesse biológico podem ser conservados. Desta forma, usualmente, a meta é que figurem todos os objetivos de conservação ou a representação até que um limite inferior (uma quantidade mínima) seja atingida ou mesmo a maior quantidade/extensão possível seja alcançada [135].

A abordagem estruturada do SCP oferece aos tomadores de decisão um sistema dinâmico de suporte à decisão, com a possibilidade de constante atualização e correção, permitindo criar diferentes cenários e desta forma identificar o melhor sistema de unidades de conservação capaz de *atingir as metas de conservação com menor custo e menos conflitos*. Além disso, o SCP deve lidar não só com a localização das reservas em relação a padrões físicos e biológicos, mas também com o projeto da reserva, que inclui variáveis como tamanho, conectividade, replicação. Uma abordagem sistemática estruturada para SCP fornece as bases necessárias para atender a esses objetivos.

Segundo Margules e Pressey [155], o SCP tem as seguintes características:

1. exige escolhas claras sobre os recursos a serem utilizados como substitutos à biodiversidade global no processo de planejamento;
2. baseia-se em objetivos explícitos, de preferência traduzidos em metas quantitativas e operacionais;
3. reconhece a medida em que as metas de conservação foram atingidas nas reservas existentes;
4. usa métodos diretos e simples para a detecção e seleção de novas reservas complementares às já existentes;
5. aplica critérios explícitos na implementação de ações de conservação do solo, especialmente no que diz respeito à gestão da proteção quando nem todas as áreas candidatas podem ser fixadas de uma só vez (o que normalmente acontece);
6. adota objetivos claros e mecanismos para a manutenção de condições dentro de reservas que são necessárias para persistência de recursos naturais chave, juntamente com o acompanhamento desses recursos e uma gestão adaptativa.

Assim, as principais características do SCP estão relacionadas ao estabelecimento de alvos e metas explícitos, cujas análises utilizam os seguintes princípios orientadores [30, 127]:

1. *Insubstituibilidade (irreplaceability)*: grandeza outorgada a uma área com o objetivo de retratar sua contribuição para a conservação da região analisada ou a probabilidade de tal porção do território fazer parte de uma solução que obedeça às metas

de conservação estabelecidas. Essa dimensão procura refletir a importância relativa dessas áreas no sistema. Por exemplo, se um objeto de conservação é representado em apenas um sítio, tal região é considerada essencial;

2. *Complementaridade*: característica almejada das localidades propostas para fazerem parte de um sistema de áreas protegidas previamente existentes, de maneira que objetos de conservação ainda não representados ou metas ainda não atingidas sejam contemplados [179]. Quando (conjuntos de) sítios são altamente complementares, quase não há sobreposição entre as características naturais neles representadas;

3. *Flexibilidade*: possibilidade de atingir a meta de proteção dos objetos de conservação por diferentes combinações de áreas equivalentes;

4. *Vulnerabilidade*: risco de destruição ou grau de ameaça a um determinado ambiente e/ou objeto de conservação;

5. *Representatividade*: junção de distintos tipos de ambientes e de metas de conservação a fim de retratar a biodiversidade em diferentes níveis;

6. *Persistência ou funcionalidade*: preservação, a longo prazo, da viabilidade e integridade biológica e ecológica dos objetos de conservação.

Para estabelecer metas de conservação é preciso pensar, para cada objeto selecionado, com base em sua viabilidade ecológica, o quanto é suficiente conservar. A porção necessária depende do tipo de dado disponível (e.g., área de *habitat* ou número de pontos de ocorrência) [127].

Estratégias de implementação, gestão e monitoramento continuam sendo problemas a serem investigados em SCP [202].

2.2 Formulando o Problema da Mínima Representação

Ao tratar o SCP, um critério importante para uma CAN é que ela represente o máximo de biodiversidade possível. Por este critério, uma CAN deveria conter pelo menos um exemplar de cada tipo de objeto de conservação presente na região de interesse.

Uma vez que há restrições, seria prudente escolher um conjunto de sítios que atingisse uma representação pelo mínimo custo, o que é chamado de *Problema da Mínima Representação*.

Quando é requerida pelo menos uma ocorrência de cada objeto de conservação e há um número finito de sítios discretos que podem ser escolhidos, este é um *Problema de Cobertura de Conjuntos* [177], conhecidamentente NP-difícil, clássico em Complexidade de Algoritmos.

O Problema da Mínima Representação é um problema de otimização que procura maximizar a *representação* da diversidade de objetos de conservação. Para atingir este propósito, ele tem sido formulado de duas maneiras [30]:

1. *O problema da cobertura mínima de conjuntos*: minimizar o número de sítios, área total ou custo, ao mesmo tempo em que se garante a representação de todas as *características naturais* (e.g., espécies) um dado número de vezes [183].

   Dada a matriz $A_{i \times j}$, onde $i = 1, ..., m$ corresponde a *sítios* e $j = 1, ..., n$ corresponde a *espécies a serem conservadas*, com a_{ij} representando a quantidade de ocorrências da característica j no sítio i.

   Seja $x_i \in \{0, 1\}$, tal que, $x_i = \begin{cases} 1, & \text{se o sítio } i \text{ foi selecionado;} \\ 0, & \text{caso contrário.} \end{cases}$

   Seja c_i o custo do sítio i.

   Seja r_j o nível de representação desejado, e.g., o tamanho de uma população mínima viável para a espécie j, ou seja, o número mínimo de indivíduos em uma população, necessário para garantir uma alta probabilidade de sobrevivência [239].

   O problema de otimização consiste em minimizar a Equação 2.1:

   $$\sum_{i=1}^{m} c_i x_i \tag{2.1}$$

   Sujeita à restrição expressa na Equação 2.2 (para todo j, cada característica seja representada pelo menos r_j vezes):

   $$\forall j \in \{1, 2, ..., n\}, \ \sum_{i=1}^{m} a_{ij} x_i \geq r_j \tag{2.2}$$

   Deste caso geral, podem ser derivados problemas particulares:

   (a) Minimizar o número de sítios necessários para a representação das características naturais. Fazendo-se:

   $c_i = 1$, todos os sítios passam a ter o mesmo custo;

   $a_{ij} = 0$ ou 1, se apenas a presença ou ausência da característica é considerada;

   Se $\begin{cases} r_j = 1, & \text{tem-se a representação única do problema;} \\ r_j \geq 1, & \text{tem-se a representação múltipla do problema.} \end{cases}$

   (b) Minimizar a área total necessária à representação de cada característica a um nível r_j. Tem-se:

$c_i = area_i$, o custo do sítio é sua área;

a_{ij} = área coberta pela característica j no sítio i;

r_j = nível de representação nas unidades de área.

(c) Minimizar o custo total de representação de características. Assemelha-se à instância (a), mas o custo de um sítio é seu preço real ($c_i = custo_i$).

2. *O problema da cobertura máxima*: maximizar a representação de características naturais dado um limite para o número de sítios, o custo ou a área. O problema, conforme originalmente enunciado por Church e ReVelle [38] e apresentado por [12, 31, 40], tem a limitação imposta sobre o número de sítios e consiste em:

   Dada a matriz $A_{i \times j}$, onde $i = 1, ..., m$ corresponde a *sítios* e $j = 1, ..., n$ corresponde a *espécies a serem conservadas*, com a_{ij} representando a ocorrência da espécie j no sítio i, da seguinte forma:

   $a_{ij} \in \{0, 1\}$, tal que, $a_{ij} = \begin{cases} 1, & \text{se a espécie } i \text{ está localizada no sítio } j \text{ ;} \\ 0, & \text{caso contrário.} \end{cases}$

   Sejam as variáveis de decisão x_i e y_j, tais que:

   Para cada sítio i, $x_i = \begin{cases} 1, & \text{se o sítio } i \text{ foi selecionado;} \\ 0, & \text{caso contrário.} \end{cases}$

   Para cada espécie j, $y_j = \begin{cases} 1, & \text{se a espécies } j \text{ está em pelo menos um sítio selecionado;} \\ 0, & \text{caso contrário.} \end{cases}$

   O problema de selecionar k sítios a fim de maximizar a cobertura das espécies consiste em maximizar a Equação 2.3:

$$\sum_{j=1}^{n} y_j \tag{2.3}$$

   Sujeita às restrições 2.4, 2.5, 2.6 e 2.7:

$$\sum_{i=1}^{m} a_{ij} x_i \geq y_j, \quad j = 1, ..., n \tag{2.4}$$

$$\sum_{i=1}^{m} x_i = k \tag{2.5}$$

$$0 \leq y_j \leq 1, \quad j = 1, ..., n \tag{2.6}$$

$$x_i = 0 \ \ ou \ \ 1, \quad i = 1, ..., m \tag{2.7}$$

Uma vez que a variável y_j é igual a 1 apenas quando a espécie j é representada, a Equação 2.3 mede o total de espécies cobertas pelos sítios selecionados. A Inequação 2.4 garante que y_j terá valor 1 apenas se aquela espécie é representada em pelo menos um sítio selecionado, o que é válido desde que para um dado j, a quantidade $\sum_i a_{ij}x_i$ seja o número de vezes que a espécie j é representada nos sítios selecionados. Assim, se a espécie j não é coberta, y_j terá valor 0; enquanto que y_j receberá valor 1, seu limite superior, se a espécie é coberta ao menos uma vez. A Equação 2.5 restringe o número de células que podem ser selecionadas (k). A Inequação 2.6 fornece os limites para y_i e indica que esta variável pode ser tratada como contínua, uma vez que as Equações 2.3 e 2.4 forçarão y_j para o valor 1 se a espécie é coberta pela seleção, mas faz com que o valor de y_j seja 0 se ela não é representada. A Equação 2.7 estabelece como obrigatória a restrição binária sobre cada variável x_i.

Outra maneira pela qual o problema pode ser enunciado é:

Seja S o conjunto de sítios a serem selecionados, n o número de características.

O problema de otimização consiste em maximizar a Equação 2.8:

$$\sum_{j=1}^{n} V_j(S) \tag{2.8}$$

Sujeita à restrição expressa na Equação 2.9:

$$\sum_{i \in S} c_i \leq k \tag{2.9}$$

Onde:

$V_j(S)$ é o valor da característica j na seleção S.

No caso mais simples:

$$V_j(S) = \begin{cases} 1, & \text{se o nível de representação desejado para a} \\ & \text{característica } j \text{ é atingido na seleção S;} \\ 0, & \text{caso contrário.} \end{cases}$$

$V_j(S)$ também pode ser utilizado para expressar a previsão de persistência da característica j na seleção S.

c_i é o custo do sítio i, podendo ser definido como 1 – se apenas o número de sítios é levado em consideração –, como a área do sítio, ou como seu custo real dependendo dos objetivos de otimização.

k é o recurso máximo disponível, contabilizado na mesma unidade de c_i.

O problema da máxima cobertura é uma extensão do problema da mínima cobertura de conjuntos na qual há uma restrição no número de sítios (ou custo, ou área, etc.) tendo em vista o valor estabelecido para k (Equação 2.9). Por causa desta restrição, nem todas as características serão preservadas.

ReVelle at. al [192] apontam que o problema da máxima cobertura pode também ser tratado como uma modificação do problema p-medianas [161, 191, 213].

Enunciado o problema, foram desenvolvidos algoritmos para lidar com ele. Tais algoritmos envolviam a seleção de sítios complementares sequenciais, até que o objetivo de representação de todas as espécies fosse atingido.

Para conjuntos grandes de dados, a solução por inspeção é difícil de encontrar. E mesmo com a implementação de algoritmos para a automatização do processo, como se trata de um problema NP-difícil, alguns conjuntos grandes de dados podem ser computacionalmente intratáveis por métodos exatos [184, 198]. Quando se considera o problema da mínima cobertura de conjuntos, para um conjunto de oito sítios, são possíveis $2^8 = 256$ soluções, variando entre a seleção de todos os sítios e a seleção de nenhum sítio. Para o exemplo trazido por Possingham et al. [177] para a ecorregião do Platô da Columbia (nos EUA), com 5.000 sítios, o número de sistemas de reservas possíveis é $2^{5.000}$, um número tão grande que é intratável. Quando se considera o problema da máxima cobertura, havendo, por exemplo, um total de n sítios, e sendo 2 a quantidade máxima de sítios que podem ser escolhidos ($k = 2$), o número de espécies cobertas por cada combinação de dois sítios do total de n pode ser calculado por $\binom{n}{2}$, e as combinações que fornecem a maior cobertura (maior número de espécies representadas) são, por definição, ótimas. Entretanto, como o número de tais combinações é $\binom{n}{k}$ o uso de tais técnicas exaustivas se torna impraticável para todos os valores não triviais de k [12].

A abordagem mais óbvia seria utilizar um *algoritmo guloso* (*greedy algorithm*) como o Algoritmo 1 (adaptado de [53]).

O algoritmo funciona da seguinte forma: dado um conjunto de sítios com espécies desprotegidas Σ, o conjunto $\Lambda \backslash \Lambda'$ contém, em cada etapa, o conjunto de elementos remanescentes não-cobertos (espécies desprotegidas ainda não selecionadas para proteção). O conjunto Γ contém a cobertura sendo construída. A linha 4 é o passo de *decisão gulosa*: um sítio σ_k é escolhido de tal maneira que "cubra" tantas espécies desprotegidas quanto possível (com "empates" sendo resolvidos arbitrariamente). Após σ_k ser selecionado, ele

Algoritmo 1: Algoritmo guloso para SCP

Dados:
Seja Σ um conjunto de sítios, com $\sigma_i \in \Sigma, i = 1, ..., m$, sítios individuais.
Seja Λ um conjunto de espécies desprotegidas, com $\lambda_j \in \Lambda, j = 1, ..., n$, correspondendo a espécies individuais.
Seja Γ o conjunto de sítios já selecionados.
Seja Λ' o conjunto de espécies $\lambda_j \in \Lambda$ representadas em Γ.
Seja $X_{ij} = \begin{cases} 1, & \text{se } \lambda_j \in \sigma_i \\ 0, & \text{se } \lambda_j \notin \sigma_i \end{cases}$

1 **begin**
2 $\Gamma \leftarrow \emptyset$
`/* executa o laço de repetição enquanto houver espécies não representadas */`
3 **while** $\Lambda \backslash \Lambda' \neq \emptyset$ **do**
4 Selecione um $\sigma_k \in \Sigma$, tal que, $\sigma_k = max(\sum_{\lambda_j \in \Lambda \backslash \Lambda'} X_{ij})$
5 $\Gamma \leftarrow \Gamma \cup \{\sigma_k\}$
6 Atualize Λ'
7 **return** Γ

é adicionado a Γ e as espécies desprotegidas nele representadas são adicionadas a Λ' e desconsideradas na próxima repetição ($\Lambda \backslash \Lambda'$). Quando o algoritmo termina, o conjunto Γ contém uma subfamília de Σ que cobre Λ, ou seja, que contém todas as espécies desprotegidas que se deseja salvaguardar.

O problema é que a solução encontrada com algoritmos gulosos é subótima e, portanto, ineficiente quando comparada a uma solução ótima. Todos algoritmos que escolhem sítios sequencialmente são ineficientes e não apresentam garantia de que se encontrará uma solução ótima [177]. Apesar disso, tal classe de algoritmos tem marcada importância na história do desenvolvimento de abordagens, algoritmos e ferramentas para SCP, o que será tratado na Seção 2.3.

2.3 Abordagens para SCP

Diversas abordagens para SCP têm sido sugeridas ao longo das últimas décadas, variando de um simples sistema de pontuação seguido de seleção [138, 181] a técnicas de otimização [31, 40, 225].

O desenvolvimento de algoritmos e ferramentas para suporte à decisão em SCP remonta à década de 1980 [198] e se tornou um componente importante de pesquisa em Biologia de Conservação [202].

O método mais comum utilizado no passado para SCP foi o *ad hoc* [15, 157, 183, 212], no qual os sítios são selecionados e adicionados à solução por motivos variados, tais como disponibilidade e baixo custo, ou mesmo com o objetivo de preservar uma espécie em particular, ou por conter algum valor não associado diretamente a biodiversidade, como algum fator geológico (e.g., jazida mineral ou sítio arqueológico), ou mesmo, apenas porque eram as regiões que restaram após as demais áreas terem sido alocadas para outros usos.

Outra abordagem corresponde à utilização de *sistemas de pontuação*, que têm a vantagem de tornar a base de decisão clara e explícita pela identificação de critérios que se mostram importantes e pela mensuração de tais critérios.

Há extenso registro de sistemas de pontuação utilizando as mais diversas medidas, variando de fatores sócio-politicos a conceitos tais como *wilderness* (que corresponderia a um valor associado a medidas estritamente biológicas, tais como presença e quantidade de espécies raras ou endêmicas, ou a presença de grande biodiversidade).

A ideia por trás destes métodos é atribuir uma pontuação a cada sítio independentemente dos demais, baseado em um número de diferentes critérios. Os critérios individuais são combinados resultando em uma pontuação única para o sítio (transformando o problema multiobjetivo em monobjetivo).

A maneira mais simples e também mais utilizada de combinação dos critérios é a soma dos mesmos, às vezes utilizando alguma ponderação com o objetivo de refletir possíveis hierarquias ou graus de importância entre os critérios. A reserva é então formada pela escolha de um dado número de sítios com as maiores pontuações.

A geração de tais pontuações individuais mostrou-se mais de uma vez complicada, em especial pela subjetividade inerente às diferentes medidas, podendo, inclusive levar a resultados diversos para o mesmo conjunto de dados [15, 156]. Além disso, quando se pensa no sistema de reservas como um todo, este sistema apresenta como limitação o fato de não refletir que o "valor" de um sítio individual depende de maneira crítica do que já está contido no sistema de reservas.

Assim, quando se pensa que uma CAN seria mais uma coleção de sítios que *funcionariam* melhor juntos do que uma coleção de sítios que seriam *bons* individualmente (sem levar-se em consideração o que já existe/integra a CAN), o sistema de pontuação acima descrito não leva em conta a primeira situação. Isso evidencia o fato da importância de um sítio não consistir apenas em valores intrínsecos, como riqueza de espécies, mas também em fatores outros, como em que o sítio *contribui* para a CAN, o que depende, por sua vez, do conteúdo de cada sítio que já integra a CAN. É possível mensurar o valor de um sítio individual no contexto do sistema de reservas, mas este valor variará toda vez que um sítio for acrescido ou removido do sitema.

A abordagem *iterativa*, que se seguiu, adotou um método simples para lidar com o

problema acima apontado: pontuar cada sítio de acordo com um critério baseado no que o sítio adiciona à CAN. Os sítios com maiores pontuações seriam selecionados para integrar a CAN e os sítios remanescentes teriam seus valores atualizados, tal processo continuaria até que um critério de parada fosse atingido (e.g., a adequada representação de um conjunto de objetos de conservação).

Este tipo de seleção utiliza a classe de *algoritmos gulosos* (Algoritmo 1), descrita na seção anterior, entretanto, conforme já foi ressaltado, tais algoritmos não apresentam a garantia de fornecer resultados ótimos. Um dos motivos reside no fato que, não surpreendentemente, espécies raras não aparecem necessariamente em sítios "ricos" em espécies [178]. Sítios ricos em espécies geralmente o são porque usualmente contêm espécies relativamente comuns (no conjunto de dados). Objetos de conservação que não aparecem nos sítos mais ricos serão incluídos no sistema de reserva mas em uma iteração bem posterior do algoritmo, depois que diversas escolhas já foram feitas.

Apesar da raridade de espécies já figurar entre os componentes da abordagem de pontuação, foi apenas no final da década de 1980 que o conceito de raridade foi incluído entre os valores da abordagem iterativa tornando-se a base de diversos algoritmos que utilizavam *heurística* [15].

Entretanto, cumpre ressaltar que soluções obtidas com heurísticas apresentam o risco substancial de serem grosseiramente subótimas, além disso, a obtenção de um bom resultado previamente utilizando uma determinada heurística não garante sua eficiência para todos os conjuntos de dados [195].

Em geral, na abordagem iterativa, não são utilizadas informações quanto ao posicionamento espacial do sítio, de maneira que as reservas resultantes tendem a ser bastante fragmentadas. Essa limitação foi contornada pela inclusão da informação de adjacência a reservas já existentes no processo de seleção.

A abordagem *exata* (que garante a produção de soluções ótimas) foi discutida inicialmente por Cocks e Baird, em 1989 [177, 198], ao apontarem que o problema SCP poderia ser formulado como o problema da máxima cobertura (Seção 2.2 conforme desenvolvido por Church e ReVelle [38]), e perceber que enunciado daquela forma, x_i poderia assumir valores contínuos entre 0 e 1, não estando restrito a 0 (ausência) ou 1 (presença), fazendo com que aquele fosse um *Problema de Programação Linear (PL)* para o qual pode ser encontrada solução em tempo polinomial [134], diferente do *Problema de Programação Linear Inteira (PLI)* (sabidamente NP-difícil [53]).

Williams e ReVelle [12, 239] ressaltaram a disponibilidade de pacotes de *software* codificando algoritmos exatos para PL e PLI (LINDO, CPLEX, OSL). Utilizando um destes pacotes para a resolução do problema original de Margules e Nicholls (tratado no Algoritmo 4), encontraram como solução ótima 11 sítios, enquanto Margules e Nicholls

encontraram 12, mostrando que a utilização de um algoritmo exato não representou uma melhora expressiva em termos de otimização.

Um crítica forte à utilização de algoritmos heurísticos em favor de métodos exatos foi imposta por Underhill [225], inclusive utilizando um tom hostil em seu artigo – sendo talvez este, junto com o fato de não ter aplicado suas observações a nenhum conjunto de dados reais, motivos para a pequena repercussão, à época, de suas considerações. Além disso, Underhill apontou uma formulação do Problema da Mínima Representação até então não enunciado: "maximizar o número de espécies que podem ser conservadas com um limite para o custo ou a área" (i.e., o Problema da Máxima Cobertura com limitação de custo ou área – Seção 2.2). Outra contribuição sua consiste na sugestão do uso de Análise Multicritérios para SCP.

Em resposta às críticas de Underhill, Pressey et al. [184] mostraram que métodos de PLI que garantem uma solução ótima, como algoritmos *branch-and-bound* – uma enumeração sistemática de todos os candidatos a solução, através da qual grandes subconjuntos de candidatos ineficazes são descartados em massa utilizando limitantes superior e inferior para *podar* ramos da árvore de enumeração que não contêm soluções ótimas –, são muitas vezes intratáveis no contexto de diversos problemas reais. Apontaram, também, que métodos heurísticos têm vantagens práticas sobre métodos clássicos e que a subotimalidade não é necessariamente um desvantagem para muitas aplicações do mundo real.

Pressey et al. [183] compararam alguns métodos heurísticos com a solução ótima encontrada por algoritmos exatos (basicamente *branch-and-bound*) a fim de determinar o *grau de subotimalidade* dos métodos não-exatos. Os resultados obtidos mostraram que, para os problemas de presença-ausência, os algoritmos exatos tiveram a vantagem de garantir a solução ótima, mas apresentaram tempos de execução muito mais longos do que os heurísticos. Soluções encontradas pelos métodos heurísticos foram 5-10% maiores que a ótima. Os algoritmos exatos falharam em resolver problemas de área proporcional enquanto os heurísticos os resolveram rapidamente. Assim, Pressey et al. [183] concluíram que a escolha do algoritmo depende do tamanho do conjunto de dados, das metas de representação, do tempo de análise disponível e da importância da garantia de que a solução seja ótima.

Outra abordagem utilizada para SCP consistiu na utilização de *meta-heurística*, um método heurístico para resolver de forma genérica problemas de otimização utilizando combinação de escolhas aleatórias e conhecimento histórico dos resultados anteriores adquirido pelo método para a realização de buscas pelo espaço de soluções. Tais técnicas foram empregadas em especial por Possingham e colaboradores [198]. Cumpre destacar que tal técnica tampouco garante a obtenção de soluções ótimas.

No programa SPEXAN (*SPatially EXplicit ANnealing*), foi usado *simulated annealing* [15] (Algoritmo 9) (uma técnica de busca local probabilística, baseada numa analogia com a termodinâmica[1] [140]), sendo esse o primeiro uso de um algorimo meta-heurístico para SCP [198]. Sua vantagem reside no controle maior sobre a configuração espacial das reservas quando comparado aos algoritmos heurísticos e no fato de permanecer tratável quando confrontado com grandes conjuntos de dados.

Outro método meta-heurístico posteriormente utilizado foi *busca tabu*[2] [39] (Algoritmo 10).

Critérios adicionais passaram a ser explorados, tais como forma, conectividade e dispersão, alinhamento (com tipos de *habitats* ou mesmo com unidades políticas), e uma diversidade de outros critérios espaciais.

Um aspecto crítico na abordagem do SCP tem sido o desenvolvimento e uso de ferramentas de suporte à decisão incorporando algoritmos especificamente desenvolvidos para solução do problema, o que será tratado nas duas subseções seguintes.

2.3.1 Algoritmos

Esforços iniciais no desenho de reservas derivam da Teoria da Biogeografia de Ilhas, cuja origem remonta à década de 1960. Tal teoria enfatizava aspectos tais como tamanho, forma e número de reservas, mas sem oferecer indicações explícitas e respostas conclusivas para a tarefa de seleção de áreas para reserva [177]. Como exemplo, tinham-se as seguintes regras [198]:

1. reservas grandes são preferíveis a reservas pequenas;
2. um única reserva grande é preferível a várias reservas pequenas cuja soma das áreas corresponda à da grande;

[1]Esta meta-heurística é uma metáfora de um processo térmico, conhecido como *annealing* ou recozimento, utilizado em metalurgia para obtenção de estados de baixa energia num sólido. O processo consiste de duas etapas: 1) a temperatura do sólido é aumentada para um valor próximo a 1.100°C (temperatura de início de transformação da austenita em ferrita); 2) o resfriamento é realizado lentamente até que o material se solidifique. Nesta segunda fase, os átomos que compõem o material organizam-se numa estrutura uniforme com energia mínima, tendo como resultado prático, uma redução dos defeitos do material. De maneira análoga, o algoritmo *simulated annealing* substitui a solução atual por uma próxima (i.e., em sua vizinhança no espaço de soluções), escolhida de acordo com uma função objetivo e com uma variável T (dita Temperatura, por conformidade com a inspiração térmica do algoritmo). Quanto maior for T, maior a componente aleatória que será introduzida na próxima solução escolhida. À medida que o algoritmo progride, o valor de T é decrementado, e o algoritmo começa a convergir para uma solução ótima.

[2]Procedimento adaptativo auxiliar, que guia um algoritmo de busca local na exploração contínua do espaço de busca. A partir de uma solução inicial, tenta avançar para uma outra solução (melhor que a anterior) na sua vizinhança até que um critério de parada seja satisfeito. O método é construído de forma a evitar o retorno a um ótimo local previamente visitado. Esta característica faz com que o método seja capaz de superar a otimalidade local em direção a um resultado ótimo ou próximo ao ótimo global. É uma técnica bastante semelhante ao *simulated annealing*.

3. reservas próximas umas das outras são preferíveis às que não o são;
4. reservas equidistantes são melhores do que aquelas que não o são;
5. reservas conectadas por corredores são melhores do que aquelas que não o são;
6. uma reserva circular é melhor do que uma alongada/comprida.

Contudo, ao longo do tempo, em especial ao final dos anos 1980, estas regras e a própria Teoria foram de certa forma abandonadas no que diz respeito ao desenho de reservas naturais [198]. Mais do que cegamente tentar maximizar o número de espécies baseado em abstrações da biogeografia de ilhas, percebeu-se a necessidade de tratar o SCP selecionando áreas que atingissem objetivos definidos, com o custo mínimo para outros possíveis usos da terra, baseando-se em informação empírica da distribuição das espécies [177]. Além disso, as regras da Teoria da Biogeografia de Ilhas jamais foram testadas, na prática, em nenhum conjunto representativo da dados [198].

O *primeiro algoritmo para SCP* de que se tem notícia, foi publicado por Kirkpatrick e colaboradores em 1980 [179, 199] (e republicado em 1983 [138]).

De um conjunto de 60 espécies de plantas endêmicas na região da costa centro-leste da Tasmânia, Kirkpatrick procurou priorizar áreas de conservação para 25 espécies (as outras 35 foram consideradas adequadamente representadas em – pelo menos duas – reservas já existentes, e por isso, não foram incluídas na análise) [139, 179, 182, 198].

Baseado em dados de ocorrência e abundância das espécies em estudo num *grid* de 460 células com aproximadamente $1 \times 1\ km^2$ cada (σ_i, no Algoritmo 2), Kirkpatrick [138] classificou as 25 espécies em 4 categorias (seguindo a abordagem de pontuação, utilizada à época) às quais foram atribuídos valores de prioridade (π_j, no Algoritmo 2):

1. ausente da reserva e bastante confinado à área de estudo ($\pi_j = 100$);
2. pobremente representado na reserva e bastante confinado à área de estudo ($\pi_j = 50$);
3. ausente da reserva e mais comum fora da área de estudo ($\pi_j = 25$); e
4. pobremente representado na reserva e mais comum fora da área de estudo ($\pi_j = 10$).

Para cada célula do *grid* (σ_i), os valores de prioridade das espécies nela presentes eram adicionados, obtendo-se um valor de conservação (κ_i, no Algoritmo 2).

Segundo Pressey [179], a análise de Kirkpatrick foi inovadora ao fazer uma importante observação: áreas com muitas espécies importantes apresentavam, na lista de valores de conservação inicial, pequeno κ_i enquanto muitas áreas com κ_i alto tinham as mesmas espécies.

Kirkpatrick, então, assumiu que os σ_i com maiores κ_i foram selecionados, ajustando, manualmente [179], todos os κ_i dos σ_i remanescentes, tirando as espécies que eram abundantes ou comuns aos σ_i já selecionados. Ele repetiu este passo de seleção de novas áreas até que as células "não selecionadas" tivessem pontuação menor que um valor K determinado (para ele $K = 30$).

Como resultado final, as células selecionadas pelo algoritmo foram agrupadas e acrescidas de uma faixa de $0,5\ km^2$ de segurança, resultando na recomendação de 7 áreas de reserva. Num decurso de aproximadamente 20 anos, todas as áreas recomendadas foram legalmente protegidas, de maneira que o *algoritmo pioneiro para SCP* foi o primeiro (e até hoje o único) completamente implementado na prática [179, 198, 203], talvez pela junção de dois elementos-chave: um método claro e aplicável, e o empenho e dedicação dos principais interessados[3].

Segundo Pressey [179], em 1984, Ackery e Vane-Wright, este último do Museu de História Natural de Londres, publicaram um livro a respeito das borboletas *Danaidae* (subfamília *Lepidoptera*, familia *Nymphalidae*), onde, em uma pequena seção, o problema SCP foi abordado por meio da descrição de um *método* para identificar *faunas críticas* (um conjunto mínimo de ilhas e zonas biogeográficas que representassem todas as 157 espécies de *Danaidae* conhecidas à época).

O método para priorização de áreas proposto por Ackery e Vane-Wright é apresentado no Algoritmo 3. Entretanto, cumpre ressaltar que não houve a proposta de um algoritmo no sentido em que tal termo é computacionalmente conhecido (como uma sequência clara, não ambígua, de instruções que são executadas até que determinada condição se verifique), dado que o passo apresentado na linha 9 do Algoritmo 3 não foi claramente especificado, o algoritmo descreve um problema de otimização mas não apresenta os passos para sua resolução [198].

Ainda assim, a contribuição do trabalho de Ackery e Vane-Wright reside na posterior introdução (em 1991) do termo *complementaridade* [132, 228], que significa a necessidade de áreas para complementar, mais do que duplicar, outras reservas segundo metas ainda não atingidas para as características que estas contêm. Assim, os algoritmos que utilizam tal conceito passaram a ser conhecidos como *algoritmos baseados em complementaridade*. Observe-se que, a já citada linha 9 do Algoritmo 3 determina a seleção do menor conjunto de áreas capaz de garantir a representação de todas as espécies que ainda não haviam

[3]Segundo Kirkpatrick citado por [179]:

> The major factor in acceptance was the desire of the forestry people to appear scientific in their conservation efforts... Everyone communicates with everyone else in Tasmania – I knew all the protagonists well. I think that the logic of the process, and its minimalism, also appealed.

Algoritmo 2: Algoritmo de Kirkpatrick (1980/1983)

Dados:
Seja Σ um conjunto de áreas (sítios, unidades de paisagem), com $\sigma_i \in \Sigma, i = 1, ..., m$, correspondendo a áreas individuais.
Seja Λ um conjunto de espécies, com $\lambda_j \in \Lambda, j = 1, ..., n$, correspondendo a espécies individuais.
Seja π_j o valor de prioridade da j-ésima espécie.
Seja κ_i o valor de conservação da i-ésima área σ_i.
Seja K o menor valor de conservação para uma área potencialmente protegida.
Seja ψ_{ij} o valor de conservação da j-ésima espécie, λ_j, na i-ésima área, σ_i.
Seja Γ o conjunto de áreas já selecionadas.
Seja $X_{ij} = \begin{cases} 1, & \text{se } \lambda_j \in \sigma_i \\ 0, & \text{se } \lambda_j \notin \sigma_i \end{cases}$

begin
  $\Gamma \leftarrow \emptyset$
  /* o algoritmo repete os passo seguintes até que todos os σ_i ainda não selecionados tenham um κ_i com valor menor que um limite mínimo K determinado */
  repeat
    /* atualiza ψ_{ij} da espécie λ_j das áreas ainda não selecionadas, com base no número de vezes em que a espécie λ_j ocorre nas áreas já selecionadas */
    forall the $\lambda_j \in \Lambda, \sigma_i \in \Sigma \backslash \Gamma$ **do**
      switch λ_j **do**
        case *ocorre uma vez em* $\mathscr{P}(\Gamma)$
          $\psi_{ij} \leftarrow \pi_{ij}$
        case *ocorre duas vezes em* $\mathscr{P}(\Gamma)$
          $\psi_{ij} \leftarrow \frac{1}{2}\pi_{ij}$
        case *ocorre mais que duas vezes em* $\mathscr{P}(\Gamma)$
          $\psi_{ij} \leftarrow 0$
    /* atualiza κ_i das áreas ainda não selecionadas */
    forall the $\sigma_i \in \Sigma \backslash \Gamma$ **do**
      $\kappa_i \leftarrow \sum_{j=1}^{m} X_{ij}\psi_{ij}$
    /* seleciona, dentre as áreas remanescentes ($\Sigma \backslash \Gamma$), aquela com κ_i mais alto */
    Selecione $\sigma_k \in \Sigma \backslash \Gamma$, tal que, $\kappa_k = max_{\kappa_i}(\Sigma \backslash \Gamma)$
    $\Gamma \leftarrow \Gamma \cup \{\sigma_k\}$
  until $\forall \sigma_i \in \Sigma \backslash \Gamma$ *tenha* $\kappa_i < K$

sido selecionadas por não serem endêmicas, ou seja, a representação das espécies complementares às selecionadas nas linhas 3–8 do Algoritmo 3.

Assim, como com Kirkpatrick, o "algoritmo" foi executado "à mão". O resultado foi um conjunto de 31 áreas necessárias para representar todas as espécies [228].

Algoritmo 3: Algoritmo de Ackery e Vane-Wright (1984)

Dados:
Seja Σ um conjunto de ilhas e zonas biogeográficas, com $\sigma_i \in \Sigma, i = 1, ..., m$, correspondendo a áreas individuais.
Seja Λ o conjunto de espécies, com $\lambda_j \in \Lambda, j = 1, ..., n$, correspondendo a espécies individuais.
Seja Γ o conjunto de área já selecionadas.
Seja $X_{ij} = \begin{cases} 1, & \text{se } \lambda_j \in \sigma_i \\ 0, & \text{se } \lambda_j \notin \sigma_i \end{cases}$

1 **begin**
2 $\Gamma \leftarrow \emptyset$
/* verifica, para toda espécie λ_j, se ela é endêmica (endemia entendida aqui como a ocorrência em uma única área σ_i) */
3 **forall the** $\lambda_j \in \Lambda$ **do**
4 **if** $\sum_{i=1}^{m} X_{ij} = 1$ **then** /* λ_j ocorre em uma única área σ_i:é endêmica */
5
6 **if** $X_{kj} = 1$ **then** /* sabendo-se que λ_j só ocorre em uma área, identifica esta área σ_k */
7
8 $\Gamma \leftarrow \Gamma \cup \{\sigma_k\}$
/* prioriza a seleção do menor conjunto de áreas σ_i ($min(|\Omega|)$) que garante a presença (cobertura) de todas as n espécies */
9 $\Gamma \leftarrow \Gamma \cup \Omega$, onde $\Omega \subseteq \Sigma$, tal que $\sum_{\sigma_i \in \Gamma} X_{ij} = n$ e $|\Omega|$ é o menor possível

Também em 1984, Margules e Nicholls [154] propuseram um algoritmo SCP baseado na probabilidade de ocorrência de comunidades de plantas (em especial eucaliptos) na Penísula Eyre, sudoeste da Austrália. Segundo Pressey [179], o algoritmo utilizava uma sequência de regras de seleção para encontrar um conjunto de porções de terrenos (*patches*) com o objetivo de representar cada comunidade com uma probabilidade de aceitação mínima. Os resultados foram apresentados em uma conferência em 1985, e o trabalho publicado dois anos depois.

O algoritmo foi aplicado como parte da priorização de 101 áreas (das quais apenas 21 foram amostradas [156]). Foi utilizado um método de classificação da vegetação que identificou 6 diferentes "comunidades" de plantas, para as quais foram empregadas, subsequentemente, técnicas para estimar a probabilidade de ocorrência de cada comunidade

em cada área. As áreas foram, então, priorizadas com base no Algoritmo 4. Nas linhas 3–5 do Algoritmo 4, as áreas com maior probabilidade de ocorrência de cada comunidade de plantas são selecionadas. Nas linhas 6–11, caso necessário, áreas adicionais são selecionadas a fim de garantir 95% de probabilidade de ocorrência de cada comunidade (para tanto, é calculado o produto da probabilidade de falha [156], $\prod_{\sigma_i \in \Gamma} (1 - \omega_{ij})$, que deve ser menor que 5%). Empates são resolvidos com base na extensão da área. O resultado final obtido apontou a necessidade de seleção de 12 áreas para representação de todas as espécies ao menos uma vez.

Algoritmo 4: Algoritmo de Margules e Nicholls (1987)

Dados:
Seja Σ um conjunto de áreas (*patches*), com $\sigma_i \in \Sigma, i = 1, ..., m$, correspondendo a áreas individuais.
Seja Λ um conjunto de comunidades de plantas, com $\lambda_j \in \Lambda, j = 1, ..., n$, correspondendo a comunidades individuais de plantas.
Seja ω_{ij} a probabilidade de ocorrência da j-ésima comunidade, λ_j, na i-ésima área, σ_i.
Seja Γ o conjunto de áreas já selecionadas.

1 **begin**
2 $\Gamma \leftarrow \emptyset$
`/* para cada comunidade de plantas` λ_j`, seleciona a área com a maior probabilidade de ocorrência */`
3 **forall the** $\lambda_j \in \Lambda$ **do**
4 Selecione $\sigma_k \in \Sigma$, tal que, $\omega_{kj} = max_{w_{ij}}(\Sigma)$
5 $\Gamma \leftarrow \Gamma \cup \{\sigma_k\}$
`/* busca garantir uma probabilidade de 95% de ocorrência de cada comunidade no conjunto da áreas selecionadas */`
6 **while** $\exists \lambda_j \in \Lambda$, *tal que,* $\prod_{\sigma_i \in \Gamma} (1 - \omega_{ij}) \geq 0,05$ **do**
7 **foreach** λ_l *na sitação acima descrita* **do**
8 Selecione $\sigma_k \in \Sigma \backslash \Gamma$, tal que, $\omega_{kl} = max_{w_{il}}(\Sigma)$
9 **if** *há empates* **then**
10 Selecione σ_k com a maior área (extensão em m^2)
11 $\Gamma \leftarrow \Gamma \cup \{\sigma_k\}$

Margules e colaboradores, em especial da *Commonwealth Scientific and Industrial Research Organisation (SCIRO)*, de Canberra, Austrália, desenvolveram outros três algoritmos diferentes (Algoritmo 5, 6 e 7), bem como duas variantes com modificações relativamente menores [198].

Entretanto, o Algoritmo 4, segundo Sarkar [198], seria o mais inovador, pela utilização de probabilidade nas análises. Os algoritmos posteriores deste grupo utilizaram apenas os

conceitos de presença/ausência das características consideradas (sem a opção de inclusão de incerteza).

Uma expressiva novidade com relação a seus predecessores foi que, pela primeira vez, os algoritmos enunciados foram computacionalmente implementados (em linguagem Fortran) [179, 198].

Em 1988, Margules, Nicholls e Pressey apresentaram dois algoritmos para SCP. O primeiro deles (Algoritmo 5), foi desenvolvido para representar cada uma das 98 espécies de plantas nativas presentes em 432 áreas pantanosas da planície inundada de Macleay River, localizada no noroeste de Nova Gales do Sul, Austrália, selecionando, para isso, o menor conjunto de áreas [156, 158].

O Algoritmo 5 inicialmente verifica se há espécies que só existem em uma única área e seleciona estas áreas (linhas 4–7). Em seguida, entra em laço de repetição que só termina quando todas as espécies estão representadas nas áreas selecionadas (linha 8-27). O laço começa pela seleção das áreas com as espécies mais raras ainda não selecionadas (linhas 9–10), se há mais de uma área com a mesma medida de raridade, é selecionada a que contribui com o maior número de espécies ainda não representadas (linhas 11–14), havendo ainda empate, opta-se pela área com o grupo menos frequente de espécies (definido como o grupo que tem a menor soma de frequências de ocorrência das áreas ainda não-selecionadas [156]) (linhas 15–18). Sarkar [198] chama atenção para este passo, considerando-o problemático na medida em que uma área com poucas espécies comuns poderia ter preferência na seleção sobre outra com muitas espécies raras; destaca, também, que em *todas* as demais variantes deste algoritmo, este passo desapareceu ou foi substituído. Se ainda assim, persiste o empate, ele é resolvido optando-se pela primeira área da lista (linhas 19–22). O resultado final obtido apontou a necessidade de seleção de 20 áreas (cobrindo 44,9% da área total) para representar todas as espécies ao menos uma vez.

O segundo algoritmo apresentado por Margules, Nicholls e Pressey em 1988 (Algoritmo 6) utilizou o mesmo conjunto de dados de entrada do algoritmo anterior e foi desenvolvido para representar, além de todas as espécies, todos os 9 tipos de *habitat* (ξ_l) nos quais as áreas pantanosas foram classificadas. O primeiro passo do algoritmo consiste em selecionar, para cada tipo de *habitat* as áreas com maior riqueza de espécies (linhas 4–7). Sarkar [198] aponta esse passo como problemático, na medida em que não testa as áreas para o número de espécies compartilhadas, apesar das similaridades provavelmente serem pequenas, devido à diferença de tipos de *habitat*. No passo seguinte (linhas 8–14), até que todas as espécies tenham sido representadas, o algoritmo continua a selecionar uma área de cada tipo, maximizando o número de novas espécies (linhas 10–13). Para representar todos os tipos de *habitats*, bem como todas as espécies de plantas, foi necessário selecionar 75,3% do total de áreas (em contraste com os 44,9%, encontrados pelo

Algoritmo 5: Algoritmo I de Margules, Nicholls e Pressey (1988)

Dados:
Seja Σ um conjunto de áreas pantanosas, com $\sigma_i \in \Sigma, i = 1, ..., m$, correspondendo a áreas individuais.
Seja Λ um conjunto de espécies de plantas, com $\lambda_j \in \Lambda, j = 1, ..., n$, correspondendo a espécies individuais de plantas.
Seja Γ o conjunto de áreas já selecionadas.
Seja Λ' o conjunto de espécies de plantas representadas em Γ.
Seja $X_{ij} = \begin{cases} 1, & \text{se } \lambda_j \in \sigma_i \\ 0, & \text{se } \lambda_j \notin \sigma_i \end{cases}$
Seja $\rho_j = \frac{1}{\sum_{i=1}^{m} X_{ij}}$ a medida da raridade de λ_j.
Seja $\phi_j = \frac{\sum_{i=1}^{m} X_{ij}}{\sum_{j=1}^{n} \sum_{i=1}^{m} X_{ij}}$ a medida da frequência de λ_j.

begin
  $\Gamma \leftarrow \emptyset$
  $\Lambda' \leftarrow \emptyset$
  `/* verifica para todas as espécies,` λ_j`, se só existem em uma única área e seleciona estas áreas */`
  forall the $\lambda_j \in \Lambda$ **do**
    forall the $\sigma_i \in \Sigma$ **do**
      if $\exists! \sigma_k$, *tal que,* $X_{kj} = 1$ **then**
        $\Gamma \leftarrow \Gamma \cup \{\sigma_k\}$
  `/* repete os passos seguintes até que todas as espécies de plantas estejam representadas em` Λ' `*/`
  repeat
    `/* seleciona a espécie mais rara dentre as ainda não selecionadas (`$\Lambda \backslash \Lambda'$`) */`
    Selecione λ_p, tal que, $\rho_p = max_{\rho_j}(\Lambda \backslash \Lambda')$
    `/* identifica a área,` σ_k`, onde` λ_p `ocorre */`
    Selecione σ_k, tal que, $X_{kp} = 1$
    `/* se é identificada mais de uma área onde` λ_p `ocorre... */`
    if *há empates* **then**
      `/* ... seleciona a área que contribui com o maior número de espécies ainda não representadas */`
      Selecione σ_q, tal que $\sum_{\lambda_j \in \Lambda \backslash \Lambda'} X_{qj}$ é máximo
    else
      $\sigma_q \leftarrow \sigma_k$
    if *há empates* **then**
      `/* seleciona a área com o grupo menos frequente de espécies */`
      Selecione σ_r, tal que, $\sum_{\lambda_j \in \Lambda \backslash \Lambda'} X_{ij}\phi_j$ é mínimo
    else
      $\sigma_r \leftarrow \sigma_q$
    if *há empates* **then**
      `/* seleciona a primeira área da lista */`
      Selecione σ_s pela ordem léxica
    else
      $\sigma_s \leftarrow \sigma_r$
    $\Gamma \leftarrow \Gamma \cup \{\sigma_s\}$
    `/* atualiza` Λ' `*/`
    forall the $\lambda_j \in \Lambda \backslash \Lambda'$ **do**
      if $\sum_{\sigma_i \in \Gamma} X_{ij} > 0$ **then**
        $\Lambda' \leftarrow \Lambda' \cup \{\lambda_j\}$
  until $\Lambda' = \Lambda$

algoritmo anterior, quando somente a presença de todas as espécies, ao menos uma vez, era exigida) [158].

O Algoritmo 5, sem a sequência representada pelas linha 15-18, tornou-se a base para os algoritmos desenvolvidos por aquele grupo. Como exemplo, Margules (em 1989), usou o Algoritmo 5 com a substituição dos passos das linhas 15–18 e com a utilização do critério da extensão da área, antes da ordem léxica[4] para resolução de empates, para selecionar reservas a fim de incluir associações de plantas em remanescentes de vegetação natural *mallee*[5] no sul da Austrália [198].

Pressey e Nicholls (em 1989) também publicaram uma variante do algoritmo no contexto de propriedades usadas para pastoreio no oeste de Nova Gales do Sul, Austrália. Havia 128 sistemas de terra e 1.026 propriedades. As propriedades foram pontuadas segundo quatro diferentes critérios: 1) diversidade; 2) raridade; 3) representação; e 4) uma combinação dos três critérios anteriores em uma única função. A alteração incorporada ao algoritmo dizia respeito à exclusão do passo contido nas linhas 15–18 do Algoritmo 5 e a inclusão de uma regra de preferência em caso de empate pela unidade com menor área, antes que fosse utilizado o critério léxico. Pressey e Nicholls, segundo o critério empregado, encontraram resultados bastante diversos [156]. No entanto, segundo Sarkar [198], o resultado mais interessante foi que, a partir de então, a elaboração do algoritmo – antes uma preocupação secundária diante do problema a ser resolvido – tornou-se um problema teórico em si mesmo.

Como resultados dos esforços que se seguiram, um algoritmo bem mais elaborado foi publicado por Nicholls e Margules em 1993 (Algoritmo 7), motivado, em especial, por fatores tais como: a hipótese de que deveria ser incentivada a diminuição da distância entre as reservas e o fato do desempate de áreas pelo critério da ordem léxica, utilizada até então, depender primordialmente da sequência de representação dos dados.

O Algoritmo 7 começa com a opção de inicialização do conjunto de reservas selecionadas (Γ), que pode já começar com algumas unidades especificadas (I), o que permite a inclusão de reservas pré-existentes (linha 2). Os primeiros passos são idênticos ao Algoritmo 5 (linhas 4–14). Da linha 15 à 46, o algoritmo estabelece critérios de resolução de empates que são sucessivamente: 1) a menor distância com relação a alguma unidade já selecionada (que pertence a Γ) (linhas 15–18); 2) a maximização da representação das características que ainda não atingiram os níveis-alvo (τ_j) (linhas 19–22); 3) níveis sucessivos de representação de raridade (linhas 23–30); 4) unidade com menor área capaz de fazer com que a característica atinja seu nível-alvo (linhas 31–38) – não é apresentada justificativa para este regra; 5) menor área (linhas 39–42); 6) ordem léxica (linhas 43–

[4] ordem determinada pela sequência dos dados no arquivo de entrada

[5] um termo usado para descrever áreas da Austrália cobertas principalmente por eucaliptos do tipo *mallee*.

Algoritmo 6: Algoritmo II de Margules, Nicholls e Pressey (1988)

Dados:
Seja Σ um conjunto de áreas pantanosas, com $\sigma_i \in \Sigma, i = 1, ..., m$, correspondendo a áreas individuais.
Seja Λ um conjunto de espécies de plantas, com $\lambda_j \in \Lambda, j = 1, ..., n$, correspondendo a espécies individuais de plantas.
Seja Ξ o conjunto de tipos de *habitat*, com $\xi_l \in \Xi, l = 1, ..., p$, correspondendo a tipos de *habitat* individuais.
Ξ é uma partição de Σ: cada tipo de *habitat* é composto por um certo número de áreas pantanosas e nenhuma área pantanosa pertence a mais de um tipo de *habitat*.
Seja Γ o conjunto de áreas já selecionadas.
Seja Λ' o conjunto de espécies de plantas representadas em Γ.
Seja $X_{ij} = \begin{cases} 1, & \text{se } \lambda_j \in \sigma_i \\ 0, & \text{se } \lambda_j \notin \sigma_i \end{cases}$

begin
  $\Gamma \leftarrow \emptyset$
  $\Lambda' \leftarrow \emptyset$
  /* para cada tipo de *habitat*, é selecionada a área com a maior riqueza de espécies */
  forall the $\xi_l \in \Xi$ **do**
    Selecione $\sigma_k \in \xi_l$, tal que, $\sum_{j=1}^{n} X_{kj}$ é máximo
    $\Gamma \leftarrow \Gamma \cup \{\sigma_k\}$
    Atualize Λ'
  /* os passos seguintes são repetidos até que todas as espécies estejam representadas em Λ' */
  repeat
    forall the $\xi_l \in \Xi$ **do**
      /* se para um determinado *habitat*, há área contendo espécie ainda não representada... */
      if $\exists \sigma_i \in \xi_l$, *tal que,* $\exists \lambda_j \in \Lambda \backslash \Lambda'$ *com* $X_{ij} = 1$ **then**
        /* ... seleciona a área do *habitat* (dentre as ainda não selecionadas) que maximima o número de novas espécies */
        Selecione σ_k, tal que, $\sigma_k = max_{\sigma_i \in \xi_l \cap (\Sigma \backslash \Gamma)} \sum_{\lambda_j \in \Lambda \backslash \Lambda'} X_{ij}$
        $\Gamma \leftarrow \Gamma \cup \{\sigma_k\}$
        Atualize Λ'
  until $\Lambda' = \Lambda$

46). Apesar do algoritmo ser aplicado à resolução de um problema (de representação de 31 comunidades de florestas no sudeste de Nova Gales do Sul), o foco do trabalho foi o algoritmo em si.

Cumpre destacar que de acordo com o critério de *eficiência* (e) de uma CAN, introduzido por Pressey e Nicholls em 1989 [198], dado por:

$$e = 1 - \frac{x}{t} \tag{2.10}$$

onde:

x = tamanho das unidades selecionadas para representar as características biológicas adequadamente;

t = tamanho total de todas as unidades.

todos os algoritmos iterativos apresentados tiveram melhores resultados do que os baseados em pontuação. Ainda assim, os algoritmos iterativos são heurísticos e não garantem a produção das soluções ótimas.

A tentativa de obter tais soluções ótimas, levou ao desenvolvimento de outra "família" de algoritmos para SCP, baseados em PLI (mencionados na Seção 2.3).

Em 1987, na Universidade da Cidade do Cabo, Rebelo (citado em [198]) havia começado a trabalhar com dados de distribuição de *Proteaceae*, uma família de plantas pertencente à vegetação *fynbos*, típica da região da Província do Cabo Ocidental, África do Sul. Ele estava interessado em usar a base de dados de que dispunha, rica em espécies endêmicas, para guiar a priorização de áreas de proteção e convenceu-se que um processo de seleção efetivo deveria basear-se no endemismo. Começou selecionando células de um *grid* onde se localizavam espécies únicas, já que estas regiões claramente integrariam qualquer conjunto de reservas, e percebeu que para prosseguir no processo de seleção, deveria considerar as espécies mais amplamente distribuídas que ocorriam concomitantemente às espécies únicas, aplicando, assim, complementaridade nas seleções seguintes [148]. O trabalho de Rebelo e Siegfried foi publicado em 1990, e foi realizado de maneira independente (sem o conhecimento) dos trabalhos que vinham sendo desenvolvidos paralelamente na Inglaterra e na Austrália.

Rebelo e Siegfried usaram um *grid* de 12×12 km^2 para mapear 326 *taxa* (espécies e subespécies distintas). Inicialmente foram distribuídos quatro secções (amostras de áreas – tomadas a partir da vegetação – sob a forma de uma faixa contínua) ao longo da região de estudo baseadas na riqueza da área. Seguindo o algoritmo proposto (Algoritmo 8), em cada secção, uma célula do *grid* foi selecionada com base na maior riqueza de espécies (linhas 4–7). No passo seguinte, células foram iterativamente selecionadas ao longo das

Algoritmo 7: Algoritmo de Nicholls e Margules (1993)

Dados:
Seja Σ um conjunto de unidades, com $\sigma_i \in \Sigma, i = 1, ..., m$, correspondendo a unidades individuais.
Seja a_i a área da unidade σ_i.
Seja Λ um conjunto de características biológicas, com $\lambda_j \in \Lambda, j = 1, ..., n$, correspondendo a características individuais.
Seja T o conjunto de níveis de representação para aquelas características, com $\tau_j \in T, j = 1, ..., n$, correspondendo ao nível-alvo para caracterísitcas individuais.
Seja I o conjunto de reservas inicialmente selecionadas (que pode ser igual a $\emptyset$).
Seja Γ o conjunto de unidades já selecionadas.
Seja Λ' o conjunto de características representadas em Γ.
Seja $X_{ij} = \begin{cases} 1, & \text{se } \lambda_j \in \sigma_i \\ 0, & \text{se } \lambda_j \notin \sigma_i \end{cases}$
Seja d_{rs} a distância entre as unidades σ_r e σ_s
Seja $\rho_j = \frac{1}{\sum_{i=1}^{n} X_{ij}}$ a raridade de λ_j.

begin
  $\Gamma \leftarrow I$
  $\Lambda' \leftarrow \emptyset$
  forall the $\lambda_j \in \Lambda$ **do**
    forall the $\sigma_i \in \Sigma$ **do**
      if $\exists! \sigma_k$, *tal que,* $X_{kj} = 1$ **then**
        $\Gamma \leftarrow \Gamma \cup \{\sigma_k\}$
  repeat
    Selecione λ_p, tal que, $\rho_p = max_{\rho_j}(\Lambda \backslash \Lambda')$
    Selecione σ_k, tal que, $X_{kp} = 1$
    if *há empates* **then**
      Selecione σ_q, tal que, $\sum_{\lambda_j \in \Lambda \backslash \Lambda'} X_{qj}$ é máximo
    else
      $\sigma_q \leftarrow \sigma_k$
    if *há empates* **then**
      Selecione σ_r, tal quel, $\forall \sigma_e \in \Gamma, \forall \sigma_f \in \Sigma \backslash \Gamma, d_{ef}$ tem d_{ek} como mínimo
    else
      $\sigma_r \leftarrow \sigma_q$
    if *há empates* **then**
      Selecione σ_s, tal que, $\sum_{\lambda_j \in \lambda \backslash \lambda_j} X_{sj}$ é máximo
    else
      $\sigma_s \leftarrow \sigma_r$
    if *há empates* **then**
      Selecione σ_t, tal que, $\rho_k = min_{\rho_j}(\Lambda \backslash \Lambda')$ e $\sum_{\sigma_i \in \Gamma \cup \{\sigma_t\}} X_{ik} > \tau_k$
    else
      $\sigma_t \leftarrow \sigma_s$
    if *há empates* **then**
      Selecione σ_u, tal que, $\rho_l = min_{\rho_j}(\Lambda \backslash (\Lambda' \cup \{\lambda_k\}))$ e $\sum_{\sigma_i \in \Gamma \cup \{\sigma_u\}} X_{il} > \tau_l$
    else
      $\sigma_u \leftarrow \sigma_t$
    if *há empates* **then**
      Selecione σ_v, tal que, $\rho_k = min_{\rho_j}(\Lambda \backslash \Lambda')$ e
      if $\sum_{\sigma_i \in \Gamma \cup \{\sigma_v\}} X_{ik} \geq \tau_k$ **then**
        a_v é minimo
      else
        a_v é máximo
    else
      $\sigma_v \leftarrow \sigma_u$
    if *há empates* **then**
      Selecione $\sigma_w \in \Sigma \backslash \Gamma$, tal que, a_w é mínimo
    else
      $\sigma_w \leftarrow \sigma_v$
    if *há empates* **then**
      Selecione σ_z pela ordem léxica
    else
      $\sigma_s \leftarrow \sigma_w$
    $\Gamma \leftarrow \Gamma \cup \{\sigma_k\}$
    /* atualiza Λ' */
    forall the $\lambda_j \in \Lambda \backslash \Lambda'$ **do**
      if $\sum_{\sigma_i \in \Gamma} Xij > \tau_j$ **then**
        $\Lambda' \leftarrow \Lambda' \cup \{\lambda_j\}$
  until $\Gamma' = \Gamma$

secções, desde que compartilhassem menos da metade de suas espécies com as áreas já selecionas e acrescentassem pelo menos um quarto de novas espécies (linhas 8–12). Finalmente, células são selecionadas a partir do resto da área em estudo com base nas novas espécies que elas contêm (linha 13). Caso haja empate, utiliza-se a riqueza para solucioná-lo (linhas 14–17). Persistindo o empate, a prioridade recai sobre a área com maior quantidade de *fynbos* (linhas 18–21), não havendo mais regras para desempate, o algoritmo assume que após este ponto não ocorrerão mais empates. Como resultado final, foi encontrado que 95% da espécies podem ser acomodadas em 16% da área de estudo e, mais importante ainda, foi o achado que menos de 17% das áreas existentes e menos de 50% das áreas propostas estavam em células com alto endemismo.

Segundo Sarkar [198], também para este grupo, os algoritmos se tornaram alvo de estudo. Em 1992 foram publicados mais dois algoritmos baseados em duas regras usadas no Algoritmo 5: 1) células que acrescentam mais espécies novas; 2) células que adicionam mais espécies raras. O primeiro algoritmo usa primeiro a Regra 1 e posteriormente a Regra 2 para desempates. O segundo algoritmo inverte esta ordem. Com relação à eficiência (*Cf.* Equação 2.10), não houve diferença significativa entre os dois algoritmos.

Em 1996, Ian Ball, da Universidade de Adelaide, Austrália, sob orientação de Hugh Possingham, escreveu, em linguagem C, o programa SPEXAN, utilizando pela primeira vez uma meta-heurística, *simulated annealing* (recozimento simulado), para tratar SCP. SPEXAN foi utilizado para estudar a dinâmica da ocorrência de *hollow bearing trees* (árvores com cavidades em seu tronco) em plantações para indústria madeireira. Posteriormente, *simulated annealing* foi utilizado em uma série de outros programas, tais como SITES e Marxan (ambos desenvolvidos por Ball e Possingham e derivados de SPEXAN).

O Algoritmo 9 mostra simplificadamente, o funcionamento do *simulated annealing*. Preliminarmente, o algoritmo começa com uma solução s_o, com uma energial inicial $E(s_o)$ (linha 2-5) e continua até que um critério de parada seja atingido (e.g., um número máximo de ciclos ou até que um estado de energia $eMax$ seja encontrado) (linhas 7–17). Durante o processo, a função $neighbor(s)$ deve gerar aleatoriamente um vizinho do estado s (vizinhos de um estado são novos estados do problema produzidos pela alteração de um determinado estado, de alguma forma especial, e.g., pela inserção de uma pequena perturbação local) (linha 9). A função $P(e, eNew, T)$ corresponde a uma probabilidade de aceitação calculada com base nos valores de energia e do estado s, da energia $eNew$ do vizinho do estado s (produzido na linha 9) e da temperatura T. A função $random()$ retorna um valor entre 0 e 1. Caso a probabilidade de aceitação de um novo estado seja maior que um valor gerado aleatoriamente desloca-se para o novo estado (linhas 11–13) e se este novo estado tiver uma energia menor (mais estável) do que o melhor estado até então encontrado, ele é armazenado (linhas 14–16) O esquema de "recozimento" é definido

Algoritmo 8: Algoritmo de Rebelo e Siegfried (1990)

Dados:
Seja Σ um conjunto de áreas, com $\sigma_i \in \Sigma, i = 1, ..., m$, correspondendo a unidades individuais.
Seja ϕ_i a quantidade de vegetação *fynbos* em cada σ_i.
Seja Λ um conjunto de espécies de plantas, com $\lambda_j \in \Lambda, j = 1, ..., n$, correspondendo a espécies individuais de plantas.
Seja T um conjunto de secções, com $\tau_l \in T, l = 1, ..., p$, correspondendo a secções individuais.
Seja Γ o conjunto de áreas já selecionadas.
Seja Λ' o conjunto de espécies de plantas representadas em Γ.
Seja $X_{ij} = \begin{cases} 1, & \text{se } \lambda_j \in \sigma_i \\ 0, & \text{se } \lambda_j \notin \sigma_i \end{cases}$

begin
  $\Gamma \leftarrow \emptyset$
  $\Lambda' \leftarrow \emptyset$
  forall the $\tau_l \in T$ **do**
    Selecione $\sigma_k \in \tau_l$, tal que, $\sum_{\lambda_j \in \Lambda} X_{kj}$ é máximo
    $\Gamma \leftarrow \Gamma \cup \{\sigma_k\}$
    Atualize Λ'
  forall the $\tau_l \in T$ **do**
    forall the $\sigma_k \in \tau_l$ **do**
      if $\sigma_k \in \tau_l$, *tal que,* $\frac{\sum_{\lambda_j \in \Gamma \cap \{\sigma_k\}} X_{ij}}{|\Gamma'|} < 0,5 \wedge \frac{\sum_{\lambda_j \in \Gamma \cup \{\sigma_k\}} X_{ij}}{|\Gamma'|} > 0,25$ **then**
        $\Gamma \leftarrow \Gamma \cup \{\sigma_k\}$
        Atualize Λ'
  `/* Ordenar as áreas remanescentes de acordo com a regra */`
  Selecione $\sigma_k \in \Sigma \backslash \Gamma$, tal que, $\sum_{\lambda_j \in \Lambda \backslash \Lambda'} X_{jk}$ é máximo
  if *há empates* **then**
    Selecione σ_p, tal que, $\sum_{\lambda_j \in \Lambda} X_{jk}$ é máximo
  else
    $\sigma_p \leftarrow \sigma_k$
  if *há empates* **then**
    Selecione σ_q, tal que, ϕ_i é máximo
  else
    $\sigma_q \leftarrow \sigma_p$
  $\Gamma \leftarrow \Gamma \cup \{\sigma_q\}$
  Atualize Λ'

pela função *temperature*(r) que deve produzir a temperatura a ser utilizada, dada uma fração r da previsão de tempo que foi gasto até então (a temperatura vai *decaindo* ao longo da execução do algoritmo) (linha 8). Ao final de sua execução, o algoritmo retorna a melhor solução encontrada.

Algoritmo 9: *Simulated annealing*

```
1  begin
2      s ← s_o                                          /* estado inicial */
3      e ← E(s)                  /* calcula a energia do estado inicial */
4      sBest ← s
5      eBest ← e
6      k ← 0
7      while not(stopCondition()) do
8          T ← temperature(k/kMax)                  /* calcula a temperatura */
9          sNew ← neighbor(s)                /* escolhe um vizinho e...    */
10         eNew ← E(sNew)                       /* ...computa sua energia */
           /* Convém mudar para o novo estado?                              */
11         if (P(e, eNew, T) > random()) then
               /* Sim, mude de estado                                       */
12             s ← sNew
13             e ← eNew
           /* É um estado melhor que o anterior?                            */
14         if e < eBest then
               /* Armazena novo estado como o melhor encontrado até então
                  */
15             sBest ← sNew
16             eBest ← eNew
17         k ← k + 1
18     return sBest
```

Segundo Sarkar [198], em 1997, como resultado da tese de mestrado de W. J. Okin, na Universidade da Califórnia, EUA, foi aplicada outra meta-heurística para SCP, *busca tabu*, que também foi utilizada por Ciarleglio em seu ConsNet [42, 43] (Seção 2.3.2).

O Algoritmo 10 mostra o funcionamento da *busca tabu*. Partindo de uma solução inicial, a busca move-se a cada iteração para a melhor solução na vizinhança, não aceitando movimentos que levem a soluções já visitadas por permanecerem armazenadas em uma *lista tabu*, que persiste na memória durante um determinado tempo ou certo número de iterações (*prazo tabu*). Como resultado final, espera-se que seja encontrado um ótimo global, ou uma solução mais próxima possível deste.

Linhas 2–4 do Algoritmo 10 representam configurações iniciais: a criação de uma solução inicial (possivelmente escolhida aleatoriamente), estabelecendo a solução inicial

Algoritmo 10: Busca tabu

```
1  begin
2      s ← s_o
3      sBest ← s
4      tabuList ← ∅
5      while not(stopCondition()) do
6          candidateList ← ∅
7          for sCandidate ∈ sNeighborhood do
8              if not containsTabuElements(sCandidate, TabuList) then
9                  candidateList ← candidateList + sCandidate
10         sCandidate ← locateBestCandidate(candidateList)
11         if fitness(sCandidate) > fitness(sBest) then
12             sBest ← sCandidate
13             tabuList ← featureDifferences(sCandidate, sBest)
14             while (size(tabuList) > maxTabuListSize) do
15                 expireFeatures(tabuList)
16         return sBest
```

como a melhor encontrada até o momento, e inicializando uma lista tabu vazia. Nesta simplificação, a lista tabu é tão somente uma estrutura de memória de curto prazo, que contém o registro dos elementos dos estados visitados.

O algoritmo propriamente dito tem início na linha 5, com um laço de repetição que continua à procura da solução ótima até que uma condição de parada especificada pelo usuário seja atendida (e.g., um limite de tempo ou um limite no valor da função *fitness* a ser atingido). Na linha 6, uma lista de candidatos é inicializada. As soluções vizinhas são verificadas no que diz respeito a elementos tabu nas linhas 7–9. Se a solução não contém elementos constantes da lista tabu, ela é adicionada à lista de candidatos (linha 9).

O melhor candidato na lista de candidatos é escolhido na linha 10 (geralmente, as soluções são avaliadas segundo uma função *fitness*). Se o candidato tem um valor maior do que o de melhor *fitness* (linha 11), é definido como o novo melhor valor (linha 12) e suas características são adicionadas à lista tabu (linha 13). Neste ponto, se a lista tabu estiver cheia (linha 14), alguns elementos poderão *expirar* (linha 15), geralmente o esquema é *primeiro a entrar, primeiro a sair (FIFO)* assim, os elementos na lista expiram na mesma ordem em que são adicionados.

Este processo continua até que critério de parada especificado pelo usuário seja atendido, e então a melhor solução encontrada ao longo do processo de busca é retornada (linha 16).

2.3.2 Ferramentas

Um aspecto crítico em SCP tem sido o desenvolvimento e uso de ferramentas de suporte à decisão incorporando algoritmos especificamente desenvolvidos para SCP [132, 198, 203].

Em 1999, Prendergast et al. [178] ressaltaram que ecologistas, conservacionistas e gestores relacionados à definição de políticas de conservação europeus e norte-americanos informaram que a razão principal para o pequeno nível de adoção de ferramentas para SCP se devia, à época, ao simples fato de tais atores desconhecerem a existência delas.

Ferramentas para SCP incluem:

1. WorldMap [166, 228];
2. TARGET [19];
3. C-Plan [185];
4. SITES [135, 177];
5. Marxan [16, 111, 235];
6. ResNet [135, 112, 199];
7. ConsNet [41, 42, 43, 200];
8. MultCSync [162, 201];
9. Zonation [164, 165].

WorldMap[6] é uma ferramenta para conservação de biodiversidade, raridade e priorização de áreas, que faz parte de um projeto de pesquisa do Laboratório de Biogeografia e Conservação do Museu de História Natural de Londres, cujo desenvolvimento teve início em 1988. Seus desenvolvedores destacam que, antes de ser mais um Sistema de Informação Geográfica (*Geographic Information Systems–GIS*) de propósito geral, WorldMap é uma plataforma de métodos de avaliação de biodiversidade com suporte cartográfico. Não foi encontrada uma descrição dos métodos e/ou algoritmos utilizados, mas segundo informação constante da página do projeto, o Laboratório responsável trabalha em colaboração com outros grupos de pesquisa para criar, comercialmente, produtos e serviços sob demanda, a fim de atender às necessidades de seus clientes, entre os quais estão incluídos gestores de recursos naturais, gestores de planejamento de conservação e especialistas em biodiversidade [21]. Da análise da informação constante em Muriuki et al. [166], WorldMap seria uma plataforma que disponibiliza um menu de serviços (em especial gráficos), mas cujos algoritmos empregados são personalizados para os clientes, no caso em apreço,

[6] http://www.nhm.ac.uk/research-curation/research/projects/worldmap/worldmap/demo2 .htm

teria sido implementado um algoritmo iterativo baseado em complementaridade para seleção do menor número de áreas que apresentassem, ao menos uma vez, todas as espécies (de aves do Quênia) contidas no banco de dados utilizado.

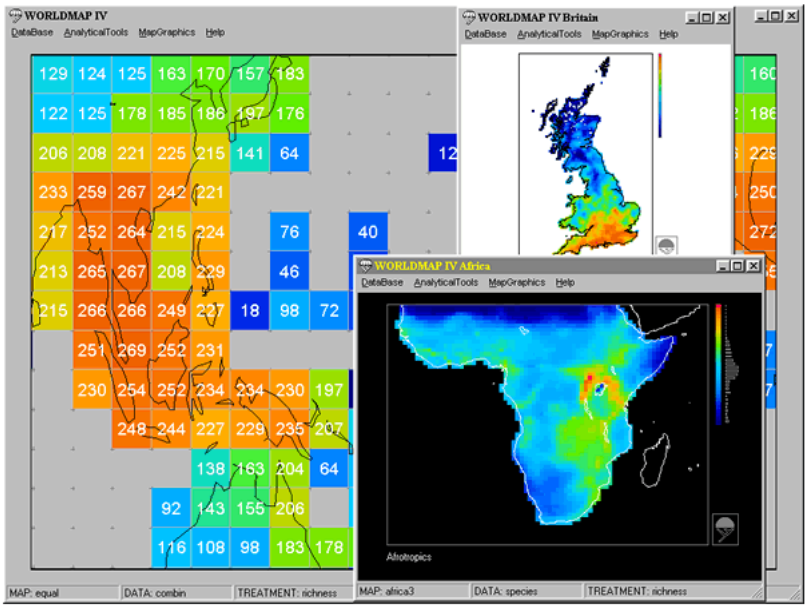


Figura 2.1: WorldMap versão 4 para Windows 95, 98 ou NT [21].

TARGET segue um procedimento de seleção com avaliação iterativa dos resultados. Utiliza um algoritmo guloso e, por isso, não garante a obtenção de um ótimo global, de maneira que requer que o usuário varie as condições iniciais objetivando testar a saída para ótimos locais em uma abordagem heurística. TARGET busca por soluções ótimas pela variação das ponderações (pesos) de custos.

C-Plan[7], desenvolvido, a partir de 1995, por Matthew Watts e Bob Pressey (com direitos pertencentes ao *New South Wales Department of Environment and Climate Change (NSW DEC)*) é um *software* de suporte à decisão que utiliza alvos e metas explícitas para análise da representatividade das unidades de conservação já existentes, através do cálculo do índice de importância biológica, ou índice de insubstituibilidade, das unidades de planejamento dentro da área em estudo, criando, desta forma, diferentes cenários de conservação. C-Plan foi descrito como o padrão-ouro em termos de *software* para SCP [2]. É possível trabalhar com o C-Plan por meio do ArcView 3.x (um programa que faz parte do *software* ArcGIS Desktop, um GIS proprietário, produzido pela empresa ESRI). Apesar de funcionar fora do ArcView 3.x, o C-Plan não possui uma interface própria de visualização gráfica e a leitura dos resultados e a manipulação de dados são feitas por meio de tabelas. Para visualizar os resultados sob a forma de mapas, é necessário operá-lo por meio do ArcView 3.x. Para entrada de dados é necessária a montagem de tabelas de alvos

[7]http://www.edg.org.au/free-tools/cplan.html

e de metas, bem como de abundância, o que não é uma tarefa simples, pois são necessários campos e formatos específicos. Para auxiliar a elaboração de tais tabelas, há uma extensão do ArcView chamada CLUZ[8]. CLUZ foi projetado originalmente para permitir a importação, análise e exibição de dados do Marxan.

C-Plan utiliza algoritmo guloso (por meio da *MINSET function*) baseado em insubstituibilidade, mas utilizando heurísticas (sistema de regras) para solução de empates [185]. C-Plan disponibiliza um *menu* para a *MINSET function* que permite ao usuário variar a ordem e ecolha das regras a serem aplicadas, diversificando, assim, o desenho do algoritmo. Por meio da funcionalidade *spattool* (ferramenta de configuração espacial), permite a análise de objetivos espaciais (tais como conectividade e tamanho). Apesar do exposto, C-Plan trata o problema SCP, que é multiobjetivo, como monobjetivo.

SITES[9] é um projeto derivado de SPEXAN 3.0 (*Spatially Explicit Annealing*) desenvolvido por Ian Ball e Hugh Possingham. Utiliza *simulated annealing* para otimizar uma função de *fitness* (referida como uma equação de custo) que computa a soma de um conjunto de componentes ponderados. O algoritmo iterativamente seleciona novos sítios, compara mudanças resultantes no que diz respeito à equação de custo e tenta minimizar o custo total. SITES possui um grande número de parâmetros opcionais que podem ser incorporados ao cálculo do custo, mas é necessário que o usuário defina seus pesos para o processo de busca. Tais parâmetros incluem um modificador de perímetro de fronteira (que permite controle quanto ao nível de agrupamento dos sítios), categorias de conservação e fatores de penalização para alvos sub-representados. Tem sido utilizado em projetos de conservação nos Estados Unidos, incluindo as regiões do sudoeste da Califórnia e o Platô da Columbia [135]. A comparação de critérios múltiplos, muitas vezes pode não ser viável pela incompatibilidade entre critérios que devem ser analisados separadamente. SITES utiliza GIS proprietário (ArcView). Possingham e Ball incorporaram o módulo SPEXAN usado em SITES ao Marxan e em novembro de 2007 a atualização de SITES foi descontinuada.

Marxan[10] foi desenvolvido com base no *software* SPEXAN. Uma versão inicial de SPEXAN correspondia ao algoritmo de seleção de sítios usado no *software* de planejamento *Environment Australia* (REST). Ambos, Marxan e SPEXAN são extensões do programa SIMAN e ALGO [15], codificados em FORTRAN77, que continham a inteligência do algoritmo, mas não eram uma versão amigável para usuários sem conhecimentos na linguagem [11]. Marxan (cujo nome provém de *MARine SPEXAN*) corresponde a uma atualização de SPEXAN desenvolvida em 2004 por Ball e Possingham, na Universidade de Queensland, Austrália, com o objetivo de auxiliar o rezoneamento da Grande Barreira

[8] http://www.kent.ac.uk/dice/cluz/
[9] http://www.biogeog.ucsb.edu/projects/tnc/toolbox.html
[10] http://www.uq.edu.au/marxan

de Coral (uma extensa faixa de corais situada entre as praias do nordeste da Austrália e Papua-Nova Guiné).

Ao longo do tempo, Marxan foi desenvolvido e aperfeiçoado – nos últimos anos, com a ajuda da *Applied Environmental Decision Analysis* (AEDA). Marxan é o *software* mais utilizado no auxílio à concepção e implementação de CANs marinhas e terrestres. Dados de 2010 informavam que Marxan contava com 2.600 usuários, em 110 países, pertencentes a 1.600 organizações que incluíam 220 universidades, a Organização das Nações Unidas e a *International Union for Conservation of Nature (IUCN)*, ONGs de conservação e 50 agências governamentais [2]. Destaca-se por buscar soluções eficazes para o problema da seleção de um sistema de áreas espacialmente coeso, com vistas a cumprir uma série de metas de conservação da biodiversidade, baseando-se numa minimização dos custos e dos efeitos de borda. Marxan assume a abordagem da cobertura mínima de conjuntos, utilizando uma combinação de *simulated annealing* e heurísticas (como minimização do perímetro da CAN, maximizando sua *compactação*) [111]. São limitações identificadas: a dificuldade em criar os arquivos de entrada (o que, como citado para C-Plan, pode ser feito com o auxílio da interface CLUZ); a dependência de GIS comerciais (proprietários) o que foi parcialmente superado pelo desenvolvimento do *Marxan Zonae Cogito*, que incorpora componentes de *softwares GIS* de código aberto; o fato de transformar um problema multidimensional (multiobjetivo) em monobjetivo pelo agregamento dos diferentes objetivos em uma única função a ser otimizada pelo algoritmo.

ResNet, *ConsNet* e *MultCSync* foram desenvolvidos no *Biodiversity and Biocultural Conservation Laboratory*, da Universidade do Texas em Austin, EUA[11].

ResNet implementa um algoritmo guloso de seleção baseado em raridade e complementaridade. O processo de priorização de sítios é iniciado pela seleção de um primeiro sítio com base na raridade, riqueza (único momento em que o critério de riqueza pode ser utilizado em ResNet é na inicialização) ou criado a partir de um conjunto de sítios definido pelo usuário (caso a CAN a ser construída tenha por base uma reserva pré-existente). A partir de então, ResNet aplica um procedimento iterativo que seleciona sítios com base na raridade, resolvendo empates fundamentado na complementaridade. Persistindo o empate, é usado o critério de adjacência e, finalmente, ordem léxica. É perceptível que ResNet se assemelha em muito à família de algoritmos desenvolvida na Austrália nos anos 1980-1990 (por exemplo, o Algoritmo 5). O usuário também tem a opção de incorporar a verificação de redundância, de infomação a respeito da área (tamanho) do sítio e de restrições quanto ao custo. A inovação de ResNet consiste na alocação dinâmica da memória [112]. ResNet foi usado em conjuntos de dados do Québec, Namíbia (Figura 2.2) e

[11] http://uts.cc.utexas.edu/consbio/Cons/Labframeset.html

Ilhas Malvinas [135]. ResNet depende da utilização de GIS proprietário e da geração de um arquivo de entrada.

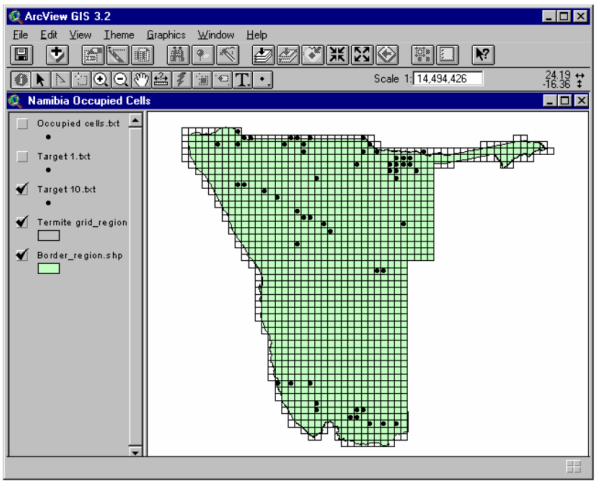


Figura 2.2: Resultado (visto no ArcView) da execução de ResNet (para a seleção de sítios representando todos os 33 gêneros de cupins da Namíbia) [112].

ConsNet[12] foi desenvolvido, em 2008, por Michael Ciarleglio como projeto de seu Doutorado na Universidade do Texas em Austin. Trata-se de um pacote de *software* para o desenho e análise de CANs para representação da biodiversidade. Segundo o autor, o SCP, apesar das muitas variações que dependem do objetivo específico do planejador, tem uma estrutura básica, e seu *software* seria genérico o bastante para contemplar tal estrutura. ConsNet emprega técnicas multicritério para desenho das CANs. Em particular, lida com uma variedade de critérios espaciais incluindo tamanho, densidade, conectividade, replicação e alinhamento [43]. Foi desenvolvido com base no *framework Modular Abstract Self-Learning Tabu Search (MASTS)* [42], que utiliza busca tabu como meta-heurística. Em MASTS, algoritmos de busca paralela podem tirar proveito de máquinas com multiprocessadores ou com memória compartilhada, além disso, também suporta novas estratégias de busca tabu como objetivos baseados em regras (*rule-based objectives–RBO*) e seleção dinâmica de vizinhos (*dynamic neighborhood selection–DNS*) [43], que podem melhorar o desempenho da pesquisa em regiões promissorar e reduzir o número de avaliações necessárias.

MultCSync [162] consiste em um arquivo executável, adicionalmente, o usuário pode usar *Gnuplot*[13], um pacote de *software* livre para apresentação gráfica dos resultados (Figura 2.3). MultCSync inicialmente computa um subconjunto de alternativas não-dominadas a partir do conjunto de alternativas factíveis. MultCSync lida com dois critérios conflitantes. Se mais de dois critérios são utilizados, são definidas preferências entre

[12]http://uts.cc.utexas.edu/consbio/Cons/consnet_home.html
[13]http://www.ncftpd.com/download/

os critérios (Figura 2.4a) para estabelecer *rankings* (Figura 2.4b). São fornecidas três opções de refinamento [201]: 1) permitir que os critérios menos importantes sejam sequencialmente desconsiderados; 2) permitir o uso de *Analytic Hierarchy Process* (AHP), para produzir um *ranking* de todas as alternativas não-dominadas; 3) utilizar um AHP modificado (onde as prioridades associadas a uma solução são normalizadas). Utiliza análise multicritério, onde os critérios são hierarquizados e a eles é atribuída uma ponderação, desta forma, MultCSync também trata o problema SCP como monobjetivo.

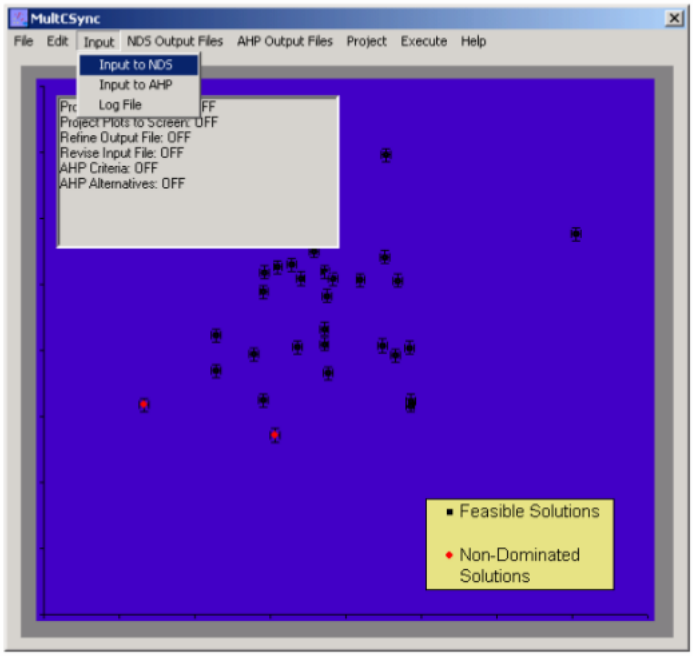
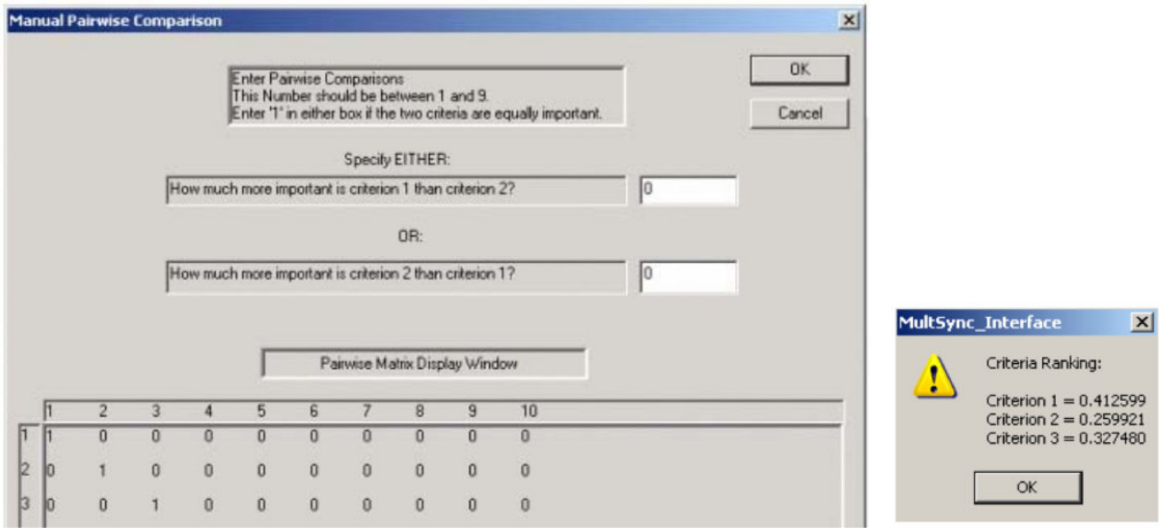


Figura 2.3: *Interface* principal do MultCSync [201].

Zonation aborda o problema da máxima cobertura, buscando a maximização dos benefícios de conservação para um custo fixo especificado pelo usuário. Começando pelo conjunto total de sítios, Zonation calcula a perda marginal para cada um dos sítios da região sob estudo. Com base em tal valor, o sítio que representa a menor perda marginal é removido, e o processo é repetido, calculando-se o valor da perda marginal para cada um dos sítios remanescentes e procedendo-se à remoção do que menos contribui para o valor de conservação da CAN. Assim, os sítios são removidos, um de cada vez, tendo como critério de decisão, a minimização da perda marginal. Como resultado, obtem-se a maximização geral dos valores de conservação dos sítios remanescentes [165]. Zonation permite a incorporação de opções de conectividade e viabilidade ao processo de priorização, por meio da utilização de heurísticas na forma de penalizações como a *boundary quality penalty* (BQP), *boundary lenght penalty* (BLP), a fim de favorecer a obtenção de CANs com maior agregação (compactação). Em vez de um conjunto ótimo de sítios que



(a) (b)

Figura 2.4: MultCSync. (a) Exemplo da caixa de diálogo mostrando o estabelecimento manual de comparação entre os critérios dois a dois (neste exemplo é mostrada a comparação parcial de dez critérios). (b) Exemplo de estabelecimento de hierarquia (ponderação) entre três critérios, de maneira automática, com utilização de AHP [201].

cumprem os objetivos-alvo, a saída de Zonation é uma hierarquia de remoção de sítios do conjunto disponível e uma curva de perda de espécies.

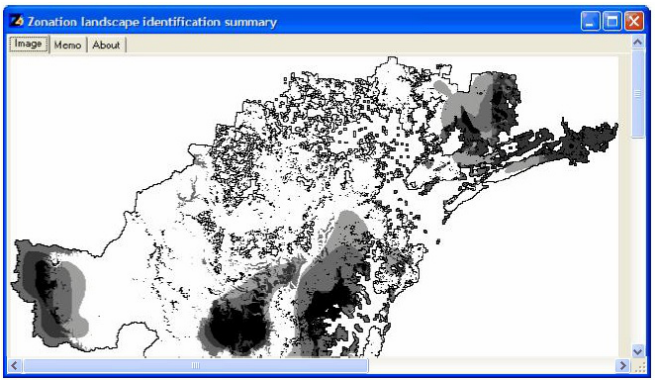


Figura 2.5: Janela do Zonation mostrando um exemplo da saída de uma execução, quando a opção de exibição do mapa foi ajustada para escala de cinza [165].

Tabela 2.1: Estratégias de solução para SCP e métodos utilizados.

	Estratégia de Solução					
	Pontuação	*Gulosa*	*PL/PLI*	*Meta-heurística*		
Métodos				*Simulated Annealing*	*Busca Tabu*	*Computação Natural*
Kirkpatrick (1983) [138]	x	x				
Ackery e Vane-Wright (1984) [5]		x				
Margules e Nicholls (1987) [154][1]		x				
Margules, Nicholls e Pressey I (1988) [158][2]		x				
Margules, Nicholls e Pressey II (1988) [158]		x				
Nicholls e Margules (1993) [168][3]		x				
Rebelo e Siegfried (1990) [189][4]		x				
Algoritmos exatos (utilização de pacotes de *software* como LINDO, CPLEX, OSL) (1994) [12, 31, 194, 225, 239]			x			
C-Plan (1995) [185][5]		x				
SPEXAN (1996) [15, 177]				x		
SITES (2000) [177]				x		
ResNet (2002) [112, 132][6]		x				
Marxan (2004) [2, 11, 177][7]				x		
Target (2004) [19]		x				
Zonation (2005) [163, 165][8]		x				
MultCSync (2005) [162][9]	x	x				
ConsNet (2008) [41][10]					x	
NSGA-II aplicado a SCP (2012) [206, 205, 209]*						x
MAIS(2015) [208]*						x

Heurística para aprimoramento dos resultados:
[1] critério de desempate (área).
[2] critério de desempate (raridade, frequência, ordem léxica).
[3] critério de desempate (raridade, distância, área, ordem léxica).
[4] critério de desempate (riqueza).
[5] critério de desempate.
[6] critério de desempate (complementaridade, adjacência).
[7] maximização da compactação, minimização do perímetro.
[8] conectividade, BQP, BLP.
[9] análise multicritérios.
[10] MCDA, RBO, MASTS.

* Propostos nesta tese.

Capítulo 3

Otimização Multiobjetivo

Problemas que podem ser modelados com o Problema da Cobertura de Conjuntos têm, naturalmente, muitos objetivos, em geral, conflitantes entre si. Buscando uma simplificação na solução, muitas vezes, utiliza-se a abordagem monobjetivo, onde os diferentes objetivos são tratados como se fossem um – agregando-os em uma única função de avaliação (*função fitness*) –, ou considerando apenas um objetivo e tratando os demais como restrições. Estes problemas de otimização com mais de um objetivo são chamados problemas de *otimização de vetores* ou *multiobjetivo* [47] e serão objeto deste capítulo.

Na Seção 3.1 são apresentados conceitos básicos de Otimização Multiobjetivo. Na Seção 3.2 é apresentado o Teorema *No Free Lunch*, que estabelece um desempenho médio igual a qualquer algoritmo de otimização (seja ele monobjetivo ou multiobjetivo) quando medido em comparação ao conjunto de problemas que é capaz de resolver.

3.1 Conceitos Básicos

Um *Problema de Otimização Multiobjetivo (MOP)* pode ser definido como o problema de se encontrar "um vetor de variáveis de decisão que satisfaça restrições e otimize um vetor de funções cujos elementos representam as funções objetivo. Estas funções formam uma descrição matemática de critérios de desempenho que normalmente estão em conflito. Assim, o termo 'otimizar' significa encontrar a solução que fornece ao tomador de decisões os valores aceitáveis para todas as funções objetivo" [172].

Assim, um MOP (minimização) pode ser descrito como:

$$Encontrar\ \overrightarrow{x}\ que\ otimiza\ F(\overrightarrow{x}) = (f_1(\overrightarrow{x}), ..., f_k(\overrightarrow{x})), \quad k = 1, ..., K \tag{3.1}$$

Sujeito a:

$$\begin{aligned} G_j(\overrightarrow{x}) \leq 0, \quad j = 1, ..., J \\ H_m(\overrightarrow{x}) = 0, \quad m = 1, ..., M \\ x_i^{(L)} \leq x_i \leq x_i^{(U)}, \quad i = 1, ..., n \end{aligned} \tag{3.2}$$

onde $\overrightarrow{x}$ é o vetor de soluções (ou variáveis de decisão) $\overrightarrow{x} = [x_1, x_2, ..., x_n]^T$, J é o número de restrições de inequações e M é o número de restrições de equações [56]. As k funções objetivo podem ser lineares ou não-lineares, contínuas ou discretas [48].

As soluções que satisfazem as funções objetivo e as restrições constituem o que se chama espaço de variáveis de decisão factíveis (ou espaço de busca, ou espaço de decisão, ou simplesmente espaço de variáveis, e corresponde a onde se faz a busca pelas soluções do problema, ou seja, é o domínio das variáveis do problema) $\Omega \subset \Re^n$. Uma das diferenças marcantes entre a otimização monobjetivo e a multiobjetivo consiste no fato que, nesta última, as funções objetivo constituem um espaço multidimensional, em acréscimo ao espaço de soluções usual. Este espaço k-dimensional é chamado espaço de objetivos, $\Lambda \subset \Re^k$. Para cada solução $\overrightarrow{x}$ no espaço de variáveis de decisão, há um ponto $\overrightarrow{y} \in \Re^k$, no espaço de objetivos, denotado por $F(\overrightarrow{x}) = \overrightarrow{y} = [y_1, y_2, ..., y_k]^T$, de maneira que, uma solução é referida como um vetor de variáveis e um "ponto" como o correspondente vetor objetivo [87] (Figura 3.1).

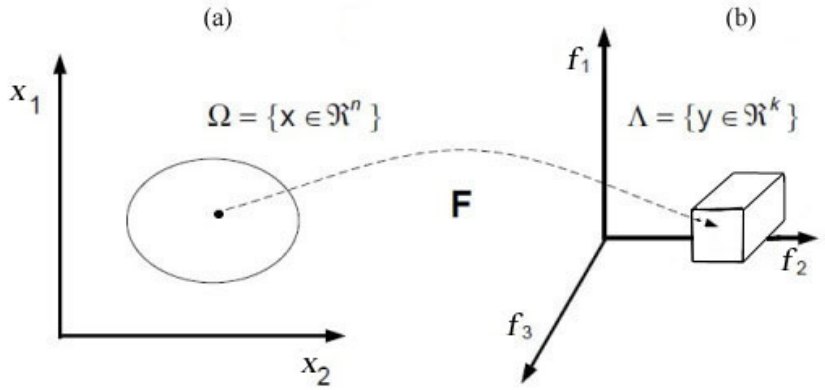


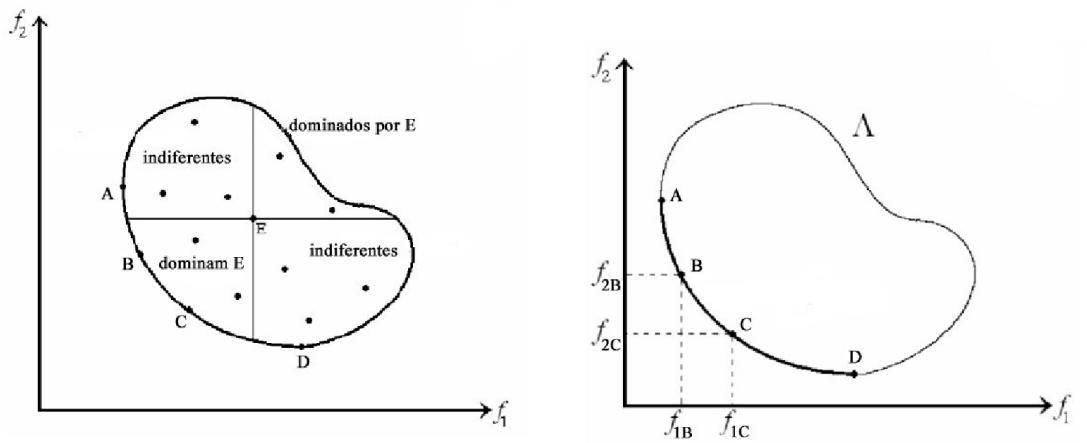
Figura 3.1: Mapa de avaliação de MOP. (a) Espaço de busca (Ω). (b) Espaço de objetivos (Λ) (adaptado de [227]).

Desta maneira, o problema MOP pode ser enunciado como segue:

Definição 1 Problema de Otimização Multiobjetivo Geral (minimização): Em geral, um MOP minimiza $F(\overrightarrow{x}) = (f_1(\overrightarrow{x}), ..., f_k(\overrightarrow{x}))$ sujeito a $G_j(\overrightarrow{x}) \leq 0, \quad j = 1, ..., J$, com $\overrightarrow{x} \in \Omega$. Uma solução para MOP minimiza os componentes de um vetor $F(\overrightarrow{x})$, onde $\overrightarrow{x}$ é um vetor de variáveis de decisão n-dimensional ($\overrightarrow{x} = x_1, ..., x_n$) de algum universo Ω [227].

Assim, um MOP consiste em n variáveis de decisão, j restrições, e k objetivos cujas funções podem ser lineares ou não-lineares, com a função de avaliação do MOP, $F : \Omega \rightarrow \Lambda$, mapeando variáveis de decisão ($\overrightarrow{x} = x_1, ..., x_n$) em vetores ($\overrightarrow{y} = y_1, ..., y_k$).

Nos MOPs, otimizar significa encontrar todos os valores aceitáveis para as funções objetivo com vistas a uma tomada de decisão. O espaço de busca Ω (que contém todos os possíveis valores de x que satisfazem $F(\overrightarrow{x})$) é parcialmente ordenado, ou seja, duas soluções arbitrárias são relacionadas de duas maneiras possíveis: ou uma domina a outra ou nenhuma delas domina (Figura 3.2a).



(a) (b)

Figura 3.2: Representação gráfica de conceitos de Pareto em duas dimensões (com $F(\overrightarrow{x}) = (f_1(\overrightarrow{x}), f_2(\overrightarrow{x}))$). (a) Dominância de Pareto no espaço de objetivos. (b) Conjunto dos ótimos de Pareto.

Se quaisquer componentes de $F(\overrightarrow{x})$ competirem, não existirá uma solução única para o problema e sim um conjunto de soluções, consistindo em todos os vetores de decisão que não podem ser simultaneamente melhorados (onde não se pode melhorar nenhum dos objetivos sem a degradação de outro). Desta maneira, o conceito de ótimo de Pareto deve ser utilizado para a caracterização e obtenção das soluções [107].

Uma solução ótima de Pareto (ou não-dominada) é aquela onde a melhora em um dos objetivos resulta necessariamente em degradação do outro.

No exemplo bidimensional da Figura 3.2b, o conjunto de ótimos de Pareto está representado entre os pontos A e D. Na mesma figura, B e C são também ótimos de Pareto. Uma melhora no objetivo f_1 resulta em degradação do objetivo f_2, ou seja, $f_{1B} < f_{1C}$ *e* $f_{2B} > f_{2C}$.

Definição 2 Dominância de Pareto: Sejam $\overrightarrow{a}$ e $\overrightarrow{b} \in \Omega$, $\overrightarrow{a}$ domina $\overrightarrow{b}$ (denotado para o problema de minimização como $F(\overrightarrow{a}) \preceq F(\overrightarrow{b})$), se e somente se:

$$\forall i \in \{1, 2, .., k\},\ f_i(\overrightarrow{a}) \leqslant f_i(\overrightarrow{b}) \ \wedge\ \exists j \in \{1, 2, ..., k\}, f_j(\overrightarrow{a}) < f_j(\overrightarrow{b})$$

Em outras palavras, $\overrightarrow{a}$ não é pior que $\overrightarrow{b}$ em nenhum dos objetivos e é melhor em pelo menos um.

Todos os vetores de decisão que não são dominados por nenhum outro vetor de decisão são chamados *não-dominados* ou *ótimos de Pareto*.

Definição 3 Otimalidade de Pareto: Uma solução $\overrightarrow{x^*} \in \Omega$ é um ótimo de Pareto, se e somente se, $\overrightarrow{x^*}$ é não-dominada em relação a Ω, ou seja, nenhum vetor do espaço de busca domina $\overrightarrow{x^*}$

A otimização multiobjetivo visa, portanto, a obtenção e seleção dos ótimos de Pareto.

A solução esperada é composta por um conjunto de pontos de equilíbrio, uma família de soluções consideradas iguais entre si e superiores em relação ao restante das soluções.

Definição 4 Conjunto Ótimo de Pareto: Para um dado MOP $F(\overrightarrow{x})$, o conjunto ótimo de Pareto (P^*) é tal que:

$$P^* = \{\overrightarrow{x} \in \Omega | \nexists \overrightarrow{y} \in \Omega, F(\overrightarrow{y}) \preceq F(\overrightarrow{x})\}$$

Definição 5 Frente de Pareto: Para um dado MOP $F(\overrightarrow{x})$ e um conjunto ótimo de Pareto P^*, a Frente de Pareto (FP^*) pode ser definida como:

$$FP^* = \{F(\overrightarrow{x}) = [f_1(\overrightarrow{x}), f_2(\overrightarrow{x}), ..., f_k(\overrightarrow{x})] \ | \overrightarrow{x} \in P^*\}$$

Da mesma maneira que nos problemas de otimização monobjetivo, há a possibilidade de que os MOPs apresentem ótimos de Pareto locais e globais. Em alguns casos, grande

parte das soluções é atraída para as frentes de Pareto locais. A Figura 3.3 mostra a diferença entre uma frente de Pareto global e uma local.

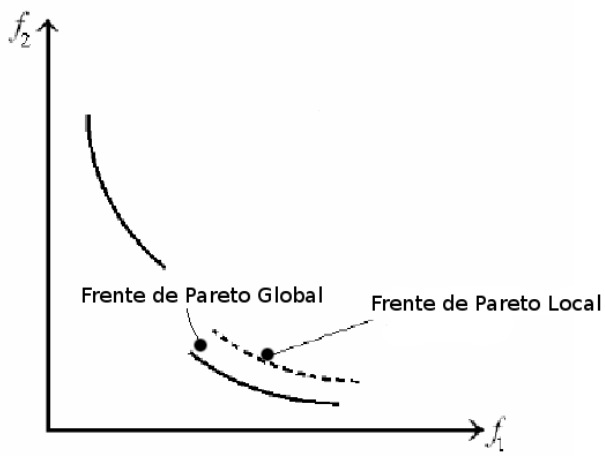


Figura 3.3: Frente de Pareto local e global (mais próxima à origem dos eixos).

Definição 6 Conjunto ótimo de Pareto Local: Se para todo $\overrightarrow{x} \in P$, não existir solução $\overrightarrow{y}$, $||\overrightarrow{y} - \overrightarrow{x}|| \leq \varepsilon$, que domine qualquer solução pertencente a P, então as soluções pertencentes ao conjunto P constituem um conjunto ótimo de Pareto local. ε é um número positivo pequeno e é obtido de uma perturbação em uma pequena vizinhaça de $\overrightarrow{x}$ [86].

Definição 7 Conjunto ótimo de Pareto Global: Se não existir solução que domine qualquer elemento do conjunto P no espaço de busca, então as soluções pertencentes a P constituem o conjunto ótimo de Pareto Global [86]

Encontrar o ótimo global de um MOP é, em geral, um problema NP-completo [13].

O tamanho e forma da frente de Pareto dependem, normalmente, do número de funções objetivo e de suas interações.

Qualquer MOP pode ser convertido em um problema de otimização monobjetivo pelo agrupamento dos objetivos em uma função escalar $s : \Re^n \rightarrow \Re$ que é a forma que a otimização clássica lida com objetivos múltiplos [246]. Existem muitos métodos para efetuar o agrupamento de objetivos em uma função escalar [129, 227], mas todos requerem profundo conhecimento e informação preferencial do decisor. No entanto, para muitos

problemas práticos, este conhecimento não está disponível e, além disso, resolver um MOP por técnicas monobjetivo sempre resulta em uma solução de ponto único.

Em diversas situações é indesejável ou difícil trabalhar exclusivamente com pontuação numérica. Quando dois critérios representam diferentes sistemas de valores, pode ser impossível agregar tais critérios de uma maneira significativa, por exemplo, como comparar a preservação de espécies em perigo e a promoção do desenvolvimento econômico? Insistir em uma função objetivo única pode levar à agregação de quantidades díspares, requerendo assunções que muitos tomadores de decisão podem considerar inaceitáveis e conduzir a resultados inacurados [43].

Na literatura, alguns métodos foram propostos para reduzir o conjunto de Pareto a um tamanho manejável. Entretanto, o objetivo não é apenas limitar um determinado conjunto, mas gerar um subconjunto representativo que mantenha a diversidade e as características do conjunto original [246].

3.2 Teorema *No Free Lunch*

O Teorema *No Free Lunch (NFL)* (numa livre tradução: Teorema "Não Há Almoço Grátis") explora a conexão existente entre algoritmos de otimização e os problemas aos quais são aplicados, demonstrando que se um algoritmo tem um bom desempenho em uma certa classe de problemas, então ele necessariamente "paga um preço" por isso com um desempenho pior no conjunto de todos os demais problemas [240].

A formulação original do Teorema NFL afirma que, em média, cada algoritmo de otimização tem o mesmo desempenho quando nenhum conhecimento prévio a respeito do custo da função *monobjetivo* f é assumido [240]. Köppen [145] estende o Teorema NFL para o caso de otimização *multiobjetivo* por meio do seguinte teorema[1]:

Teorema 1. *Dados dois algoritmos determiniśticos* [2] *a e b, um valor de desempenho* $k \in \mathbb{R}$ *e uma medida de desempenho c:*

$$\sum_{f} \delta(k, c(m, a, f)) = \sum_{f} \delta(k, c(m, b, f)) \tag{3.3}$$

onde δ *é o delta de Kronecker* [3] *e* (m, a, f) *representa a sequencia de m sucessivas aplicações do algoritmo a ao problema multiobjetivo* f*.*

[1] Para prova do Teorema 1 *v.* [145].

[2] A análise pode ser estendida para algoritmos estocásticos [146, 240].

[3] O delta de Kronecker $\delta(i, j)$ é a função das variáveis i e j que assume valor 1 se as variáveis são iguais, e 0 caso contrário: $\delta(i, j) = \begin{cases} 0, & \text{se } i \neq j \\ 1, & \text{se } i = j \end{cases}$

De maneira simplificada, o que o Teorema mostra é que, em média, cada algoritmo tem o mesmo desempenho quando aplicado a todos os possíveis problemas f. Assim, um algoritmo desenvolvido especificamente para um problema A terá um desempenho melhor comparado a outros algoritmos para os quais não foi introduzido conhecimento prévio do domínio do problema, mas seu desempenho se deteriorará rapidamente para problemas bem diferentes de A. Um algoritmo genérico terá sempre um desempenho razoável, mas será incapaz de superar o algoritmo específico na classe de problemas para as quais este é desenvolvido.

O Teorema NFL implica que se conhecimento do domínio do problema não for incorporado ao domínio do algoritmo, não há garantias formais que um algoritmo geral robusto exista para aquele problema [48]. Implica, ainda, que incorporar muito conhecimento do domínio do problema no algoritmo reduz sua efetividade em outros problemas fora e mesmo dentro de uma classe particular de problemas, o que, de maneira mais ampla, indica que pode existir uma dicotomia entre um conjunto de testes e uma possível aplicação ao mundo real.

Assim, ao mesmo tempo que o Teorema NFL evidencia o perigo de se comparar algoritmos por seu desempenho em uma pequena amostra de problemas, estes mesmos resultados indicam a importância de se incorporar conhecimento específico do problema no algoritmo a ser desenvolvido [240].

Diante do exposto, Köppen [145] sustenta que o Teorema NFL também pode ser visto como capaz de estabelecer a impossibilidade de se obter uma definição matemática concisa do desempenho de um algoritmo, desta forma, propõe o chamado *tournament performance*, um procedimento para construção de diferentes medidas de desempenho para diferentes algoritmos multiobjetivo da seguinte forma:

1. Execute o algoritmo a por k evoluções/gerações da função f e faça o teste M_1 para identificar os pontos não-dominados obtidos pelo algoritmo;
2. Selecione k pontos aleatórios do domínio e compute o Conjunto de Pareto M_2 dos valores correspondentes de f;
3. Compute o conjunto M_3 dos elementos de M_2 que não são dominados por nenhum elemento de M_1.

A relação entre $|M_1|$ e $|M_2|$ fornece uma medida de como o algoritmo a tem um desempenho melhor do que uma busca aleatória (*random search*) (o que corresponderia ao *modelo nulo*, *Cf.* Seção 6.4).

Capítulo 4

Inspiração Biológica: Algoritmos Evolutivos e Sistemas Imunológicos Artificiais

Nesta tese, é proposto um algoritmo baseado em *Sistemas Imunológicos Artificiais* (*Artificial Immune Systems–AIS*). Cumpre destacar que, em que pese alguns autores [55, 58, 62, 136] incluírem os AIS entre os Algoritmos Evolutivos (*Evolutionary Algorithms–EA*) (Figura 4.1a), considera-se mais adequada a abordagem adotada por de Castro [75], que os coloca dentro do grupo maior da Computação Natural (Figura 4.1b).

A Computação Natural pode ser considerada como a utilização de materiais e mecanismos naturais a fim realizar algum tipo de computação [76]. Ela contempla, em essência, três métodos [75]:

1. os que tomam inspiração da natureza com o objetivo de desenvolver novas técnicas para solução de problemas (Algoritmos Bioinspirados);
2. os baseados no uso de computadores para simular e emular fenômenos naturais, podendo resultar, desta forma, em um melhor entendimento de mecanismos naturais;
3. os que empregam materiais e mecanismos naturais (e.g., cadeias de DNA) para computar (novos paradigmas de computação).

Sem querer ser exaustiva, uma lista dos principais campos de pesquisa que integram os métodos da Computação Natural inclui: neurocomputação, *algoritmos evolutivos*, inteligência de enxames, *sistemas imunológicos artificiais*, geometria fractal, vida artificial, computação de DNA, computação quântica [76].

No intuito de obter uma maior compreenssão da natureza, a Computação Natural se beneficia da integração de linhas de pesquisa experimental e teórica das mais variadas áreas do conhecimento (e.g., Biologia, Física, Química, Matemática, etc.).

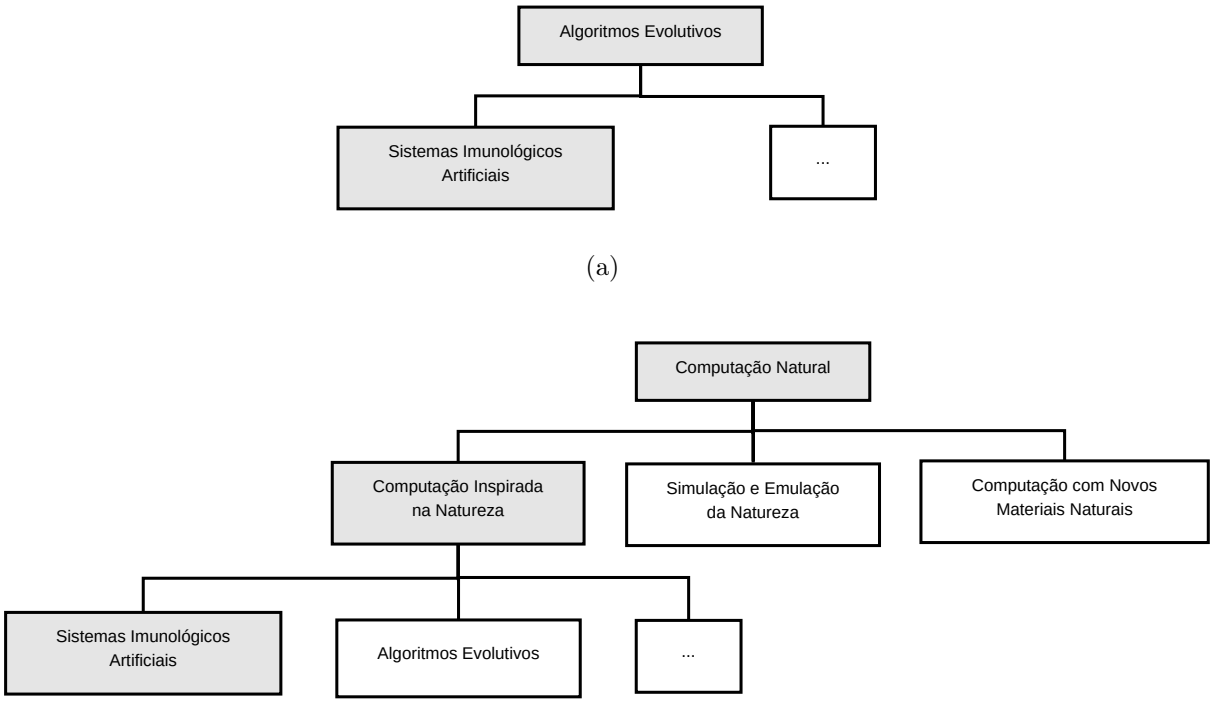


Figura 4.1: Como classificar os Sistemas Imunológicos Artificiais? (a) Como uma variante dos Algoritmos Evolutivos. (b) Como um ramo da Computação Inspirada na Natureza, uma das três principais vertentes da Computação Natural (abordagem considerada mais adequada).

A maioria das abordagens computacionais para Computação Natural são baseadas em versões simplificadas de mecanismos e processos existentes na natureza. Estas simplificações são necessárias para tornar a computação com um grande número de elementos manejável. Além disso, a simplificação pode ser vantajosa na medida em que permite destacar os itens mínimos necessários para reproduzir um aspecto particular do sistema e para observar algumas propriedades emergentes. Some-se a isso o fato que muitas vezes um modelo simplificado é suficiente para obter o fim almejado e, frequentemente, detalhes acerca do fenômeno natural observado são desconhecidos [75].

Não se pode perder de vista, entretanto, que as classificações são abstrações criadas pelo homem para facilitar sua esquematização do mundo a fim de melhor analisá-lo e interpretá-lo, assim que essencialmente são artificiais. Fato é que Stepney et al. [221] apresentaram um esforço em estabelecer um *framework* para AIS e, ao fazê-lo, estenderam-no para outros algoritmos como EA, evidenciando que ao mesmo tempo que se pode especializar um algoritmo, pode-se generalizar suas definições e com isso mostrar que compartilham de um arcabouço comum [10, 222].

Na Seção 4.1 serão tratados os conceitos básicos da Teoria Evolutiva apropriados pelos Algoritmos Evolutivos, bem como sua aplicação no contexto da Otimização Multiobjetivo, em especial serão abordados os algoritmos NSGA-II e SPEA2, usados como *baseline* de comparação do algoritmo proposto nesta tese. A Imunologia é um campo vasto e na Seção 4.2 serão abordados os aspectos considerados essenciais à compreensão dos princípios utilizados no algoritmo proposto para lidar com o problema SCP.

4.1 Algoritmos Evolutivos

Algoritmos Evolutivos (EA) integram a vertente da Computação Natural que se inspira na natureza para a solução de problemas, mais especificamente, baseiam-se em princípios da Teoria da Evolução das Espécies e da Genética [13]. Os primeiros trabalhos com EA datam da década de 1930, quando sistemas evolutivos naturais passaram a ser estudados com algoritmos de exploração de múltiplos picos de uma função objetivo [85].

No livro "A origem das espécies" de 1859, Charles Darwin propôs as bases da Teoria Evolutiva. Um dos fundamentos do modelo evolutivo de Darwin foi a sobrevivência do mais apto, i.e., aqueles indivíduos melhor adaptados ao seu ambiente terão maior probabilidade de sobrevivência. Este fenômeno corresponde à *seleção natural* e ocorre necessariamente se uma população de indivíduos tem que competir por uma quantidade limitada de recursos e tem que escapar de predadores.

Além disso, espécies são capazes de se adaptar a mudanças no ambiente por meio de *mutações*, responsáveis pela introdução de variações aleatórias nos indivíduos. Seus

efeitos ocorrem com uma certa probabilidade durante o processo reprodutivo e resultam em modificações na geração subsequente. Como consequência de processos reprodutivos, mutações e seleção natural, que ocorrem iterativamente, características da geração atual são combinadas e levemente modificadas na geração seguinte. Apenas as caracterísitcas presentes naqueles indivíduos mais adaptados, que sobrevivem e conseguem produzir uma nova geração, são preservados.

Em suma, uma espécie evolui através da reprodução, mutação e seleção natural. Assim, a seleção natural afeta indivíduos, enquanto a população como um todo evolui. Os três mecanismos evolutivos citados estão presentes nos EA por meio de *loops* iterativos (correspondendo às sucessivas gerações de uma população); operadores de mutação, de *crossover*, de seleção; além disso, indivíduos podem ser considerados soluções candidatas.

EA têm sido aplicados com sucesso em uma grande variedade de problemas de otimização [13, 48, 70]. Uma das grandes vantagens na utilização dos EA é sua generalidade, i.e., a possibilidade de lidar com problemas complexos, e ainda assim ser possível encontrar boas soluções – inclusive sem que se tenha necessariamente um conhecimento prévio aprofundado das peculiaridades do problema. Em contrapartida, não se tem a garantia de obtenção de uma solução ótima, pois esta técnica heurística pode convergir para ótimos locais, dos quais nem sempre é fácil sair [70].

4.1.1 Conceitos Básicos

Em essência, num EA, dada uma *população* de *indivíduos* (i.e., um conjunto de possíveis soluções), pressões do ambiente desencadeiam um processo de *seleção natural* que privilegia as melhores soluções até então encontradas, causando um incremento na adequação dessas soluções.

Os principais componentes dos EA são [85]:

1. população de indivíduos que concorrem por recursos limitados;
2. aptidão, que reflete a habilidade do indivíduo para sobreviver e reproduzir-se;
3. mudanças dinâmicas na população devido ao nascimento e morte de indivíduos;
4. hereditariedade e variabilidade, de maneira que novos indivíduos possuem muitas características de seus pais, mas sem serem idênticos a eles.

Desta maneira, dada uma função a ser otimizada (maximizada ou minimizada), é gerado, aleatoriamente, um conjunto de soluções (i.e. elementos pertencentes ao domínio da função). Tal conjunto corresponde a uma população, onde cada solução é um indivíduo.

Uma função de medida de adequação (*fitness*) é aplicada a cada indivíduo a fim de testar a sua aptidão, ou seja, medir a qualidade da solução candidata [98].

Tomando por base o *fitness*, algumas das melhores soluções são selecionadas para dar origem a uma nova população pela aplicação de operadores de variação (*mutação* e *recombinação*, também referida como *crossover*) responsáveis por criar variabilidade entre as soluções. A recombinação é um operador aplicado a duas ou mais soluções candidatas (chamadas *pais*) e resulta em dois ou mais novos indivíduos (os *descendentes* ou *filhos*) que são combinações dos pais. A mutação é aplicada em uma solução candidata a fim de gerar outra.

A seleção natural é consequência dos seguintes fatores: 1) indivíduos de uma população diferem entre si; e, 2) é produzida uma descendência em número maior de indivíduos do que de fato os que podem sobreviver. Assim, os indivíduos mais aptos a sobreviverem são aqueles que, graças à variabilidade genética, herdaram a combinação gênica mais adaptada para determinadas condições naturais (que têm melhor *fitness*).

Desta forma, dada uma população de tamanho N, seja N_d o número de descendentes, então, para a próxima geração, são selecionados N novos indivíduos entre as $N + N_d$ possíveis soluções, ou entre somente os N_d novos indivíduos (cada EA desenvolve, com base nesse princípio, uma estratégia de seleção), de maneira que ao final do processo, as novas soluções candidatas competem com as soluções da geração anterior (e/ou entre si), com base no *fitness*, para assumir um lugar na nova população.

Esse processo é repetido até que seja atingido um critério de parada, que pode ser o fato de um indivíduo apresentar uma solução que seja suficientemente qualificada ou mesmo que um número máximo de iterações (*gerações*) seja atingido.

O Algoritmo 11 mostra um EA típico [98].

Algoritmo 11: Algoritmo Evolutivo

Resultado: *população_final*: conjunto otimizado de soluções.

```
1  begin
2      Inicializar a população com soluções candidatas aleatórias (indivíduos)
3      Avaliar cada indivíduo com a função fitness
4      repeat
5          Selecionar pais
6          Recombinar pares de pais
7          Mutar os descendentes resultantes
8          Avaliar novos indivíduos
9          Selecionar indivíduos para a nova geração
10     until not(critério de parada)
11     return população_final
```

Vários componentes do processo evolutivo são estocásticos: a seleção favorece os indivíduos mais bem adaptados (com melhor *fitness*), mas há a possibilidade de serem selecionados outros indivíduos; o *crossover* entre indivíduos é aleatório, assim como a

mutação, ou seja, o(s) indivíduo(s) resultante(s) da aplicação dos operadores de variação depende(m) de uma série de escolhas aleatórias. Tal característica faz com que não haja garantias de que se alcance um resultado ótimo.

4.1.2 Algoritmos Evolutivos e a Otimização Multiobjetivo

EA parecem ser particularmente apropriados para a tarefa de encontrar soluções ótimas de Pareto pois processam um conjunto de soluções em paralelo. Fonseca e Fleming [107] bem como Valenzuela-Rendón [226] sugerem que a otimização e pesquisa multiobjetivo podem ser áreas onde EA seja capaz de produzir melhores resultados quando comparado a outros métodos computacionais.

Desde 1896, quando o conceito de ótimo de Pareto foi introduzido, diversas técnicas para resolução de MOP, tradicionais (Programação Matemática) ou alternativas (Algoritmos Genéticos, Algoritmos Evolutivos, Sistemas Imunológicos Artificiais, Enxame de Partículas), têm sido desenvolvidas. O objetivo é que os métodos apresentem soluções não-dominadas bem distribuidas pela frente de Pareto (global), facilitando o conhecimento do problema e a escolha da(s) solução(ões) mais adequada(s) pelo decisor.

Considerando que a solução de um MOP é constituída por um conjunto de pontos, a utilização de uma heurística baseada em populações (o que inclui EA, mas também AIS), permite encontrar vários pontos do conjunto ótimo de Pareto em uma única execução do algoritmo [231]. Outro grande potencial destes algoritmos é a integração da ampla exploração do espaço de busca com um processo de busca mais localizado resultando em um alto grau de robustez, que permite sua aplicação a diversos problemas práticos, junto aos quais outras estratégias de solução se mostram inócuas [14].

Além disso, os problemas reais tornaram-se cada vez mais complexos, sem funções definidas, na maioria das vezes descontínuas e com domínios não-convexos, dificultando a utilização de métodos exatos na resolução. Some-se a isto, o fato de heurísticas baseadas em populações serem menos suscetíveis à forma ou à continuidade da frente de Pareto [47].

As técnicas tradicionais exigem um conhecimento prévio do problema, uma especificação detalhada ou uma indicação de preferências, o que quase sempre não é possível. Mas os EA (e de fato Algoritmos Bioinspirados, de uma maneira geral) requerem apenas a descrição aproximada das características que representam o comportamento global, como uma função (ou medida) de afinidade, adaptabilidade ou desempenho [48, 231].

Estas são algumas das razões que têm impulsionado cada vez mais a utilização dos métodos bioinspirados em aplicações, além de sua flexibilidade, generalidade e robustez.

Os EA têm demonstrado bom desempenho na resolução de MOPs e, nos últimos anos, diversas abordagens e algoritmos foram apresentados dentre os quais se destacam: *Vector Evaluated Genetic Algorithm* (VEGA) [204], *Multi-Objective Genetic Algorithm*

(MOGA) [106], *Niched Pareto Genetic Algorithm* (NPGA) [122], *Pareto Archived Evolution Strategy* (PAES) [143], *Nondominated Sorting Genetic Algorithm* (NSGA) [220] e NSGA-II [89], *Streght Pareto Evolutionary Algorithm* (SPEA) [248] e SPEA2 [245], entre outros.

O primeiro Algoritmo Evolutivo Multiobjetivo (MOEA) foi proposto por Schaffer em 1985 [204] e denominado *Vector Evaluated Genetic Algorithm (VEGA)*. É um MOEA no qual a população é dividida em um número de subpopulações igual à quantidade de objetivos que se deseja otimizar. A idéia é fazer com que cada subpopulação seja avaliada com base em um único objetivo, e, posteriormente, mesclar as subpopulações em uma única à qual se aplicam operadores genéticos convencionais (*crossover* e mutação). Este é um algoritmo simples, e não garante a geração de soluções não-dominadas. Como foi observado posteriormente, unir todos os indivíduos das subpopulação a fim de obter uma nova população equivale a combinar linearmente os componentes do *fitness* para obter uma única função de *fitness*, ou seja, equivale a uma solução monobjetivo onde os coeficientes de peso dependem da população atual. Isto significa que, no caso geral, dois indivíduos não-dominados não só serão amostrados com taxas diferentes, como no caso de uma superfície côncava, a população pode se dividir em diferentes espécies, cada qual particularmente forte em um dos objetivos. Schaffer havia antecipado esta propriedade de VEGA e chamou-a de *especiação*. A especiação é indesejada, na medida em que se opõe ao objetivo de encontrar soluções não-dominadas que satisfaçam da melhor maneira possível todos os objetivos ao mesmo tempo [106].

No *Multi-Objective Genetic Algorithm* (MOGA), proposto por Fonseca e Fleming [106], a aptidão de um indivíduo está relacionada à quantidade de indivíduos que o dominam, de maneira que os indivíduos não-dominados têm a mesma aptidão, que é a maior da população.

Nondominated Sorting Genetic Algorithm (NSGA) foi proposto por Srinivas e Deb [220]. Neste algoritmo os indivíduos não-dominados globais (ou seja, não-dominados com relação a toda a população) são os primeiros a serem classificados, sendo-lhes atribuído um valor *falso*. Estes indivíduos são removidos da população para que o processo de classificação dos indivíduos restantes tenha continuidade. NSGA-II [89] é uma versão mais eficiente do algoritmo, na qual foram incorporados um operador de *crowding* e um mecanismo elitista (para preservar as melhores soluções encontradas).

Niched Pareto Genetic Algorithm (NPGA), proposto por Horn et al. [122], usa uma versão modificada de seleção por roleta associada à dominância de Pareto que funciona da seguinte maneira: são selecionados dois pais possíveis e eles são comparados com um subconjunto da população (selecionado de maneira aleatória), aquele pai que seja não-dominado com relação ao subconjunto será o ganhador. Se há um empate (ambos são

dominados, ou não-dominados), o torneio é decidido com base em um critério de *crowding*, o indivíduo com menos vizinhos ganha. Uma versão melhorada, chamada NPGA2, utiliza ordenamento aplicado ao conceito de ótimo de Pareto enquanto a seleção continua sendo feita por meio de sorteio (utilizando um esquema diferente de comparação de aptidão) [101].

Strength Pareto Evolutionary Algorithm (SPEA), proposto por Zitzler e Thiele [244], usa um arquivo externo para armazenar todas as soluções não-dominadas encontradas previamente. Cada indivíduo tem a sua força (*strength*) calculada de maneira semelhante à feita por MOGA, uma vez que depende da quantidade de indivíduos que os dominam. A aptidão de um indivíduo da geração atual é calculada segundo a força das soluções contidas no arquivo externo que o domimam. Para obter uma boa distribuição das soluções dominadas, SPEA usa um método de ligação média (*average linkage method*). A versão melhorada deste algoritmo, SPEA2 [245], possui três principais diferenças com relação à versão original: 1) na atribuição da aptidão são considerados aspectos relacionados não só à quantidade de indivíduos que dominam o indivíduo sendo avaliado, como quantos ele domina; 2) é utilizada uma estimativa de densidade populacional com relação aos vizinhos a fim de deixar a busca mais eficiente; e, 3) é utilizado um método de quebra do arquivo externo, com o objetivo de assegurar que as soluções pertencentes aos extremos da frente de Pareto não se percam.

Pareto Archived Evolution Strategy (PAES), proposto por Knowles e Corne [142, 143] é uma estratégia evolutiva (1+1), ou seja, tem-se uma população de um indivíduo pai que gera apenas um filho por mutação. Usa um arquivo externo para manter as soluções não-dominadas obtidas ao longo do processo evolutivo e, para manter a diversidade, usa um mecanismo de *crowding* baseado em uma divisão recursiva do espaço das funções objetivo, de maneira que os indivíduos são "alocados" em uma rede que facilita determinar e controlar a distribuição das soluções obtidas.

4.1.3 Padrão Ouro dos Algoritmos Evolutivos Multiobjetivo

NSGA-II e SPEA2 são considerados algoritmos que representam o estado da arte em termos de algoritmos MOO [48, 149] e por esse motivo foram escolhidos para constituir o *baseline* de comparação do algoritmo proposto nesta tese.

NSGA-II

NSGA-II (*Nondominated Sorting Genetic Algorithm II*) [89] é uma versão revisada do NSGA [220]. É baseado em camadas de classificação de indivíduos e utiliza seleção

$(\mu + \lambda)$[1], como mecanismo elitista em vez de uma população secundária [47].

O Algoritmo 12 mostra o funcionamento do NSGA-II. A população P_0 de tamanho N é inicializada de maneira aleatória (linha 33) e classificada em categorias (*ranks*) tomando por base a relação de dominância ($\preceq$) (linha 34), pela função $fastNonDominatedSort$ (linhas 13–19). A referida função atribui a todos os indivíduos não-dominados da população um valor de $rank = 1$, significando que pertencem à primeira frente de Pareto (F_1), o que faz com que tenham o mesmo potencial de serem selecionados. Estes individuos são retirados da população e a função $findNonDominatedFront$ (linhas 1–12) é repetida em $(P \setminus F_1)$ para dar continuidade ao processo de classificação dos indivíduos restantes nas respectivas frentes de Pareto (os indivíduos da segunda frente recebem $rank = 2$ e são armazenados em F_2 e assim sucessivamente).

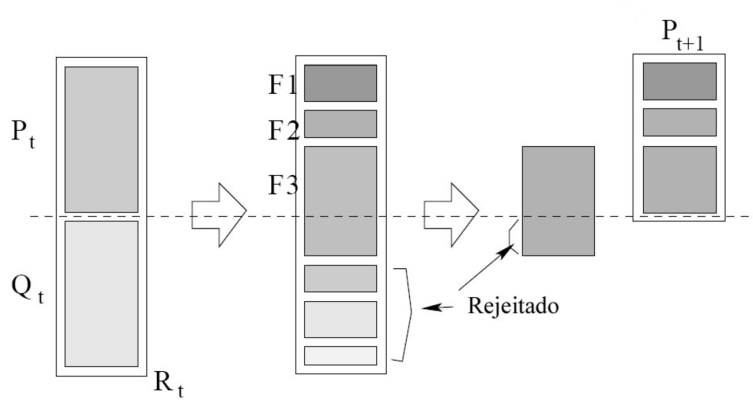


Figura 4.2: Esquema do funcionamento do algoritmo NSGA-II (adaptado de [89]).

Os operadores de seleção, *crossover* e mutação são usados para criar uma população filha Q_t de tamanho N (linha 35).

A união das populações P_t e Q_t dá origem à população R_t, de tamanho $2N$ (linha 37), após o que, R_t é ordenada pela chamada à função $fastNonDominatedSort$.

Dado que todos os indivíduos da população atual são incluídos em R_t, é garantido o elitismo.

Se o tamanho de F_1 é menor do que N, todo os indivíduos de F_1 são selecionados para a nova população P_{t+1} e os demais membros são selecionados das frentes subsequentes na ordem de seus *rankings*, i.e., as soluções de F_2 são escolhidas, seguidas das soluções do conjunto F_3 e assim sucessivamente, até que nenhum outro conjunto F_i inteiro possa ser acomodado (linhas 41–44). Nesse processo é calculado o valor de *crowding* (um operador

[1] $(\mu + \lambda)$ significa que: 1) a população contém μ indivíduos; 2) λ novos indivíduos são criados por operadores de variação (e.g., *crossover*, mutação) em cada iteração; e 3) os melhores μ indivíduos dentre pais e filhos sobrevivem.

Algoritmo 12: NSGA-II

Dados:
Seja F o conjunto de todas as frentes não-dominadas de R_t, tal que $F = (F_1, F_2, ...)$.
Seja $I[i]m$ o valor da função objetivo m para o i-ésimo indivíduo pertencente ao conjunto I.

Function $findNonDominatedFront(P)$ **is**
  $P' \leftarrow \{1\}$ /* inclui o primeiro elemento em P' */
  foreach $p \in P \wedge p \notin P'$ **do**
    $P' \leftarrow P' \cup \{p\}$ /* inclui temporariamente p em P' */
    /* compara p com outros elementos de P' */
    foreach $q \in P' \wedge q \neq p$ **do**
      if $p \preceq q$ **then** /* se p domina um elemento q de P' */
        $P' \leftarrow P' \setminus \{q\}$ /* deleta q de P' */
      else if $q \preceq p$ **then** /* se p é dominado por um elemento de P' */
        $P' \leftarrow P' \setminus \{p\}$ /* não inclui p em P' */
    return P'

Function $fastNonDominatedSort(P)$ **is**
  $i \leftarrow 1$ /* i é o contador do ***rank*** */
  while $P \neq \emptyset$ **do**
    $F_i \leftarrow findNonDominatedFront(P)$
    $P \leftarrow P \setminus F_i$ /* remove as soluções não-dominadas de P */
    $i \leftarrow i + 1$
  return F

Function $crowdingAssignment(I)$ **is**
  $l \leftarrow |I|$ /* número de soluções em I */
  foreach i **do**
    $I[i] \leftarrow 0$
    foreach *objetivo* m **do**
      $I \leftarrow sort(I, m)$ /* ordena usando cada função objetivo */
      /* seleciona os pontos extremos e lhes atribui valor infinito */
      $I[1] \leftarrow \infty$
      $I[l] \leftarrow \infty$
      for $i \leftarrow 2$ **to** $(l - 1)$ **do**
        $I[i] \leftarrow I[i] + (I[i + 1]m - I[i - 1]m)$
  return I

begin
  $t \leftarrow 0$
  $P_t \leftarrow inicializePop()$
  $P_t \leftarrow fastNonDominatedSort(P_t)$
  /* aplica os operadores de ***crossover*** e mutação para criar uma nova população Q_t */
  $Q_t \leftarrow makeNewPop(P_t)$
  while *not(stopCondition())* **do**
    $R_t \leftarrow P_t \cup Q_t$ /* combina a população-pai com a população-filha */
    $F \leftarrow fastNonDominatedSort(R_t)$
    $P_{t+1} \leftarrow \emptyset$
    $i \leftarrow 1$
    /* repete o laço até que a população-pai esteja completa */
    while $|P_{t+1}| + |F_i| \leq N$ **do**
      $crowdingAssignment(F_i)$ /* calcula o ***crowding*** em F_i */
      /* inclui a i-ésima frente não-dominada na população-pai */
      $P_{t+1} \leftarrow P_{t+1} \cup F_i$
      $i \leftarrow i + 1$ /* checa próxima frente para inclusão */
    /* coloca F_i em ordem crescente de frente e decrescente de ***crowding***, utilizando $\prec_n$ */
    $sort(F_i \prec_n)$
    /* escolhe os primeiros $(N - |P_{t+1}|)$ elementos de F_i */
    $P_{t+1} \leftarrow P_{t+1} \cup F_i[1 : (N - |P_{t+1}|)]$
    $Q_{t+1} \leftarrow makeNewPop(P_{t+1})$
    $t \leftarrow t + 1$
  return P_t

de comparação de aglomeração para priorizar os pontos que estão menos aglomerados) (chamada à Procedure *crowdingAssignment*, linha 42).

Na Procedure *crowdingAssignment* (linhas 22–30), para o cômputo do *crowding* é necessário que a população seja ordenada de acordo com o valor dos indivíduos em cada função objetivo em ordem crescente (linha 25). Em seguida, para cada função objetivo (m), as soluções extremas, i.e, com o maior e o menor valores, recebem um valor de *crowding* infinito (linhas 26–27). Todas as demais soluções intermediárias recebem *crowding* igual à diferença absoluta dos valores das duas soluções que lhes são adjacentes (Figura 4.3) (linhas 28–29). O valor final do *crowding* é a soma dos *crowdings* parciais obtidos para cada uma das funções objetivos daquele indivíduo. Uma solução com um valor menor de *crowding* está, de certa forma, *mais rodeada* por outras soluções, i.e., está em uma região mais *povoada*.

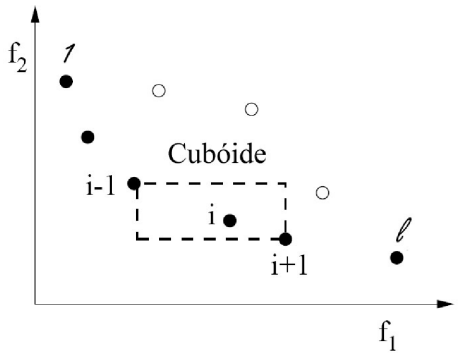


Figura 4.3: Para cada uma das funções objetivo, as soluções são ordenadas e a distância entre dois pontos vizinhos ao ponto de referência i é utilizada para o cálculo do *crowding*. Tal distância é uma estimativa do tamanho do maior cubóide que contém o ponto i, mas que não contém nenhum outro ponto da população. Na frente de Pareto denotada pelos círculos sólidos, o cubóide da i-ésima solução está representado pelo retângulo tracejado. Dados 2 pontos, aquele com menor valor de *crowding* está em uma região mais povoada (tem vizinhos mais próximos a ele) (adaptado de [89]).

Assim, cada indivíduo i da população tem:

1. um *rank* (i_{rank}); e,
2. um valor de *crowding* ($i_{crowding}$).

O *operador de crowding* ($\prec_n$) guia o processo seletivo em direção a uma frente de Pareto mais uniformemente distribuída e pode ser definido como uma ordem parcial onde:

$$i \prec_n j \Leftrightarrow (i_{rank} < j_{rank}) \vee ((i_{rank} = j_{rank}) \wedge (i_{crowding} > j_{crowding})) \tag{4.1}$$

Entre duas soluções com diferentes *ranks*, a preferência recai sobre a que tem o menor *rank* (não-dominada). Entretanto, se ambas soluções pertencem à mesma frente, a preferência é pela que esteja localizada em uma região menos povoada (com maior *crowding*).

Para completar a P_{t+1}, os indivíduos da frente seguinte à última acomodada em P_{t+1} são classificados em ordem crescente de *ranking* e em ordem decrescente de *crowding* (linha 45). Os melhores indivíduos são selecionados (linha 46).

Sobre a população resultante da seleção são aplicados operadores de *crossover* e mutação (linha 47) dando origem à população-filha Q_{t+1} (linha 47).

O processo constante das linhas 36– 48 é repetido até que um critério de parada seja atingido, quando é retornada a população P_t.

NSGA-II com Restrições

O Algoritmo 12 corresponde à implementação padrão de NSGA-II. Para efeito de comparação com o algoritmo proposto foi utilizada a versão de NSGA-II que lida com restrições, proposta por Deb et al. [89].

Na presença de restrições, cada solução é considerada *factível* ou *não-factível*, conforme atenda ou não às restrições, respectivamente.

Desta forma, dadas duas soluções distintas, há três situações possíveis:

1. ambas soluções são factíveis: neste caso a escolha recai sobre a que tem menor *crowding*;
2. uma solução é factível e a outra não: escolhe-se a solução factível;
3. ambas são não-factíveis: opta-se pela solução com a menor violação geral de restrições.

Diante deste novo cenário, é necessária uma pequena modificação na definição de *dominância*, definindo-se a *dominância na presença de restrições*.

Definição 8 Dominância na presença de restrições (*constrained-dominance*) [89]: uma solução i é dita dominar uma solução j (denotada como $i \preceq_{cd} j$) se é verificada qualquer uma das seguintes condições:

1. A solução i é factível e a solução j não;
2. As soluções i e j são ambas não-factíveis, mas a solução i tem, no geral, menos violação de restrições;
3. As soluções i e j são factíveis e a solução i domina[2] a solução j.

O efeito de usar o princípio da *dominância na presença de restrições* é que qualquer solução factível tem um melhor *rank* que qualquer solução não-factível.

Todas as soluções são *ranqueadas* de acordo com seu nível de não-dominância baseada nos valores das funções objetivo. Mas, entre duas soluções não-factíveis, a solução com menor violação de restrições tem um *rank* melhor (menor). Além disso, esta modificação no princípio de dominância não muda a complexidade computacional de NSGA-II.

SPEA2

SPEA2 (*Strength Pareto Evolutionary Algorithm 2*) [242, 245] usa o critério de dominância de Pareto para a atribuição do *fitness* e a seleção de indivíduos. Em cada geração, os indivíduos não-dominados da população atual são copiados em uma população externa, também chamada memória externa ou arquivo.

O Algoritmo 13 mostra o funcionamento do Algoritmo SPEA2. No *Passo 1*, a população P_0 é inicializada de maneira aleatória (linha 30) e a população externa $\overline{P}_0$ é criada vazia (linha 32). No *Passo 2*, é feita a atribuição do *fitness* $F(i)$ para cada indivíduo i (da população atual e do arquivo):

1. é calculada a *força de Pareto* (*Pareto strength*), $S(i)$, que representa o número de indivíduos j que são dominados por i, de maneira que $S(i) = |\{j | j \in (P_t \cup \overline{P}_t) \wedge i \prec j|$;
2. é calculado o *raw fitness*, $R(i)$, somando-se a força de Pareto de todos os indivíduos j que dominam i, de maneira que $R(i) = \sum_{j \in (P_t \cup \overline{P}_t), j \prec i} S(j)$.

Quanto menor o valor de $R(i)$, melhor é o indivíduo – independente do fato do problema ser de minimização ou de maximização (aqui chama-se atenção ao fato que a atribuição do *fitness* no contexto do algoritmo SPEA2 não deve ser confundida com a atribuição do *fitness* no contexto das funções objetivo a serem otimizadas).

Além disso, a informação da densidade populacional é incorporada para discriminar entre indivíduos com mesmo $R(i)$. A densidade utilizada é uma adaptação do algoritmo k-ésimo vizinho mais próximo (*k-nearest neighbor*) (Figura 4.4), onde a densidade de um ponto qualquer é uma função (decrescente) da distância para o k-ésimo ponto mais próximo. Para cada indivíduo i, as distâncias (no espaço de objetivos) para todos os indivíduos $j \in (P_t \cup \overline{P}_t)$ é calculada e armazenada em uma lista, que é colocada em ordem crescente, o k-ésimo elemento dá a distância procurada σ_i^k. É comum usar-se $k = \sqrt{N + \overline{N}}$, mas $k = 1$ geralmente é suficiente e fornece uma implementação mais eficiente [242]. Para o indivíduo i, a densidade é dada por $D(i) = \frac{1}{2+\sigma_i^k}$. No denominador, adiciona-se dois para garantir que $0 < D(i) < 1$. Por fim, o *fitness* $F(i)$ é obtido fazendo-se, $F(i) = R(i) + D(i)$. Torna-se evidente que indivíduos não-dominados terão $F(i) < 1$.

Algoritmo 13: SPEA2

Dados:
N: tamanho da população.
$\overline{N}$: tamanho do arquivo (memória secundária).
Result: A: arquivo de soluções não-dominadas

Function $fitnessAssignment(P)$ **is**
  /* calcula a força de Pareto $S(i)$ e a densidade populacional $D(i)$ dos elementos de P */
  foreach $i \in P$ **do**
    $S(i) \leftarrow 0$
    foreach $j \in P \wedge i \neq j$ **do**
      /* cria uma lista das distâncias de i em relação a cada um do demais indivíduos de P */
      $distance(i,j) \leftarrow norm(i-j)$
      $distance(j,i) \leftarrow distance(i,j)$
      /* calcula a força de Pareto de i */
      if $i \prec j$ **then**
        $S(i) \leftarrow S(i)+1$
    /* ordena $distance(i,:)$, vetor das distâncias de i aos demais elementos de P, em ordem crescente */
    $d \leftarrow sort(distance(i,:))$
    /* σ_i^k recebe a distância de i ao k-ésimo vizinho */
    $\sigma_i^k = d[k]$
    /* calcula a densidade populacional de i */
    $D(i) \leftarrow \frac{1}{2+\sigma_i^k}$
  /* calcula o *raw fitness* $R(i)$ dos elementos de P */
  foreach $i \in P$ **do**
    $R(i) \leftarrow 0$
    foreach $j \in P \wedge i \neq j$ **do**
      if $j \prec i$ **then**
        $R(i) \leftarrow R(i)+S(j)$
  /* calcula o *fitness* $F(i)$ dos elementos de P */
  foreach $i \in P$ **do**
    $F(i) = R(i)+D(i)$
  /* coloca P em ordem crescente de $F(i)$ */
  $P \leftarrow sort(P, F(i))$
  return P

Function $truncate(P,N)$ **is**
  if $|P| > N$ **then**
    foreach $i \in P$ **do**
      foreach $j \in P \wedge i \neq j$ **do**
        if $i \prec_d j$ **then**
          $q \leftarrow i$
    $P \leftarrow truncate((P \setminus q), N)$
  return P

begin
  /* Passo 1: inicialização */
  $t \leftarrow 0$
  $P_t \leftarrow inicializePop()$
  $\overline{P}_t \leftarrow \emptyset$
  repeat
    /* Passo 2: atribuição do fitness */
    $P \leftarrow P_t \cup \overline{P}_t$
    $P \leftarrow fitnessAssignment(P)$
    /* Passo 3: seleção */
    /* copia os indivíduos não-dominados (i.e., com $F(i) < 1$) de $P = P_t \cup \overline{P}_t$ em $\overline{P}_{t+1}$ */
    $i \leftarrow 1$
    while $F(i) < 1$ **do**
      $\overline{P}_{t+1} \leftarrow P[i]$
      $i \leftarrow i+1$
    /* se o arquivo é muito "grande" e "falta" espaço, remove indivíduos por meio da função $truncate$ */
    if $|\overline{P}_{t+1}| > |\overline{N}|$ **then**
      $\overline{P}_{t+1} \leftarrow truncate(\overline{P}_{t+1})$
    /* se o arquivo é muito "pequeno" e "sobra" espaço, copia os melhores $N - |\overline{P}_{t+1}|$ indivíduos dominados de P em $\overline{P}_{t+1}$ */
    else if $|\overline{P}_{t+1}| \leq |\overline{N}|$ **then**
      $\overline{P}_{t+1} \leftarrow \overline{P}_{t+1} \cup (P \setminus \overline{P}_{t+1})[1 : N - |\overline{P}_{t+1}|)]$
    $P^* \leftarrow \overline{P}_{t+1}$
    /* Passo 4: finalização */
    if $stopCondition()$ **then**
      return P^*
      exit
    /* Passo 5: formação do mating pool */
    /* faz seleção por torneio binário com substituição em $\overline{P}_{t+1}$ para criar um *mating pool* */
    $P_{mp} \leftarrow binaryTournmentSelection(\overline{P}_{t+1})$
    /* Passo 6: variação */
    /* aplica os operadores de *crossover* e mutação ao *mating pool* para criar uma nova população atual P_{t+1} */
    $P_{t+1} \leftarrow makeNewPop(P_{mp})$
    $t \leftarrow t+1$
  until

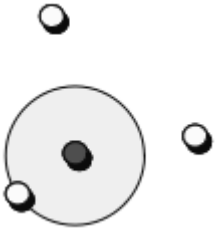


Figura 4.4: k-ésimo vizinho mais próximo. A região circular cinza corresponde à área englobada pela distância ao k-ésimo vizinho mais próximo, onde $k = 1$. Tal informação é utilizada por SPEA2 para o cálculo da densidade $D(i)$, utilizada, por sua vez, no cálculo do *fitness* $F(i)$ (adaptado de [242]).

No *Passo 3*, é feita a seleção dos indivíduos que integrarão a população externa/arquivo. Todos os indivíduos i não-dominados são copiados para o arquivo da próxima geração $\overline{P}_{t+1} = \{i | i \in (P_t \cup \overline{P}_t) \wedge F(i) < 1\}$. Se ($|\overline{P}_{t+1}| = \overline{N}$), este passo está completo, caso contrário, pode haver duas situações:

1. $|\overline{P}_{t+1}| < \overline{N}$: o arquivo é muito pequeno e os melhores $\overline{N} - |\overline{P}_{t+1}|$ indivíduos dominados de $P_t \cup \overline{P}_t$ são copiados no novo arquivo. Isso é feito tomando-se os primeiros $\overline{N} - |\overline{P}_{t+1}|$ indivíduos com $F(i) \geq 1$ do conjunto de soluções $P_t \cup \overline{P}_t$ ordenado por $F(i)$ e adicionando-os a $\overline{P}_{t+1}$;

2. $|\overline{P}_{t+1}| > \overline{N}$: o arquivo é muito grande e uma função *truncate* é chamada para remover iterativamente indivíduos de $\overline{P}_{t+1}$ até que $|\overline{P}_{t+1}| = \overline{N}$. A cada iteração, é escolhido um indivíduo i para remoção, de maneira que $i \prec_d j, \forall j \in \overline{P}_{t+1}$, com:

$$i \prec_d j \Leftrightarrow (\sigma_i^k < \sigma_j^k) \vee ((\sigma_i^l = \sigma_j^l) \wedge (\sigma_i^k < \sigma_j^k)) \tag{4.2}$$

   onde σ_i^k é a distância de i para o k-ésimo vizinho mais próximo em $\overline{P}_{t+1}$, i.e., a cada estágio, é escolhido o indivíduo com a menor distância com relação a outro indivíduo, se há empate, ele é resolvido considerando a segunda menor distância e assim por diante (Figure 4.5).

Neste momento (*Passo 4*), é verificada a condição de terminação do algoritmo, se ela foi atingida, o algoritmo retorna P^* que é o arquivo contendo os indivíduos não-dominados que resolvem o problema e a execução do algoritmo é finalizada, caso contrário, segue para o *Passo 5*, que corresponde à formação do *mating pool*, i.e., a população selecionada por torneio binário em $\overline{P}_{t+1}$ e sobre a qual serão aplicados, no *Passo 6*, os operadores de *crossover* e mutação, resultando na população P_{t+1}. O contador de gerações é atualizado e o processo retorna ao *Passo 2*.

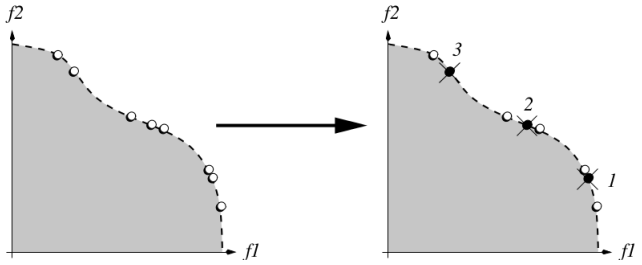


Figura 4.5: Representação esquemática do funcionamento da função *truncate*. À esquerda: conjunto de soluções não-dominadas. À direita: considerando que $N = 5$, soluções removidas por *truncate* e sua ordem de remoção [242].

4.2 Sistemas Imunológicos Artificiais

4.2.1 Conceitos Básicos

O *Sistema Imunológico (SI)* dos vertebrados é um complexo de células, moléculas e órgãos que representam um mecanismo de identificação capaz de perceber e combater disfunções das próprias células do organismo e a ação de microrganismos infecciosos exógenos.

A interação entre o SI e outros sistemas e órgãos permite a regulação do corpo, garantindo a manutenção de seu estado de equilíbrio e seu funcionamento estável [130].

A função do SI é geralmente associada à proteção contra patógenos e doenças, mas ele tem sido implicado em um leque muito maior de atividades, incluindo a formação de tecido cicatricial, a eliminação de células anormais ou que sofreram injúria, o crescimento de novos vasos sanguíneos, a regeneração de tecidos do corpo, e a limpeza de debris[3] intercelulares [49].

Sua complexidade já foi comparada à do cérebro em diversos aspectos: é capaz de reconhecer sinais externos e internos; controlar a ação de seus componentes; influenciar o comportamento de outros sistemas, como o nervoso e o endócrino; e mais importante, aprender como combater agentes causadores de doenças e extrair informações deles [75].

Padrões moleculares presentes nos patógenos são chamados *antígenos* e podem ser reconhecidos pelo SI provocando sua resposta.

A identificação dos materiais estranhos ocorre por intermédio de dois sistemas (mecanismos de resposta) que integram o SI [130]:

1. o *sistema imunológico inato*: também conhecido como resposta imune inata ou inespecífica, assim chamada porque o organismo nasce com a habilidade de reconhecer certos microrganismos e imediatamente destruí-los. Ao longo da evolução

[3]Membrana ou filamento que se forma entre duas superfícies serosas, geralmente após um processo inflamatório [1].

dos vertebrados, o SI desenvolveu a capacidade de identificar estruturas fortemente conservadas, responsáveis pela sobrevivência de patógenos, de maneira que diferentes agentes patogênicos da mesma classe (e.g., fungos, vírus, bactérias, parasitas), vão expressar essas estruturas. As células do SI inato estão prontamente disponíveis para o combate contra uma grande variedade de patógenos, sem que seja necessária uma exposição prévia a eles, podendo, portanto, destruir muitos microrganismos no primeiro encontro; além disso, tal reação (como é inata) vai ocorrer da mesma maneira em todos os indivíduos normais [3]; e

2. o *sistema imunológico adaptativo*: também conhecido como *resposta imune específica*, é requisitada quando há uma infeção não controlada pela resposta imune inata. Ele é capaz de reconhecer os mesmos estímulos antigênicos numa futura reexposição do organismo ao antígeno, o que permite ao SI melhorar-se a cada encontro com um determinado antígeno.

Enquanto a resposta imunológica adaptativa resulta em imunidade contra reinfecção no decorrer da vida de um indivíduo, a resposta imune inata segue constante, independente da exposição a antígenos.

A separação conceitual entre sistema imunológico inato e adaptativo não significa que ambos operam de maneira independente, de fato, há uma grande iteração entre as células de ambos os sistemas. Ambos dependem da atividade de *células brancas do sangue*, os *leucócitos*.

Dentre os leucócitos, as células mais importantes são os *linfócitos*, que mediam a resposta imunológica adaptativa, responsáveis pelo *reconhecimento* e *eliminação* de agentes patogênicos. Os linfócitos reagem a condições do ambiente em que habitam e que são por ele percebidas por meio de *receptores*, estruturas complexas de proteína ligadas a sua membrana (Figura 4.6). A ligação de um receptor a uma estrutura a qual ele tem *afinidade* induz reações químicas ou mudanças na transcrição do DNA dentro da célula, alterando seu comportamento ou composição. Os linfócitos também são capazes de desenvolver uma *memória* imunológica, ou seja, são capazes de reconhecer os mesmos estímulos antigênicos quando estes são apresentados novamente ao organismo. Há dois tipos de linfócitos: B e T. Os linfócitos B, também chamados de *células B* têm sua origem na medula óssea, os linfócitos T (ou *células T*), originam-se do timo.

Cada linfócito possui apenas um tipo de receptor específico, capaz de reconhecer um *epítopo* ou *determinante antigênico* (menor porção de antígeno com potencial de gerar a resposta imune) com especificidade distinta para cada receptor, o número de linfócitos que pode se conectar a um antígeno é restrito (Figura 4.7).

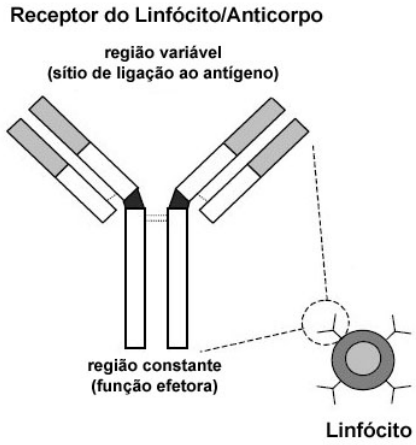


Figura 4.6: Estrutura esquemática de um receptor/anticorpo. Células B e T possuem receptores em sua superfície, entretanto, apenas células B têm a capacidade de secretar anticorpos. Anticorpos são a forma solúvel (capaz de circular no organismo) dos receptores de antígenos das células B. Os dois braços da molécula em forma de Y contêm a região variável (em cinza) que se liga de maneira específica ao antígeno. O receptor se liga à célula B ou T por meio da região constante, ou, no caso de anticorpos, esta é a região que se liga ao mecanismo efetor que é ativado pelo anticorpo para eliminar o patógeno (adaptado de [78]).

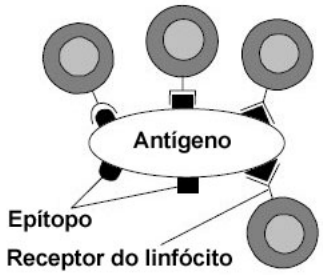


Figura 4.7: Reconhecimento de antígeno por receptores de linfócitos (adaptado de [80]).

As células T têm sua ação iniciada quando epítopos lhes são apresentados por moléculas especializadas integrantes do *Complexo Principal de Histocompatibilidade (Major Histocompatibility Complex–MHC)* e não serão aqui tratadas.

Se uma célula B encontra um antígeno com uma afinidade suficiente, ela prolifera e se diferencia em *células efetoras* e *de memória*, em um processo chamado *seleção clonal*. As células efetoras já estão preparadas para atuar na defesa do organismo enquanto as células de memória ficam de prontidão, à espera de que haja nova invasão pelo mesmo antígeno.

O *Princípio da Seleção Clonal* (ou *Princípio da Expansão Clonal*) é a teoria usada para descrever as propriedades básicas de uma resposta imune adaptativa a um estímulo

antigênico.

A área de pesquisa em Sistemas Imunológicos Artificiais (*Artificial Immune Systems–AIS*) é extensa. Existem diversos modelos tais como: seleção clonal, redes imunes, seleção positiva, seleção negativa. A escolha dos melhores modelos depende dos objetivos e das características do problema a ser estudado. Devido principalmente à capacidade de variação genética, seleção e adaptação dos linfócitos B, o algoritmo empregado neste trabalho baseia-se no princípio da seleção clonal que será explanado em mais detalhes a seguir.

4.2.2 Seleção Clonal

O reconhecimento antigênico é o primeiro requisito para a ativação de uma resposta imunológica. O receptor da célula imune tem que reconhecer o antígeno com uma certa *afinidade* (uma ligação entre o receptor e o antígeno ocorre com uma força proporcional a esta afinidade). Se a afinidade é maior que um limite inferior, o SI é ativado [79].

Apenas as células capazes de reconhecer um estímulo antigênico proliferarão e se diferenciarão em células efetoras, sendo então, selecionadas em lugar das demais. A maior diferença entre a expansão clonal dos linfócitos B e T é que os linfócitos B sofrem mutação somática durante a reprodução e os linfócitos B efetores são células que secretam anticorpos ativamente, enquanto linfócitos T não sofrem mutação durante sua reprodução e são principalmente secretores de linfocinas (moléculas envolvidas na emissão de sinais entre células durante a respostas imunes) ou se diferenciam em linfócitos que realizam fagocitose ativamente. A presença de eventos de mutação e seleção no processo de expansão dos linfócitos B permite que estes linfócitos aumentem sua diversidade e também se tornem progressivamente mais capazes de reconhecer seletivamente antígenos.

O mecanismo da seleção clonal pode ser visto na Figura 4.8. Quando receptores antigênicos que estão em um linfócito B conectam-se a um antígeno (em presença de um sinal estimulatório), isto faz com que o antígeno ative o linfócito B. A estimulação dos linfócitos B povoca sua proliferação e diferenciação em células terminais (que não se dividem) secretoras de anticorpos, chamadas *células plasmáticas*. Cada célula B secreta apenas um tipo de anticorpo que é relativamente específico para o antígeno (propriedade chamada *monoespecificidade*). Enquanto as células plasmáticas são as células mais ativas quanto à secreção de anticorpos, os linfócitos B em divisão também secretam anticorpos, mas em uma taxa menor. Os linfócitos B, além de proliferarem e se diferenciarem em células plasmáticas, podem diferenciar-se em *células de memória* de longa vida [78, 130].

A proliferação no caso de células imunes se dá por mitoses somáticas sucessivas, um processo de reprodução assexuada onde a célula se divide generando clones, não há *crossover*.

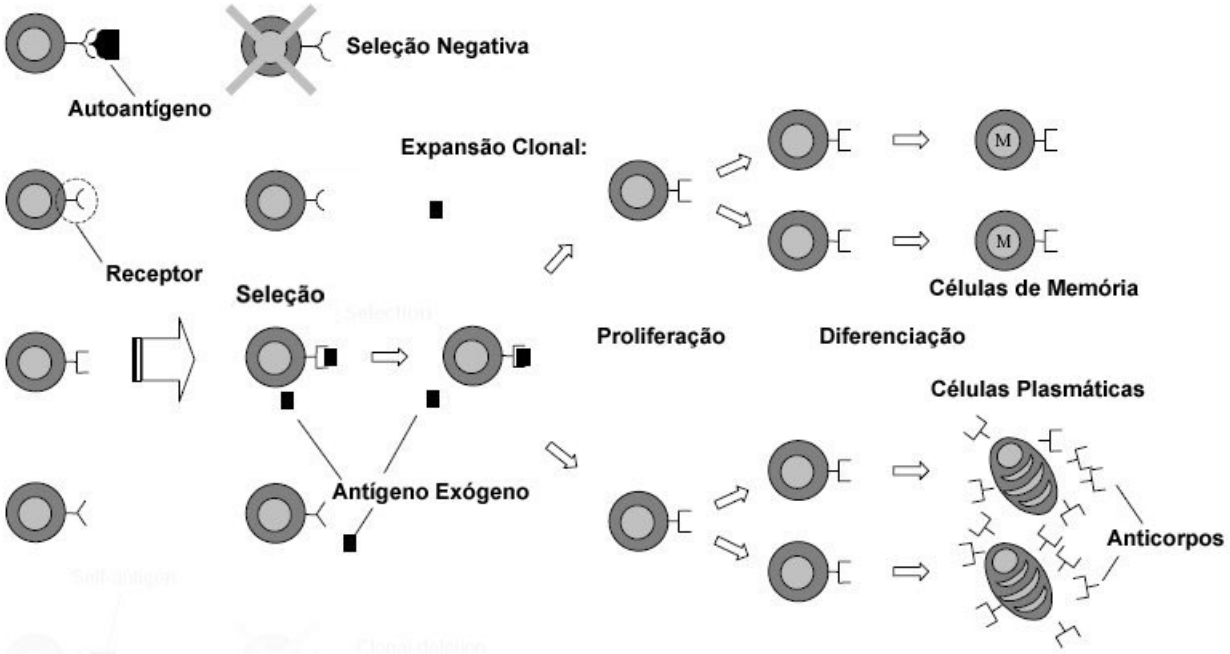


Figura 4.8: Seleção Clonal (adaptado de [78]).

Enquanto a evolução biológica adaptativa ocorre por seleção natural entre organismos, a pesquisa em imunologia forneceu a primera evidência clara de que mudanças adaptativas ontogenéticas – mudanças observadas desde a origem e desenvolvimento de um organismo, do embrião até sua plena forma desenvolvida – podem ser alcançadas por meio de variação e seleção em um mesmo organismo [82].

Durante a reprodução responsável por introduzir mudanças aleatórias nos genes que determinam o receptor do anticorpo, ocasionalmente tais mudanças podem levar a um aumento da afinidade do anticorpo. Este processo, junto a uma forte pressão seletiva, resulta em células com receptores antigênicos com maior afinidade em relação ao antígeno em questão. As variações de alta-afinidade são selecionadas para entrar no conjunto de células de memória. Todo o processo de seleção e mutação é conhecido como *maturação da resposta imune* e é análogo à seleção natural de espécies [79].

Células de memória circulam pelo organismo e provavelmente não produzem anticorpos. Entretanto, quando expostas a um segundo estímulo antigênico, rapidamente começam a diferenciar-se em células plasmáticas capazes de produzir anticorpos de alta afinidade [71, 82], sendo *pré-selecionadas* por antígenos específicos que estimularão a *resposta primária* (resposta obtida no primeiro contato com o antígeno, necessária para que ocorra a sensibilização do sistema imune e para que a memória seja gerada) (Figura 4.9). A teoria também propõe que linfócitos auto-reativos (que reagem contra alvos existentes no próprio indivíduo – autoantígenos) são removidos do repertório antes de sua maturação (Figura 4.8).

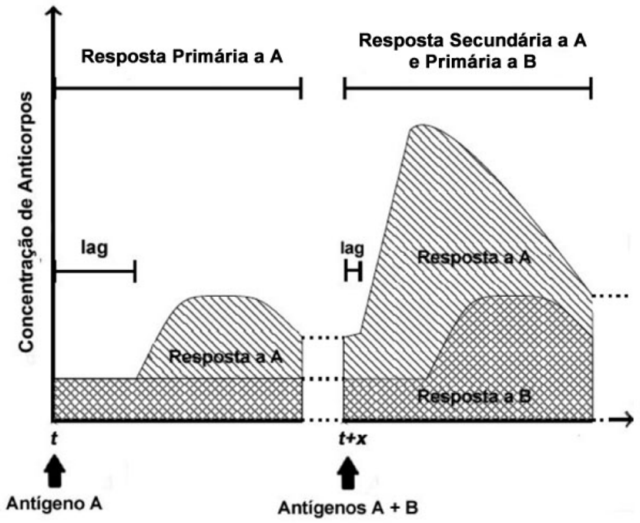


Figura 4.9: Uma típica resposta imunológica adaptativa mediada por anticorpos. O primeiro encontro com um antígeno produz uma *resposta primária*. O antígeno A é introduzido no tempo $t = zero$, encontrando poucos (ou nenhum) anticorpos específicos. Após uma fase *lag* (onde está em ação principalmente o SI inato), aparecem anticorpos contra o antígeno A; sua concentração cresce até um platô, a partir do qual declina. Quando é introduzida uma mistura de antígenos A e B, observa-se uma rápida e intensa resposta a A, evidenciando a *resposta secundária* a este antígeno e a *memória imunológica* (a habilidade do SI em responder a uma segunda exposição ao mesmo antígeno de maneira mais eficiente e efetiva, conferindo ao indivíduo uma defesa específica contra a infecção). A resposta ao antígeno B assemelha-se à resposta primária a A, uma vez que é o primeiro encontro com o antígeno B. A partir do tempo $t+x$, a comparação da resposta imunológica ao antígeno A com a resposta ao antígeno B evidencia também a especificidade da resposta imunológica (adaptado de [130]).

O repertório não só é diversificado pelo processo de hipermutação, como existem mecanismos nos quais linfócitos B raros, com receptores mutados com alta afinidade podem ser selecionados para dominar a resposta.

Dada a natureza aleatória do processo de mutação, uma larga proporção de genes mutantes tornam-se não funcionais ou possivelmente desenvolvem especificidades prejudiciais às células do próprio organismo. As células com receptores de baixa afinidade e as células auto-reativas são eliminadas [169]. Isso é necessário para que não contribuam significativamente para o conjunto de células de memória.

Durante o processo de diferenciação, um mecanismo de hipermutação opera com uma taxa próxima de 10^{-3} por par de bases (bp) da região variável, por geração. Já que o tamanho de combinação destas regiões variáveis está em 700 bp, em média, uma mutação a cada divisão de células será introduzida [78]. Essa média tem como resultado o fato de que todos clones de uma única célula podem ser similares, mas jamais idênticos, a seus pais. Assim, a diversidade no repertório imunológico é mantida.

Mutações pontuais são boas para explorar regiões locais, enquanto que a edição de receptores pode resgatar a resposta imunológica que esteja *presa* em um ótimo local. A edição de receptores e as mutações pontuais têm papéis complementares no processo de maturação de afinidades [82]. É de se esperar que qualquer clone de alta afinidade desenvolvido por hipermutação somática ou edição de receptores fosse preferencialmente expandido, mas o que se observa é que algumas poucas células de baixa afinidade também entram no repertório, mantendo, desta forma, a diversidade populacional. Além disso, a fração de células novas que vêm da medula óssea também é acrescentada ao conjunto de linfócitos com o mesmo objetivo de manter a diversidade.

Um rápido acúmulo de mutações é necessário para uma rápida maturação da resposta imunológica. Entretanto, a maior parte das mudanças levará a uma coleção de anticorpos débeis ou não-funcionais. Se uma célula que conseguiu uma mutação útil continua mutando a uma mesma taxa durante a próxima resposta imunológica, então a acumulação de mudanças deletérias pode causar a perda das vantagens da mutação. Assim, um surto de hipermutação somática seguida de uma pausa para permitir a seleção e expansão clonal pode ser a base do processo de maturação.

As principais propriedades da teoria da seleção clonal são:

1. *monoespecificidade*: restrição que determina um padrão fenotípico para cada célula diferenciada e retenção deste padrão pelos descendentes/clones;

2. *expansão clonal*: proliferação e diferenciação de linfócitos em linfócitos maduros ao entrar em contato com antígenos exógenos ao corpo. Um antígeno provoca a proliferação de linfócitos B diferentes (cada um é capaz de reconhecer uma porção distinta do patógeno e com uma afinidade diferente), a taxa de proliferação de cada linfócito é proporcional a sua afinidade com o antígeno seletor, quanto maior a afinidade, maior o número de clones gerados;

3. *hipermutação somática*: geração de mudanças genéticas aleatórias por uma forma de mutação acelerada, expressa como diferentes padrões de anticorpos. Estas mutações experimentadas pelos clones são inversamente proporcionais à afinidade do anticorpo com o antígeno (clones com afinidade maior sofrem menor mutação e vice-versa);

4. *seleção negativa*: eliminação de linfócitos reativos recém-diferenciados com padrões antigênicos representados pelos próprios componentes, chamados *autoantígenos*;

5. *autoimunidade*: o conceito de um *clone proibido* resistente à eliminação prematura por autoantígenos como a base de *doenças autoimunes*.

Enquanto o número total de linfócitos no SI é regulado, aumentos nos tamanhos de alguns clones significa que outros clones devem diminuir em quantidade. Entretanto, o número total de linfócitos não é mantido absolutamente constante [82].

Se o SI *aprende* aumentando o tamanho da população de linfócitos específicos, ele deve: 1) esquecer antígenos previamente aprendidos; 2) aumentar de tamanho; ou 3) constantemente reduzir a porção de seu repertório que é gerada aleatoriamente e que é responsável por responder a novos antígenos [82].

É importante destacar que com um número limitado de células e moléculas, o SI é capaz de detectar um número quase ilimitado de antígenos. Como destaca de Castro [72], isto se deve ao fato de que cada antígeno tem um número de características que lhe permite ser reconhecido por mais de um receptor. Além disso, apesar de todos os receptores de um linfócito terem a mesma especificidade (reconhecerem o mesmo tipo de antígeno), a diversidade de receptores é grande no SI (os receptores são gerados por recombinação aleatória de segmentos de DNA).

4.2.3 Métodos Computacionais

Segundo Dasgupta [67], os Sistemas Imunológicos Artificiais (AIS) surgiram da tentativa de modelar e aplicar metodologias inteligentes inspiradas no sistema imunológico biológico à solução de problemas do mundo real.

É possível listar e discutir diversas propriedades do SI que são bastante interessantes sob uma perspectiva computacional: reconhecimento de padrões, qualidade de ser único, auto-identidade, diversidade, disponibilidade, autonomia, ser multi-camadas, cobertura dinamicamente mutável, detecção distribuída, tolerância ao ruído, resiliência, tolerância à falha, robustez, aprendizagem e memória, padrão de resposta predador-presa, auto-organização [67, 78, 108, 218].

Os AIS apresentam, em sua estrutura básica, as principais características requeridas para a resolução de MOPs: elitismo, diversidade, memória, repertório dinâmico, mutação e clonagem proporcionais à afinidade. Estas características inerentes fazem com que o método tenha vantagem em relação a outras estratégias populacionais.

Sob a perspectiva da otimização multiobjetivo, são interessantes:

1. *Detecção distribuída*: os anticorpos devem estar dispersos no espaço de busca, sendo cada um responsável pela exploração de uma região, evitando a aglomeração (centralização) de soluções. O objetivo do processo de otimização é determinar múltiplos ótimos (soluções não-dominadas) dentro do repertório de anticorpos. Neste caso, o interessante é que todo repertório de anticorpos (ou a maior parte dele) seja selecionado para clonagem, sendo cada anticorpo analisado localmente, resultando em

um algoritmo capaz de executar uma busca por diversas soluções distintas. Além disso, soluções ótimas locais são preservadas;

2. *Diversidade*: o espaço de busca tem infinitas possibilidades que devem ser exploradas. Um Algoritmo MOO deve apresentar um conjunto de soluções diversas, de maneira que toda a Frente de Pareto (ou a maior parte dela) possa ser obtida. Os AIS são inerentemente capazes de manter a diversidade da população. Num algoritmo genético clássico, por exemplo, seria necessário algum mecanismo para a manutenção de diversidade;

3. *Memória*: muitos algoritmos bioinspirados para MOO trabalham com uma memória externa elitista, para que não ocorra a perda do melhor indivíduo (solução não-dominada). Esta já é uma característica instrínseca ao SI, pois no processo de seleção clonal, é a melhor célula (entre anticorpo-pai e clones) que sobrevive para a próxima iteração. O processo de seleção clonal é elitista, logo, o AIS não apresenta efeito retrógrado em suas iterações.

Outras características dos AIS, interessantes para resolução de problemas de otimização, podem ser ressaltadas:

1. A amplitude de mutação sofrida pelos clones é inversamente proporcional à afinidade do anticorpo-pai, de maneira que, quanto maior a afinidade, menor será a mutação;

2. Uma vez que os antígenos tenham sido eliminados do organismo, existirá uma quantidade excedente de anticorpos e o sistema deve voltar ao seu nível normal. É por isso que algumas células são eliminadas (auto-regulação), mas algumas se conservam, circulando pelo organismo como células de memória. O nível normal de anticorpos no repertório não se manterá necessariamente constante, muito pelo contrário, o tamanho do repertório é dinâmico. Em otimização, o objetivo é que o tamanho seja estabelecido de acordo com a necessidade do problema.

Alguns autores propuseram que um algoritmo genético sem *crossover* seria um modelo razoável de seleção clonal, mas o algoritmo genético padrão não leva em consideração aspectos importantes do AIS, como reprodução e mutação proporcionais à afinidade [79].

Frente a isso, de Castro e Von Zuben [80, 81] propuseram um algoritmo de seleção clonal, chamado CLONALG, inicialmente utilizado para o reconhecimento de padrões, mas que serviu como modelo para os algoritmos inspirados em seleção clonal utilizados em outras aplicações.

Dado um conjunto de padrões a serem reconhecidos **Ag** (antígenos), os passos básicos de CLONALG (Algoritmo 14) são:

1. Gerar um conjunto de soluções candidatas (inicializar aleatoriamente uma população de indivíduos **Ab** (anticorpos) que corresponde ao repertório de células e moléculas imunes);
2. Para cada padrão em **Ag**, apresentá-lo à população **Ab** e determinar sua afinidade com cada elemento da população **Ab**;
3. Selecionar os n melhores indivíduos de **Ab** com base em suas afinidades com o antígeno correspondente;
4. Reproduzir (clonar) estes indivíduos proporcionalmente a suas afinidades com os antígenos (quanto maior a afinidade, maior o número de cópias). Esta corresponde à fase de expansão clonal;
5. Mutar todas as cópias a uma taxa inversamente proporcional à afinidade (quanto maior a afinidade, menor a taxa de mutação). Esta corresponde à fase de mutação somática;
6. Acrescer os indivíduos mutados à população **Ab** e re-selecionar n indivíduos maduros (otimizados) para serem mantidos como memória do sistema (para permanecer no sistema);
7. Repetir os pasos desde 2 até que um certo critério seja atingido (e.g., um número mínimo de gerações, ou seja alcançada uma taxa de erro mínimo ou um número predeterminado de avaliações).

Este algoritmo permite ao AIS tornar-se cada vez melhor em sua tarefa: baseado em um comportamento evolutivo, CLONALG "aprende" a reconhecer o antígeno.

De Castro e Von Zuben [72] destacam que de fato este é um novo tipo de algoritmo evolutivo inspirado no SI que envolve três principais processos evolutivos: reprodução, variação genética e seleção. Ainda que apresente similaridades com estratégias evolutivas e técnicas de algoritmos genéticos, a sequência de passos não é a mesma e o desempenho é qualitativamente diferente.

Em que pese a quantidade de pesquisa realizada nos últimos anos em AIS, esta ainda pode ser considerada uma área nova, aberta a contribuições. Os primeiros AIS foram: *Clonal Selection Algorithm* (CLONALG) [82], *Artificial Immune Network for Optimization* (opt-aiNet) [77, 84]. Citamos, em seguida, algoritmos AIS Multiobjetivos (Tabela 4.1): *Multiobjective Immune System Algorithm* (MISA) [47, 55], *Multiobjective lmmune Algorithm* (MOIA) [151], *Multiobjective Clonal Selection Algorithm* (MOCSA) [32], *Vector Artificial Immune System* (VIS) [110], *Population Adaptative Based Immune Algorithm* (PAIA) [37], *Artificial Immune Network for Omni-Optimization* (omni-aiNet) [46, 91], *Multiobjective Bayesian Artificial Immune System* (MOBAIS) [34, 35].

Algoritmo 14: CLONALG

Dados: **Ag**: conjunto de antígenos (problema a ser resolvido).
Resultado: **Ab**: população de células imunes capazes de reconhecer **Ag**.

```
1  begin
2      Initialização: inicializar aleatoriamente um repertório Ab (população) de
       células imunes
3      while not(critério de parada) do
4          for each antígeno ∈ Ag do
5              Seleção: selecionar as células com maior afinidade em relação ao
               antígeno;
6              Reprodução/Clonagem: gerar cópias das células imunes de maneira
               diretamente proporcional a sua afinidade, i.e, quanto maior a afinidade,
               mais clones são gerados;
7              Variação genética: mutar cada célula de maneira inversamente
               proporcional a sua afinidade, i.e., quanto maior a afinidade, menor a
               taxa de mutação;
8              Avaliação da afinidade: avaliar a afinidade de cada célula mutada com
               relação ao antígeno;
9          Retornar a população Ab ao seu tamanho normal;
10     return Ab
```

O algoritmo *Artificial Immune Network for Optimization* (opt-aiNet) [77, 84] é um AIS monobjetivo que pode ser considerado uma extensão do CLONALG, contemplando princípios de redes imunes (passos envolvendo interações anticorpo-anticorpo e não apenas interações antígeno-anticorpo). Possuiu um ciclo de maturação por afinidade que é repetido até atingir-se a estabilidade, medida pela diferença entre a afinidade média da população atual com a anterior. O algoritmo inicia com uma população gerada aleatoriamente, o critério de parada é guiado pela variação do tamanho da população, sendo o processo interrompido quando o tamanho da população não varia entre duas iterações consecutivas. A cada iteração, é gerado um número fixo de clones de cada anticorpo da rede, assim, diferente de CLONALG, a quantidade de clones não é proporcional à afinidade do anticorpo. Posteriormente, ocorre a hipermutação, esta sim, aplicando taxas inversamente proporcionais à afinidade de cada clone. Os clones de um anticorpo formam uma subpopulação onde o melhor anticorpo é selecionado (caso sua afinidade seja maior, ele substitui o anticorpo que o originou). Quando a estabilidade é alcançada, ocorre uma etapa de supressão, com o objetivo de eliminar a redundância de anticorpos na população.

Multiobjective Immune System Algorithm (MISA) [47, 55] foi de fato o primeiro AIS para otimização multiobjetivo, proposto por Cortés e Coello Coello. Baseia-se no princípio da seleção clonal e utiliza uma população secundária para armazenar soluções não-dominadas encontradas ao longo da busca. MISA tem sido aplicado a diversos problemas

Tabela 4.1: Algoritmos AIS Multiobjetivo e respectivos métodos.

	Método	
Algoritmo	*Seleção Clonal*	*Redes Imunes*
MISA [34, 55]	x	
MOIA [151]	x	
MOCSA [32]	x	
VIS [110]	x	x
PAIA [37]	x	x
omni-aiNet [46, 91]		x
MOBAIS [34, 35]	x	

multiobjetivos com resultados competitivos [210]. O algoritmo proposto neste trabalho é livremente baseado em MISA.

Multiobjective Immune Algorithm (MOIA) [151] também é inspirado no princípio da seleção clonal. As funções objetivo são avaliadas e com base nisso é definido um *rank* que mede a dominância. Anticorpos não-dominados são escolhidos para mutação, após o que, soluções não-dominadas são armazenadas na memória. Supressão e geração de novos anticorpos são passos complexos em MOIA.

Multiobjective Clonal Selection Algorithm (MOCSA) [32] é a versão multiobjetivo com codificação real (não-binária) de CLONALG. Anticorpos são avaliados por meio das funções objetivo e classificados com base na dominância em sucessivas frentes (*ranks*). São gerados clones de toda a população numa quantidade inversamente proporcional ao *rank* que o anticorpo ocupa. Uma perturbação aleatória é adicionada aos clones a fim de explorar o espaço de busca. Toda a população (original e mutada) é avaliada e classificada novamente em *ranks*, neste momento, é aplicada supressão para retornar a população ao seu tamanho original. A cada geração uma porcentagem de anticorpos gerados aleatoriamente substitui anticorpos com baixa afinidade visando a manutenção da diversidade da população.

Vector Artificial Immune System (VIS) [110] é uma extensão do algoritmo opt-aiNet para problemas multiobjetivos baseado em seleção clonal e teoria de redes imunes. VIS tem dois laços de repetição aninhados: o estágio de seleção clonal é repetido um certo número de vezes antes de avaliar as interações (afinidades) na rede imune. Anticorpos são classificados segundo a quantidade dos demais anticorpos que os dominam, formando um *rank*. O mecanismo de seleção não considera toda a população, mas o melhor anticorpo considerando-se pais e descendentes. Soluções não-dominadas são copiadas para a memória. Assim como MOCSA, após um número fixo de iterações, novas soluções são geradas aleatoriamente para manter a diversidade da população.

Population Adaptative Based Immune Algorithm (PAIA) [37] também tem inspiração

em princípios de seleção clonal e teoria de redes imunes. Nele, é simulada a variação da concentração de anticorpos no sangue por meio da utilização de uma população principal de tamanho variável. A manutenção dos anticorpos não-dominados é feita na população dinamicamente adaptada. A população inicial é gerada aleatoriamente, a afinidade é calculada com base na distância do anticorpo no espaço de busca e é definida para garantir que anticorpos não-dominados sempre tenham a menor afinidade. Ao índice definido até então, associa-se informação acerca das afinidades entre anticorpo-anticorpo e anticorpo-antígeno. O conjunto de clones é gerado com base no índice de afinidade, de maneira que um número variável de clones é gerado, mas tomando como base um limite máximo permitido de clones. Os anticorpos não-selecionados são clonados apenas uma vez. A mutação corresponde a uma perturbação aleatória com desvio padrão proporcional à afinidade dos pais. Ao final da iteração, é aplicada supressão aos anticorpos para os quais a distância com relação a outros anticorpos é maior que um valor predefinido.

Artificial Immune Network for Omni-Optimization (omni-aiNet) [46, 91] é um algoritmo baseado na teoria de redes imunológicas, capaz de resolver otimizações mono e multiobjetivo, com restrições ou sem restrições. Omni-aiNet trabalha com uma população codificada com valores reais (não-binários) e incorpora duas técnicas de variação genética: mutação polinomial (mecanismo que conduz a hipermutação tomando por base uma taxa de mutação inversamente proporcional à "avidez", uma medida calculada de acordo com um índice de dominância) e duplicação gênica (geração aleatória de uma coordenada – um valor para uma das funções objetivo – que substituirá uma coordenada da solução caso verifique-se que, com esta substituição, há melhora no anticorpo). O processo de seleção baseia-se em um *grid* para seleção de anticorpos com uma boa distribuição no espaço de objetivos, contribuindo para a diversidade de soluções na população. O mecanismo de supressão tem por base a distância euclideana no espaço de variáveis aplicando torneio binário para os anticorpos muito próximos. As regras para o torneio binário são baseada na violação de restrições. Novos anticorpos são introduzidos a cada iteração. O tamanho da população é dinamicamente ajustado durante a evolução do algoritmo.

Multiobjective Bayesian Artificial Immune System (MOBAIS) [34, 35] estende, no campo da otimização multiobjetivo, o trabalho anterior *Bayesian Artificial Immune System* (BAIS), de de Castro [36]. MOBAIS substitui os operadores de clonagem e hipermutação pela aplicação de um modelo probabilístico capaz de identificar, ou aprender, as relações entre as variáveis dos melhores anticorpos da população corrente. O modelo aplicado é o de redes bayesianas. Após a etapa de aprendizado, as informações obtidas a partir da rede bayesiana são usadas para montar novas soluções. A etapa de supressão é feita com base na avaliação da afinidade (*fitness*) do anticorpo (soluções com *fitness* baixo são descartadas) e na supressão de anticorpos similares, onde a similaridade é medida pela

distância euclideana.

4.2.4 Operadores

O comportamento de uma AIS é ditado, em geral, pelo seguintes operadores, cujas variantes mais comuns são descritas a seguir.

1. ***Operador de clonagem***: gera uma nova população (Pc) de cópias dos indivíduos da população atual (P). Os operadores de clonagem mais comuns são:

   (a) *operador de clonagem estática* [59]: clona cada Ab N vezes, produzindo uma população intermediária Pc de tamanho $|Pc| = |P| \times N$;

   (b) *operador de clonagem proporcional* [82, 83]: clona os Ab's proporcionalmente a sua afinidade antigênica;

   (c) *clonagem probabilística* [60]: Ab's são escolhidos na população atual P dependendo de uma taxa de seleção clonal p_c.

2. ***Operador de hipermutação***: este operador atua na população atual de clones (Pc), aplicando a cada indivíduo, uma quantidade M de mutações, determinada por um processo aleatório [58]. Os mais comuns são [62]:

   (a) *hipermutação estática*: o número de mutações é independente da função *fitness* F, assim, cada Ab, a cada geração, sofre M mutações.

   (b) *hipermutação proporcional*: o número de mutações é proporcional ao valor da função *fitness*.

   (c) *hipermutação inversamente proporcional*: o número de mutações é inversamente proporcional ao valor da função *fitness*.

   (d) *hipermutação convexa*: cada posição x_i (cada *gene*) do Ab, $\overrightarrow{x}$, executa uma hipermutação dependente de uma taxa de hipermutação p_h.

   (e) *hipermutação somática contínua (somatic contiguous hypermutation – CHM)* [136, 137]: também conhecida por hipermutação de regiões contínuas (*contiguous region hypermutation operator – CRH*) [44] e hipermacromutação [62]. O número de mutações independe da função *fitness* e do parâmetro M. Neste caso, dois inteiros i e j, tais que $(i+1) \leq j \leq l$ são escolhidos aleatoriamente e o operador muta no máximo $M_m(\overrightarrow{x}) = j - i + 1$ valores no intervalo $[i, j]$. Apesar de evidências de um mecanismo biológico semelhante na literatura imunológica serem esparsas [197], elas indicam que dentro das células, mutações ocorrem em *clusters* de regiões, o que seria análogo a uma região contínua.

3. ***Operador de idade (aging)***: este operador elimina indivíduos "velhos", para tanto, utiliza o parâmetro $\tau_B > 0$ que determina o número máximo de gerações que um Ab pode permanecer na população.

   (a) *aging estático*: quando o Ab tem idade $\tau_B + 1$ ele é eliminado da população atual, independentemente do *fitness* que ele possa ter. Durante a expansão clonal, um Ab herda a idade de seu pai. Após a fase de hipermutação, apenas os Ab's clonados que tenham aumentado seu *fitness* receberão $age = 0$. Obtém-se uma versão elitista deste operador atribuindo-se ao(s) melhor(es) Ab('s) da população, a cada geração, $age = 0$.

   (b) *aging aleatório*: a probabilidade que o Ab permaneça na população atual (P_{viver}) é governada pela seguinte regra com parâmetro τ_B (expectativa de vida média do Ab):

$$P_{viver}(\tau_B) = e^{-\frac{ln(2)}{\tau_B}} \Leftrightarrow P_{morrer}(\tau_B) = 1 - e^{-\frac{ln(2)}{\tau_B}} \tag{4.3}$$

   Uma versão elitista deste operador é obtida atribuindo-se, a cada geração, $P_{viver} = 1 \Rightarrow P_{morrer} = 0$, ao melhor indivíduo na população.

4. ***Operador de nascimento***: este operador substitui os piores indivíduos da população com novos indivíduos gerados aleatoriamente (fase de nascimento). Pode apresentar variantes como *não-redundância*, por meio do qual evita-se que cópias do mesmo indivíduo sobrevivam para a próxima geração.

4.2.5 Convergência dos Algoritmos AIS

Em que pese a convergência de Algoritmos AIS (em especial os baseados em seleção clonal) ter sido demonstrada em diferentes trabalhos, principalmente utilizando cadeias de Markov [44, 45, 231, 232], aqui será feita uma abordagem utilizando analogia com a prova de convergência para EA. Assim, da mesma forma que os EA se beneficiaram de resultados e técnicas da bem-estabelecida teoria de processos estocásticos [95], proceder-se-á com os algoritmos AIS: inicialmente serão apresentadas as condições consideradas suficientes para a convergência de um EA [242], em seguida, por analogia, tais condições serão estendidas para os algoritmos AIS, demostrando-se, assim, sua convergência.

Esta seção foi baseada em Cutello et al. [62].

Diz-se que um EA *converge* para um ótimo global de um problema de otimização se é possível assegurar que o algoritmo encontra a solução em um número finito de passos e se tal solução é mantida nas populações subsequentes.

Como os estados de transição de uma EA têm natureza estocástica, não se pode usar um conceito determinístico de convergência para definir o comportamento relativo ao tempo limite deste tipo de algoritmo, para tanto, utiliza-se a *convergência completa* [62], uma medida de convergência estocástica.

Definição 9 Convergência completa (I): Seja X uma variável aleatória e $(X_t : t > 0)$ uma sequência de variáveis aleatórias. Então diz-se que a sequência X_t *convergiu completamente* para X se, para qualquer $\epsilon > 0$,

$$\lim_{t\to\infty} \sum_{i=0}^{t} P(|X_i - X| > \epsilon) < \infty$$

Assim, pode-se estabelecer a seguinte definição para a convergência de um EA:

Definição 10 Convergência completa (II): Seja $X_t : t \geq 0$ a sequência de populações gerada pelo EA e seja F_t o *fitness* do melhor indivíduo da população no tempo t. Diz-se que um EA *convergiu completamente* para um ótimo global f^* do problema de otimização definido pela função $f : X \to \Re$, se a sequência aleatória não negativa $D_t = f^* - F_t$ converge completamente para zero.

Uma única iteração de um EA geral pode ser descrita como segue:

$$\forall i \in \{1, ..., n\} : x_i^{'} = mut(rec(reprod(x_1, ..., x_n)))$$

$$(y_1, ..., y_n) = sel(x_{\pi(1)}, ..., x_{\pi(q)}, x_1, ..., x_n, x_1^{'}, ..., x_n^{'})$$

Onde $(x_{\pi(1)}, ..., x_{\pi(q)}) \in \chi^q$ é a população de indivíduos gerada pelo operador de nascimento, $(x_1, ..., x_n) \in \chi^n$ é a população de pais atual, $(x_1', ..., x_n') \in \chi^n$ é a população de filhos, *reprod*, *rec*, *mut* e *sel* são funções representando respectivamente operadores de reprodução, recombinação (*crossover*), mutação e seleção.

Sob as seguintes condições, um EA converge completamente para o ótimo global de qualquer problema de otimização:

***Condição 1**: Todo indivíduo na população pode ser tranformado em um outro indivíduo qualquer em um único passo de mutação com probabilidade $p > 0$.*

$$\forall x, y \in \chi \qquad P\{y = mut(x)\} \geq \delta_m > 0 \tag{4.4}$$

***Condição 2**: O melhor indivíduo da população sobrevive, em cada geração, com probabilidade $p = 1$*

$$P\{v_m^*(sel(x_{\pi(1)}, ..., x_{\pi(q)}, x_1, ..., x_n, x_1^{'}, ..., x_n^{'})) = v_n^*(y_1, ..., y_n)\} = 1 \tag{4.5}$$

Onde v_i^* retorna o melhor indivíduo da população de i indivíduos.

Se apenas a condição 1 é válida, pode ser provado que o EA visita o ótimo global após um número finito de passos com probabilidade $p = 1$, independente de sua inicialização, mas não é possível assegurar sua convergência, uma vez que não se pode garantir que o ótimo permanece na população após ter sido encontrado. Entretanto, se a condição 2 também é válida, pode-se provar que o EA converge para o ótimo global.

Analogamente um algoritmo AIS pode ser descrito como segue:

$$\forall i \in \{1, ..., n\} : x_i^{'} = hip(clone(x_1, ..., x_n))$$

$$(x_1^{''}, ..., x_k^{''}) = aging(x_1, ..., x_n, x_1^{'}, ..., x_n^{'})$$

$$(y_i, ..., y_n) = sel(x_1^{''}, ..., x_k^{''})$$

Onde $(x_1, ..., x_n) \in \chi^n$ é a população de anticorpos/células B atual e *clone*, *hip*, *aging*, *sel* são funções representando respectivamente operadores de clonagem, hipermutação, *aging* e seleção.

Aqui, o operador de *aging* pode ou não ser usado (assim como nem todos os EAs usam *crossover*), e k pode ser maior ou menor que n.

Examinando-se o Algoritmo 15 (no Capítulo 5), percebe-se que o operador de *aging* na verdade toma parte no mecanismo de seleção do processo evolutivo, uma vez que define se o indivíduo sobreviverá para a próxima geração de acordo com sua idade, o que significa que, enquanto a condição 1 pode ser aplicada ao AIS apenas considerando o operador de hipermutação, o operador de *aging* precisa ser considerado na descrição formal da condição 2, assim, considerando-se os algoritmos AIS, a condição 2 pode ser corretamente descrita como segue:

$$P\{v_m^*(aging(x_1, ..., x_n, x_1^{'}, ..., x_k^{'})) = v_k^*(x_1^{''}, ..., x_k^{''})\} = 1$$

$$P\{v_k^*(sel(x_1^{''}, ..., x_k^{''})) = v_n^*(y_1, ..., y_n)\} = 1$$

Teorema 2. *O algoritmo AIS coverge completamente para o ótimo global de um problema de otimização, independentemente de sua inicialização, desde que um operador elitista (e.g., de aging) seja aplicado.*

Demonstração. Para provar esse teorema, é necessário mostrar que ambas condições 1 e 2 são satisfeitas pelo algoritmo AIS.

Uma vez que nem a clonagem nem o *aging* modificam os indivíduos existentes ou criam indivíduos diferentes, apenas dois operadores podem ser responsáveis pela introdução, pela primeira vez, de ótimos na população: os operadores de hipermutação e/ou de nascimento.

Considerando uma cadeia de caracteres de comprimento λ, com cada ponto do espaço de busca representado por um vetor $\{0,1\}^\lambda$. Se um indivíduo da população comparado à cadeia de caracteres ótima tem similaridade $\gamma - c$ bits, sendo a divergência igual a c bits, a probabilidade do operador de hipermutação atingir o ótimo global em um passo é:

$$P_c^{(\gamma)} = \frac{c!}{\gamma^c}\frac{1}{\gamma} \tag{4.6}$$

Onde as escolhas favoráveis $c!$ são diferentes permutações dos c elementos considerando-se as γ^c possíveis escolhas. Esta probabilidade precisa ser multiplicada pela probabilidade do operador (que é aleatório) mudar c bits. $\frac{1}{\gamma}$ é a probabilidade que r seja igual a c $(r = c)$, onde r é o número de bits escolhido aleatoriamente para ser mutado.

A Equação 4.6 pode ser estendida para alfabetos com cardinalidade K, com cada ponto do espaço de busca representado por um vetor $\{0, 1, 2, ..., K-1\}^\gamma$, a probabilidade de que cada um dos dígitos seja realmnete mutado para o correspondente dígito da cadeia de caracteres ótima é:

$$P_c(\gamma) = \frac{c!}{\gamma^c}\frac{1}{(K-1)^c}\frac{1}{\gamma} \tag{4.7}$$

Onde $\frac{1}{(K-1)^c}$ é a probabilidade de um único dígito mutar para o valor correto.

Como $P_c(\gamma)$ é sempre positivo, prova-se a *condição 1.* Tem-se também a probabilidade $P_s > 0$ que o ótimo seja aleatoriamente introduzido na população pelo operador de nascimento, que deve ser também considerado, apesar de $P_c(\gamma)$ ser suficiente para provar a condição 1.

Para provar que a *condição 2* também é válida, é necessário levar em consideração todos os operadores atuando na população e mostrar que nenhum deles será responsável pela perda da solução ótima uma vez que ela tenha sido encontrada.

O operador de clonagem cria cópias de indivíduos, mas não os modifica, assim, por sua ação o ótimo não pode ser perdido.

O operador de hipermutação apenas age nos indivíduos da população Pc, introduzidos pelo operador de clonagem, mas não modifica indivíduos criados por qualquer outro operador (inclusive ele mesmo).

O operador de *aging* elimina indivíduos velhos mas à melhor solução candidata em cada geração é dada $age = 0$ (i.e. $P_{morrer} = 0$), desta forma, é impossível que o operador *aging* perca o ótimo, a menos que ele delete indivíduos com $age = 0$, o que por definição não ocorre.

Por fim, o operador de seleção elimina os indivíduos com pior *fitness*, de maneira que o ótimo não corre o risco de ser perdido. Isto é suficiente para garantir a *condição 2*, do que segue a prova do teorema.

□

Desta forma, a convergência de um grande elenco de algoritmos AIS pode ser provada, contanto que um operador (geralmente de hipermutação ou de nascimento) satisfaça a condição 1 e que se possa mostrar que a condição 2 também é válida, o que é feito geralmente pela verificação do elitismo no operador de *aging* e/ou de seleção.

Teorema 3. *O algoritmo AIS não converge completamente para o ótimo global de um problema de otimização, independentemente de sua inicialização, se um operador de hipermutação muta apenas bits distintos e uma variante não elitista do operador de aging é aplicada.*

Demonstração. No Teorema 2 a condição 1 foi provada, contanto que o algoritmo AIS use um operador (e.g., hipermutação ou nascimento) capaz de alcançar qualquer outro indivíduo possível no espaço de busca. Isto assegura que o algoritmo "visitará" o ótimo. Para provar a convergência, a condição 2 também deve ser satisfeita.

Para provar que o algoritmo AIS não converge para o ótimo global, é suficiente provar que sempre que o ótimo tenha sido encontrado, existe uma geração subsequente na qual o algoritmo não apresenta o ótimo em sua população.

Se há uma probabilidade, não importa o quão baixa ela seja, de que o algoritmo AIS possa perder o ótimo sem ter encontrado uma outra solução ótima no meio tempo, então ele não converge com probabilidade 1 (e muito menos completamente). Isso é garantido pelo não-elitismo do operador de *aging* (por simplicidade, considerar-se-á o *aging* estático), uma vez que quando o indivíduo representando o ótimo atinge uma determinada idade ele será eliminado da população.

Num determinado tempo t, dada uma população $P(t)$ consistindo de X_1 indivíduos ótimos e $X_2 = P(t) - X_1$ indivíduos não-ótimos. No tempo $t + \tau_B$, todas as X_1 soluções ótimas terão sido removidas da população.

A prova consiste em mostrar que há sempre uma probabilidade positiva que todas as soluções ótimas sejam removidas da população antes que uma nova seja introduzida, quaisquer que sejam as quantidades de indivíduos ótimos e não-ótimos. O operador de clonagem criará uma população de clones consistindo de cópias das X_1 soluções ótimas e das X_2 soluções não-ótimas. Sejam m_{X_2} os clones de X_2. Uma vez que o operador de hipermutação sempre mudará pelo menos um bit, os clones das soluções ótimas não produzirão ótimos globais. Por outro lado, cada clone m_{X_2} pode se tornar um ótimo global no próximo passo com probabilidade p, tal que $p \leq p_{d=1}$ é a probabilidade de atingir-se o ótimo a partir da posição mais favorável, que é a distância $d = 1$. Assim, a probabilidade que um deles se torne o ótimo global é maior que $(1 - p_{d=1})^{m_{X_2}}$. Esta probabilidade é minimizada quando m_{X_2} é maximizado. Isso ocorre quando há apenas um ótimo global na população $P(t)$ (i.e. $X_1 = 1$ e $X_2 = n - 1$).

Se um novo ótimo global não foi gerado, o operador de seleção levará, no máximo, todos os X_1 indivíduos ótimos para a próxima geração (e sua idade será incrementada em 1), juntamente com os indivíduos não ótimos. Além disso, podem ser introduzidos novos indivíduos aleatoriamente criados. Assim, um limite inferior para a probabilidade P_{NO} do operador de hipermutação não gerar o ótimo global antes que os ótimos globais na população atual sejam perdidos é:

$$P_{NO} \geq (1 - p_{d=1})^{m_{n-1}\tau_B} > 0 \tag{4.8}$$

A probabilidade do operador de nascimento também não introduzir aleatoriamente um ótimo global em τ_B gerações deve ser considerado juntamente com a probabilidade acima descrita, apesar de se esperar que para funções não triviais a probabilidade P_{NO} seja muito pequena. Portanto, a soluções ótimas serão em algum momento perdidas (i.e., em um tempo infinito).

□

Capítulo 5

MAIS: Multi-Objective Artificial Immune System Algorithm

A proposta de um algoritmo AIS multiobjetivo para lidar com SCP foi motivada por resultados obtidos em estudo cotejando este tipo de algoritmo comparado a algoritmos que representam o estado da arte em MOO para resolver o problema de cobertura de antenas [210], tal estudo demonstrou que AIS obteve soluções melhores não só em termos quantitativos (de otimização absoluta, i.e., os resultados estavam mais próximo à origem dos eixos), mas também em termos qualitativos (na medida em que tais soluções encontravam-se distribuídas de maneira mais regular no espaço de soluções).

Neste capítulo, após algumas definições iniciais introduzidas à Seção 5.1, o Algoritmo *Multi-Objective Artificial Immune System* (MAIS), proposto nesta tese, é apresentado em detalhes na Seção 5.2.

5.1 Definições Iniciais

No contexto dos algoritmos AIS, o antígeno (Ag) é o problema a ser resolvido, o anticorpo (Ab) é a solução gerada. No início da *resposta primária* o problema/antígeno é reconhecido por soluções candidatas "pobres". No final da resposta imunológica o problema/antígeno é resolvido/atacado por "boas" soluções candidatas.

Formalmente, Ag é o conjunto de variáveis que modela o problema, e células B[1] são definidas como *strings binárias de tamanho finito*, i.e., $Ab = \mathbb{B}^l$, onde l é o comprimento da *string* e $\mathbb{B} = \{0, 1\}$. A entrada é o antígeno/problema, a saída é basicamente o anticorpo/solução candidata que reconhece/resolve Ag.

Em um algoritmo AIS genérico (Algoritmo 15), por $P(t)$ define-se a população de d anticorpos de comprimento l, que representa o subconjunto do espaço de soluções possíveis

[1] Aqui não se fará distinção entre anticorpo Ab e a célula B que o produz.

de tamanho l obtida no tempo t. A população inicial de anticorpos, $P(0)$, é criada aleatoriamente (linhas 2–3). Após a inicialização, há três fases diferentes.

A *fase de iteração* onde a população $P(t)$ é avaliada (linha 5). A *função de avaliação (função fitness)* que fornece o valor do anticorpo ($Ab = \overrightarrow{x}$) é $F_i(\overrightarrow{x})$ (Equação 3.1), e indica quão bom Ab é em reconhecer o Ag, i.e., a capacidade da solução candidata em resolver o problema.

A *fase de expansão clonal* é composta de dois passos: *clonagem* (linha 6) e *hipermutação* (linha 7). O operador de clonagem produz a população Pc. O operador de mutação escolhe aleatoriamente posições em $\overrightarrow{x}$ e muda seu valor de 0 para 1 e vice-versa. A função de hipermutação aplicada à população Pc gera a população $Phyp$. O mecanismo de mutação do anticorpo é modelado pelo número de mutações, que é inversamente proporcional ao valor da função *fitness*. A fase de expansão clonal desencadeia o crescimento de uma nova população de anticorpos com alta afinidade (melhor valor da função *fitness*).

Na *fase de amadurecimento* (*aging phase*), após a avaliação de $Phyp$ no tempo t (linha 8), o algoritmo elimina os anticorpos "velhos" (linha 9). Tal processo é regulado pelo *operador de idade* (*operador de aging*).

Os melhores anticorpos são então selecionados para integrar a população seguinte (linha 10), ao passo que, dependendo da implementação, pode-se substituir os piores indivíduos da população com novos indivíduos gerados aleatoriamente (fase de nascimento). A *resposta imune adaptativa* (linhas 5–11) prossegue até que um critério de parada seja atingido (linha 4).

Algoritmo 15: Algoritmo AIS genérico

Resultado: **P(t)**: população de Ab capazes de reconhecer Ag

```
1  begin
2      t ← 0
3      P(t) ← initializePop()
4      while not(stopCondition()) do
5          evaluate(P(t))
6          Pc ← clone(P(t))
7          Phyp ← hypermutate(Pc)
8          evaluate(Phyp)
9          (P(t)_a, Phyp_a) ← aging(P(t), Phyp, τ_B)
10         P(t+1) ← select(P(t)_a, Phyp_a)
11         t ← t+1
```

5.2 MAIS

Conforme citado, não se fará distinção entre linfócito B e seu receptor (anticorpo–Ab), assim, cada elemento do AIS será chamado genericamente Ab.

No contexto do algoritmo empregado, não há uma população de antígenos explícita a ser reconhecida, mas as funções objetivo a serem minimizadas. Assim, a afinidade de um anticorpo corresponde à mensuração das funções objetivo: cada anticorpo representa um elemento do espaço de entradas.

É utilizada uma população secundária (população externa ou memória) como mecanismo elitista para manter as melhores soluções encontradas ao longo do processo. Os anticorpos selecionados para serem "células de memória" (para integrar a população externa) são todos não-dominados entre si e, além disso, são não-dominados em relação a todos os anticorpos que em algum momento da evolução tentaram entrar na população secundária (o que simula uma versão elitista do operador de *aging*). Desta forma, na memória está armazenada a aproximação à verdadeira frente de Pareto do problema.

O algoritmo AIS Multiobjetivo proposto neste trabalho (Algoritmo 16) é baseado no princípio da seleção clonal e inspirado em *Clonal Selection Algorithm (CLONALG)* [72, 80, 81] e *Multiobjetive Immune System Algorithm (MISA)* [47].

Algoritmo 16: MAIS

Data: **Ag** – conjunto de antígenos
Result: Pm – população de **Ab**/células imunes capazes de reconhecer **Ag**

```
1  begin
2      P ← generateNewAb()
3      Pm ← ∅
4      while not(stopCondition()) do
5          evaluate(P)
6          Psel ← select(P)
7          Pm ← updateMemory(Pm, Psel)
8          Pc ← clone(Psel)
9          Phyp ← hypermutate(Pc)
10         Pmut ← mutate(P \ Psel)
11         P ← P ∪ Phyp ∪ Pmut
12         if (numberOfGenerations mod X) = 0 then
13             Pnew ← generateNewAb()
14             P ← P ∪ Pnew
15         return2OriginalSize(P)
16     Pm ← updateMemory(Pm, P)
17     return Pm
```

Para a resolução do problema, admite-se que não são conhecidas informações de preferência entre os objetivos, de maneira que o algoritmo deve gerar alternativas para a posterior escolha do decisor. Busca-se um conjunto de soluções que esteja o mais próximo possível do conjunto ótimo de Pareto e que tenha a maior diversidade possível no espaço de objetivos (respeitadas as restrições, se houver) [86]. Em outras palavras, o método de resolução deve guiar a busca pela frente de Pareto global e manter a diversidade das soluções não-dominadas atuais [86].

Na tentativa de se obter uma distribuição uniforme e diversa de soluções não-dominadas na frente de Pareto do problema o *grid adaptativo* proposto por [142] foi implementado e aplicado à população secundária.

No Grid Adaptativo, o espaço das funções objetivos é dividido segundo o número de subdivisões determinado pelo usuário, os extremos do *grid* correspondem aos valores máximos e mínimos disponíveis no momento para cada função objetivo. Como esses valores variam conforme a população muda dinamicamente, o *grid* é dito "adaptativo".

Uma vez determinados os limites do *grid*, para cada indivíduo da população é calculada sua posição no *grid* (qual célula ele ocupa). Idealmente, o tamanho da população seria infinita. Entretanto, diante da inviabilidade do caso ideal, é estabelecido um limite para a capacidade de armazenamento das soluções não-dominadas na memória secundária (que corresponde ao tamanho da memória secundária). Alcançado este limite, se há anticorpos não-dominados tentando entrar na população secundária, sua entrada é regulada por um critério adicional: a densidade da região. Indivíduos pertencentes a regiões menos povoadas têm preferência (Algoritmo 16, linha 7).

A manutenção da diversidade se dá no espaço de objetivos, assim, não é permitida a coexistência, na memória, de soluções que são geneticamente diferentes mas que dão origem a valores iguais para as funções objetivo (fenótipos iguais). O *grid* também leva em consideração apenas o espaço de objetivos.

Os passos de MAIS (Algoritmo 16) são:

1. Gerar a população inicial aleatoriamente (linha 2).
2. Inicializar a população secundária (memória), que neste momento deve estar vazia (linha 3).
3. Para cada indivíduo da população determinar (linha 5):
   (a) Se o indivíduo é factível ou não (aqui são avaliadas as funções objetivo e as restrições, se houver).
   (b) Dentro de uma mesma classe:

i. Determinar se o indivíduo é não-dominado ou dominado (dominância é determinada apenas entre indivíduos de uma mesma classe, ou seja, um indivíduo factível é comparado com outro também factível).

4. Determinar os melhores anticorpos, que serão clonados (linha 6):

   (a) Selecionar todos os indivíduos factíveis não-dominados. Se são menos que $d\%$ da população (foi utilizado $d = 10\%$), selecionar indivíduos adicionais até completar os $d\%$ da seguinte maneira:

      i. Factíveis dominados (preferência: dominados por um número menor de indivíduos);

      ii. Não-factíveis não-dominados (preferência: menor quantidade de violações de restrições);

      iii. Não-factíveis dominados (preferência: dominados por um menor número de indivíduos).

   Na prática, os dois últimos casos normalmente não são necessários, mas são incluídos por completude do método.

5. Copiar os melhores anticorpos obtidos no passo anterior na população secundária (memória) (linha 7). O ingresso na memória é regulado utilizando-se o *grid adaptativo* [142, 143, 144]. Para cada um dos anticorpos selecionados no passo anterior, é verificada a relação de dominância quanto aos que já estão na memória:

   (a) Se o anticorpo é dominado por qualquer anticorpo já presente na memória, ele é descartado (não lhe é permitido entrar na população secundária);

   (b) Caso contrário, todos os anticorpos pertencentes à memória que são dominados pelo novo anticorpo são eliminados e verifica-se a possibilidade de ingresso do novo indivíduo:

      i. Se a população secundária não está cheia, é permitida a entrada do novo anticorpo;

      ii. Caso contrário, é determinada qual a célula mais povoada e a posição que o novo anticorpo ocuparia no *grid*:

         A. Se ele pertence à região mais povoada, não lhe é permitido entrar;

         B. Caso contrário, ele ingressa na memória, mas um indivíduo é eliminado da célula mais povoada para que haja uma posição livre para o novo anticorpo (e seja mantido o tamanho da população secundária).

6. Determinar para cada "melhor" indivíduo selecionado no passo 4, o número de clones que se deseja criar usando os seguintes critérios (linha 8):

(a) Primeiro é determinado o número total de clones a serem produzidos por todos os indivíduos selecionados. O número total recomendado por Coello Coello e Cortéz [47] (empiricamente determinado) é de 600% do total da população (ou seja, 6 vezes o tamanho da população), mas utilizou-se 400%, que se mostrou um valor adequado para o problema SCP.

Esta quantidade total de clones é distribuída igualmente entre todos os indivíduos selecionados.

(b) A quantidade inicial de clones a serem produzidos para cada indivíduo é alterada com base nas seguintes regras:

i. Quando a população secundária não está cheia, é calculada a distância euclideana para cada um dos indivíduos selecionados em relação aos demais indivíduos também selecionados. Estas distâncias são tomadas no espaço de objetivos. Depois, é calculada a distância euclideana média em relação a todos os indivíduos selecionados para clonagem ($DEM_{general}$) – este valor será usado como referência – e a distância euclideana média para cada um dos anticorpos selecionados ($DEM_{individuo}$), é também calculada:

A. Se $DEM_{individuo} < DEM_{general}$, ele está em uma região bastante povoada (pertence a uma região cuja população está acima da média populacional) e o número de clones (determinado no passo anterior) é diminuído em 50%;

B. De maneira semelhante, se pertence a uma região com população abaixo da média ($DEM_{individuo} < DEM_{general}$), o número de clones é aumentado em 50%;

C. Se o indivíduo tem $DEM_{individuo} = DEM_{general}$, o número de clones permanece inalterado.

ii. Quando a população secundária está cheia, há 3 possíveis situações:

A. Ao indivíduo a ser adicionado à população secundária não é permitida a entrada, seja porque já pertence a ela, seja porque pertence à região mais povoada do espaço das funções objetivo: então o número do clones criado para este anticorpo é zero;

B. Quando o indivíduo pertence a uma célula cujo número de soluções está abaixo da média populacional (em relação a todas as células ocupadas na população secundária, e não mais somente em relação aos indivíduos selecionados, o número de clones a ser gerado é duplicado);

C. Quando o indivíduo pertence a uma célula cujo número de soluções está acima da média (também em relação à memória secundária), o

número de clones é reduzido à metade.
A média é usada como um limite (gatilho) para regular o número de clones a serem produzidos.

7. Clonar os melhores anticorpos baseados na informação obtida no passo anterior

8. Hipermutar os clones com base na afinidade (linha 9):

   (a) Indivíduos no *ranking* principal (determinado com base nos critérios do passo 4 de factibilidade e dominância): mutados em k posições (em geral, utiliza-se k igual ao número de objetivos, sendo que as posições a serem mutadas são escolhidas aleatoriamente);

   (b) Quando se move para baixo em uma hierarquia:

      i. Factíveis dominados: k somado ao número de indivíduos que dominam o indivíduo em questão;
      ii. Não-factíveis não-dominados: $k+$ *quantidade de violações*;
      iii. Não-factíveis dominados: o número de indivíduos que dominam o indivíduo em questão somado a $k+$ *quantidade de violações*;

   É importante destacar que se está utilizando uma medida de afinidade antigênica correspondente à dominância de Pareto e factibilidade de um indivíduo, ainda que não se tenha explicitamente uma população de antígenos a ser reconhecida, mas um conjunto de funções objetivo a serem otimizadas e um conjunto de restrições a serem satisfeitas. A taxa de hipermutação para cada anticorpo é inversamente proporcional ao valor de afinidade correspondente aos *ranks* descritos no passo 4 (quanto maior a afinidade[2], menor a mutação e vice-versa).

9. Aplica-se uma mutação uniforme aos anticorpos da população que não foram selecionados no passo 4 (linha 10). A porcentagem no início da execução do algoritmo é $pm = 60\%$ e ao final $pm = \frac{1}{L}$, onde L é a longitude de uma cadeia do anticorpo. Os decrementos em pm são feitos de maneira uniforme ao longo da execução. Verificou-se empiricamente que com este passo houve melhora no resultado final das soluções fornecidas pelo algoritmo, no caso do problema SCP.

[2] A afinidade é determinada pela posição na hierarquia:

1º) Factíveis não-dominados;
2º) Factíveis dominados;
3º) Não-factíveis não-dominados;
4º) Não-factíveis dominados.

Quanto mais acima na hierarquia, maior a afinidade.

10. A cada conjunto de X gerações (foi utilizado $X = 20$), um certo número de anticorpos novos (uma porcentagem do tamanho da população principal) é gerado e adicionado à população principal (como forma de gerar diversidade, realizar exploração do espaço de objetivos e resgatar a busca de possíveis máximos locais) (linhas 12–14). Se os indivíduos acrescidos apresentarem melhor afinidade que os já existentes na população, serão mantidos, caso contrário, serão eliminados no próximo passo.

11. Retornar a população ao tamanho original (selecionar tantos indivíduos quanto o tamanho necessário, usando o critério de hierarquias descrito no passo 4) (linha 15).

12. Repetir o processo desde o passo 3 enquanto a condição de parada não for cumprida (e.g., um número predeterminado de execuções do laço de repetição (gerações), ou de avaliações da função *fitness*[3]).

[3]A função *fitness*, neste contexto, corresponde ao conjunto das funções objetivo do problema.

Capítulo 6

Estudo de Caso

Como estudo de caso para nosso método, foi usado um conjunto de dados da espécie vegetal *Dipteryx alata* (o baru), cujas características biológicas serão abordadas na Seção 6.1. As características do dados coletados serão tratadas na Seção 6.2. Ao lidar com a conservação da biodiversidade, o problema SCP pode assumir várias formas, neste estudo de caso, o problema é apresentado na Seção 6.3. As Seções 6.4, 6.5 e 6.6 abordam, respectivamente, o modelo nulo, a ferramenta *Spartan* e as métricas utilizadas. Tais elementos serão empregados no Capítulo 7 para a análise dos dados obtidos.

6.1 O Baru

A espécie *Dipteryx alata* Vog. é uma leguminosa arbórea da família *Fabaceae*, conhecida por diversos nomes populares tais como baru, barujó, castanha-de-burro, castanha-de-ferro, coco-feijão, cumaru-da-folha-grande, cumaru-verdadeiro, cumaru-roxo, cumaru-ana, cumbary, emburena-brava, feijão-coco, fruta-de-macaco, meriparagé, pau-cumaru [96] (Figura 6.1). Ocorre nas matas, cerrados e cerradões do Brasil Central, envolvendo terras dos estados de Mato Grosso, Mato Grosso do Sul, Goiás, Minas Gerais e Distrito Federal, ocorrendo também, em menor frequência, nos estados do Maranhão, Tocantins, Pará, Rondônia, Bahia, Piauí e norte de São Paulo [33].

O baru é encontrado em terras férteis e seus ecossistemas naturais têm sido massivamente desmatados. Em função da procura pela madeira (para fabricação de carvão vegetal, instalação de cercas, indústria moveleira, construção civil) e pelo nível de desmatamento do Cerrado (em especial, em função do avanço da fronteira agropecuária), o baru está ameaçado de extinção.

O barueiro é uma árvore de grande porte, chega a medir 25 metros de altura, podendo atingir 70 cm de diâmetro. Tem vida útil em torno de 60 anos. Exibe copa densa e arredondada, possui crescimento rápido, sendo importante para fixação de carbono da

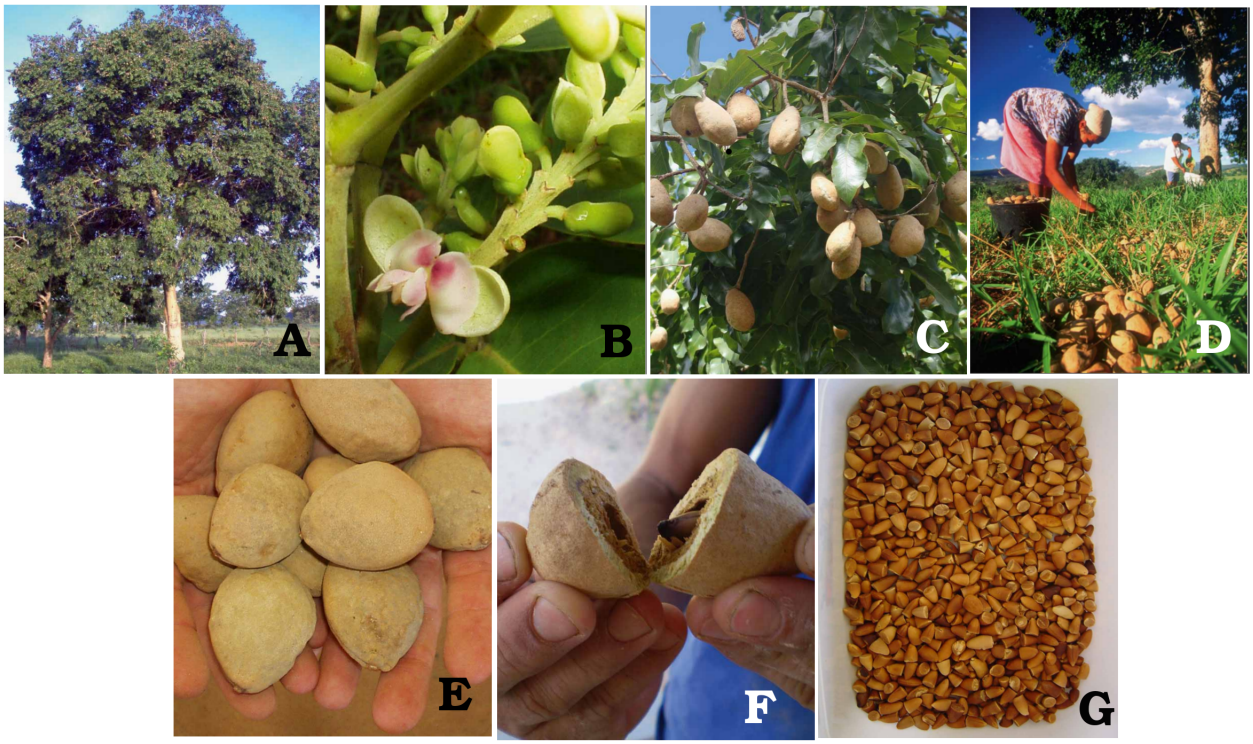


Figura 6.1: Baru (*Dipteryx alata*). A. barueiro; B. Flor do baru; C. Frutos do baru no pé; D. Colheita do baru (são colhidos apenas os frutos caídos); E. Frutos do baru; F. Fruto aberto do baru mostrando a castanha/amêndoa; G. castanha de baru processada (torrada) [33].

atmosfera. Tem sua primeira frutificação com cerca de 6 anos, sendo este período bastante variado em função das condições de solo e água. A época da floração e frutificação variam de acordo com a região, mas em geral ocorrem em fins de outubro a meados de dezembro, mas sua safra é intermitente, com grandes variações na intensidade de produção de frutos de um ano para o outro, entretanto, para efeitos práticos quanto à previsão de sua utilização comercial, pode-se considerar que o barueiro apresenta uma safra produtiva a cada 2 anos. Uma árvore adulta produz aproximadamente 150 kg de fruto por safra produtiva, o que corresponde a cerca de 1.000 a 3.000 frutos. O baru possui apenas uma semente por fruto, do qual se pode aproveitar a polpa, o endocarpo e a semente (castanha/amêndoa) [33, 54].

O baru está relacionado entre as 110 espécies nativas do Cerrado com maior potencial econômico para a população da região, e entre as 10 mais promissoras para cultivo, devido ao seu potencial [33, 69, 173, 193]: alimentício (polpa do fruto e das sementes, rica em minerais – cálcio, potássio, fósforo, ferro e zinco –, proteínas, fibras alimentares e lipídeos de alta qualidade [33, 54, 63, 69]), madeireiro (madeira clara e de alta densidade, altamente resistente ao ataque de fungos e cupins, apresentando, também, potencial para a exploração de celulose para fabricação de certos tipos de papel [54]), medicinal/-

farmacêutico/cosmético (óleo das sementes [54]), ornamental (planta para paisagismo), forrageiro (frutos caídos e sombra para o gado [96]), de elevado potencial tecnológico (devido ao alto redimento do óleo extraído de suas sementes [96]) e mesmo para utilização em artesanato [33], além disso, é uma planta melífera [69].

Como a exploração do baru ocorre por extrativismo, e como as informações sobre sua biologia e manejo ainda são insuficientes, é indispensável a realização de estudos que contribuam para direcionar estratégias mais eficientes para sua domesticação, conservação e uso sustentável [215].

Estudos anteriores mostraram significativa diferenciação genética espacial entre e dentro das populações locais de baru, e sua estrutura de larga escala geográfica está aparentemente associada a padrões históricos de fragmentação de seu *habitat* no Cerrado [50, 215].

6.2 Dados

Os dados de *D. alata* basearam-se no polimorfismo presente em 9 *loci* de microssatélites avaliados para um total de 642 árvores amostradas em 25 populações locais (Tabela 6.1) ao longo da distribuição geográfica da espécie (Figura 6.2), com amostras em cada localidade variando entre 12 a 32 indivíduos (Tabela 6.2). Cada uma das plantas coletadas foi georreferenciada por aparelho GPS. Neste conjunto de dados, identificou-se um total de 55 alelos distintos (Tabela 6.3)[1].

Os indivíduos foram caracterizados pelos tipos de alelos identificados nos *loci*-alvo, bem como pela heterozigose.

Tabela 6.1: Representação parcial e esquemática do dados de *D. alata* utilizados.

	Alelo																	
Indivíduo	*Bm164*		*DaE06*		*DaE12*		*DaE20*		*DaE34*		*DaE41*		*DaE63*		*DaE67*		*DaE46*	
1CMT	158	158	216	216	220	220	154	154	110	110	126	126	208	208	176	170	253	253
2CMT	170	158	216	216	220	220	154	154	116	114	126	126	210	210	176	170	253	253
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
29CAMT	176	174	220	220	220	218	154	154	114	110	132	132	208	208	176	176	250	250
30CAMT	156	156	220	220	220	218	154	154	114	110	124	124	208	208	176	176	250	250

Com base nos dados amostrados, foram produzidas as seguintes matrizes:

1. Matriz A: presença-ausência de alelos por indivíduo (indivíduo pode ser aqui entendido como sítio, posto que, conforme exposto, cada uma das plantas coletadas foi georreferenciada por aparelho GPS).

[1]Para detalhes acerca da metodologia de amostragem, v. Soares et al. [216]

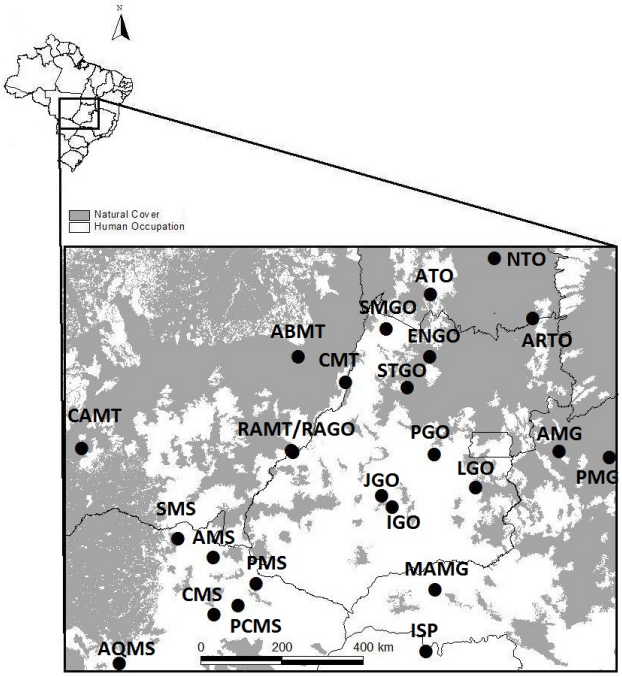


Figura 6.2: Localização geográfica, no Brasil Central, das 25 populações locais de *D. alata* amostradas. As regiões sombreadas são áreas ainda cobertas por remanescentes naturais da vegetação do Cerrado (adaptado de [92]).

Tabela 6.2: Populações e quantidade de indivíduos amostrados.

No. População	*Nome População*	*No. Indivíduos Amostrados*
1	CMT	32
2	ABMT	32
3	PGO	32
4	SMS	31
5	AMS	32
6	ATO	32
7	SMGO	32
8	LGO	32
9	ISP	31
10	MAMG	32
11	ENGO	12
12	STGO	12
13	AMG	32
14	PMG	32
15	PMS	13
16	PCMS	13
17	CMS	13
18	IGO	13
19	RAMT	27
20	RAGO	37
21	JGO	32
22	NTO	12
23	ARTO	15
24	AQMS	31
25	CAMT	30
Total	–	642

A matriz A é tal que $A_{k \times a}$, onde $k = 642$ (*indivíduos*) e $a = 55$ (*alelos*), com a_{ij} representando a ocorrência do alelo j no indivíduo i.

Tabela 6.3: Alelos identificados para os 9 loci sequenciados.

Bm164	*DaE06*	*DaE12*	*DaE20*	*DaE34*	*DaE41*	*DaE63*	*DaE67*	*DaE46*
156	212	216	146	104	118	208	170	244
158	216	218	154	106	120	210	176	247
165	220	219	156	108	122	214		250
168		220	158	110	124			253
170		222		112	126			
174				114	128			
176				116	130			
178				118	132			
				120	134			
				122	136			
				124	138			
					142			
					146			
					148			
					150			

2. Matriz B: heterozigose, por indivíduo, por alelos.

   A matriz B é tal que $B_{k \times a}$, onde $k = 642$ (*indivíduos*) e $a = 55$ (*alelos*), com b_{ij} representando a quantidade de ocorrências do alelo j no indivíduo i.

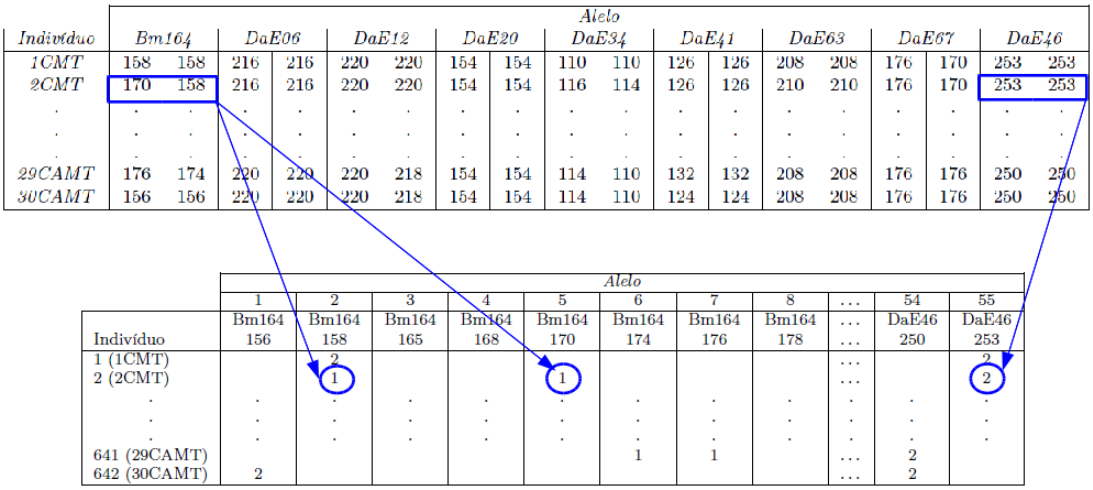
   Dessa forma, como para cada *locus*, um indivíduo possui dois alelos em seu genoma, um indivíduo i, homozigoto para um determinado *locus*, teria $b_{ij} = 2$, enquanto um indivíduo m, heterozigoto, teria $b_{mr} = 1$ e $b_{ms} = 1$, com $j, r, s \in \{1, ..., 55\}$ e $r \neq s$ (Figura 6.3.a).

   Como optou-se por trabalhar com minimizações, baseado no princípio da dualidade, a matriz B foi convertida em sua minimização equivalente, para tanto, foi necessário processar B para obter a matriz final B'. Tal processamento consistiu em, para indivíduos heterozigotos, substituir o valor 1 por -2; para homozigotos, substituir o valor 2 por -1. Procedendo desta forma, buscou-se beneficiar soluções com maior quantidade de alelos em heterozigose (que teriam valor total menor comparadas a soluções com homozigose) (Figura 6.3.b).

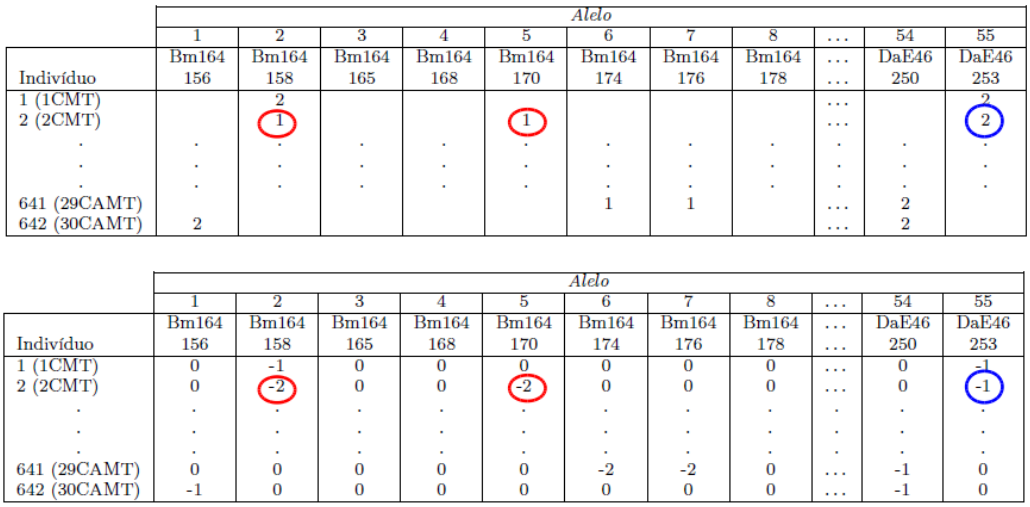
6.3 Problema: *Core Collections* de Bancos de Germoplasma

Um foco de pesquisa importante no contexto da preservação da biodiversidade diz respeito à manutenção de germoplasmas.

Quando se fala em plantas, germoplasma corresponde ao tecido vivo a partir do qual pode-se cultivar novas plantas. Há diversas maneiras de conservar-se o germoplasma, e.g., coleções de sementes, armazenamento de pólen, em enfermarias (plantações), *in*



(a) Obtenção da Matriz B.



(b) Obtenção da Matriz B'.

Figura 6.3: Obtenção da Matriz B' (de heterozigose, por indivíduo, por alelos). (a) Inicialmente a Matrix B é preenchida com a quantidade de vezes que um determinado alelo está presente no indivíduo: considerando-se um indivíduo i, $i \in \{1, ..., 642\}$ e um alelo j, $j \in \{1, ..., 55\}$, se o indivíduo é heterozigoto para aquele *locus*, i.e., possui dois alelos distintos, sendo um deles j, então $b_{ij} = 1$; se o alelo j está em homozigose (aparece duas vezes no genoma do indivíduo), $b_{ij} = 2$; e se o alelo não está presente no indivíduo, $b_{ij} = 0$. (b) Como optou-se por trabalhar com minimizações, a matriz B é processada, obtendo-se B'. Para tanto, nos indivíduos heterozigotos, substituiu-se o valor 1 por -2; nos homozigotos, substituiu-se o valor 2 por -1. Desta forma, buscou-se beneficiar soluções com maior quantidade de alelos em heterozigose

vitro [100], formando o que é conhecido como *bancos de germoplasma*. Bancos de germoplasma são uma estratégia importante para manutenção de recursos genéticos uma vez que preservam a diversidade genética de plantas que pode ser usada para estudos posteriores ou para projetos de restauração de *habitats* [68].

A manutenção de bancos de germoplasma é onerosa e seus custos estão mormente relacionados à manutenção do espaço de armazenadmento das amostras, ao controle da temperatura e da umidade, bem como à quantidade de equipamentos necessária [120]. Tais fatores dependem fortemente da quantidade de germoplasma a ser armazenada, o que nos leva ao conceito de *core collection*, um subconjunto de um banco maior de uma espécie, capaz de representar, com mínima repetição, a diversidade genética da espécie [29, 109]. O *core* não é um substituto do banco completo, mas busca capturar toda a diversidade do banco do qual é derivado, sendo assim, é uma ferramenta útil para organizar e analisar conjuntos representativos de genótipos em um banco de germoplasma.

O termo *accession* é usado para referir uma amostra presente no banco completo, ao passo que *entrada* corresponde a qualquer *accession* selecionado para inclusão no *core* [29].

Metodologicamente falando, esforços para criar *core collections* a partir de bancos de germoplasma geralmente utilizam análises estatísticas complexas e métodos de *clusterização*, entretanto, tais processamentos são feitos de uma maneira não automatizada, utlizando a aplicação sucessiva de passos repetitivos [20, 234, 241].

No desenvolvimento de uma *core collection*, busca-se a minimização do custo geral de conservação enquanto se maximiza a diversidade genética. O problema pode ser enunciado como a maximização do número de alelos ao mesmo tempo que se minimiza o número de entradas necessárias para representar esses alelos, maximizando, simultaneamente, a heterozigose.

Uma solução candidata para o problema é o vetor $\overrightarrow{x} = \{x_1, ..., x_k\}$, onde k é o número de *accessions* (neste caso, árvores amostradas, $k = 642$).

Seja $x_i \in \{0, 1\}$, tal que, $x_i = \begin{cases} 1, & \text{se o } \textit{accession } i \text{ foi selecionado para compor a solução;} \\ 0, & \text{caso contrário.} \end{cases}$

Cada *accession* i tem um custo c_i, e cada alelo j, um nível de representação desejado r_j. O objetivo é obter-se:

$$min\left(\sum_{i=1}^{k} c_i x_i\right) \tag{6.1}$$

Sujeito a:

$$\forall j \in \{1, 2, ..., m\},\ \sum_{i=1}^{k} a_{ij} x_i \geq r_j \tag{6.2}$$

Onde m = número total de alelos (i.e., $m = 55$) e $r_j = 1$ (i.e., cada alelo deve estar representado pelo menos uma vez).

Como função *fitness*, obtém-se tantas representações da Eq. 6.1 quanto objetivos a serem otimizados, variando-se c_i de acordo com o objetivo sob consideração. Isto nos permite otimizar simultaneamente objetivos distintos, em vez de agregá-los em uma única função. Desta forma, com relação à abordagem MOO, neste problema, há três objetivos a serem simultaneamente otimizados:

1. Minimizar o número de *accessions* selecionados para compor a solução (i.e., o número de entradas para o *core*);

$$min(f_1(\overrightarrow{x})) = min(selected\_accessions(\overrightarrow{x})) \tag{6.3}$$

2. Maximizar a representação de alelos;

$$\begin{aligned} max(f_2'(\overrightarrow{x})) = max(alleles(\overrightarrow{x})) \Leftrightarrow \\ min(f_2(\overrightarrow{x})) = min(lacking\_alleles(\overrightarrow{x})) \\ = min(m - alleles(\overrightarrow{x})) \end{aligned} \tag{6.4}$$

3. Maximizar a heterozigose[2].

$$max(f_3'(\overrightarrow{x})) = max(heterozygosity(\overrightarrow{x})) \tag{6.5}$$

Optou-se por trabalhar com minimizações, assim, baseado no princípio da dualidade, sem perda de generalidade, todos os objetivos foram convertidos em seu equivalente de minimização (e.g., para Equação 6.4, a otimização consistiu em minimizar o número de alelos faltantes, o que é o mesmo que maximizar o número de alelos).

De acordo com a literatura, o *core* escolhido para representar tanto quanto possível a diversidade do banco é composto de aproximadamente 10% do banco completo [29]. Assim, definiu-se $\approx 10\%$ de alelos faltantes, e $\approx 15\%$ do número de *accessions* como restrições para o espaço de objetivos. Corroborou esta decisão a informação dos especialistas (biólogos) que soluções com mais de 5 alelos faltantes e 107 entradas não seriam aceitáveis na prática. Como resultado, foram definidas restrições cujo cômputo foi baseado nas funções objetivo conforme segue:

$$c_1(\overrightarrow{x}) = \begin{cases} f_1(\overrightarrow{x}) - 107, & \text{se } f_1(\overrightarrow{x}) > 107; \\ 0, & \text{caso contrário.} \end{cases} \tag{6.6}$$

[2]A conversão da maximização da heterozigose em uma minimização se deu pelo processamento de B e consequente obtenção de B', conforme descrito na Seção 6.2.

$$c_2(\overrightarrow{x}) = \begin{cases} f_2(\overrightarrow{x}) - 5, & \text{se } f_2(\overrightarrow{x}) > 5; \\ 0, & \text{caso contrário.} \end{cases} \tag{6.7}$$

Neste contexto, a solução que viola qualquer uma das restrições é dita não-factível, caso contrário é factível.

Foi usado o conceito de dominância na presença de restrições, conforme a Definição 8 apresentada à Seção 4.1.3.

6.4 Modelo Nulo

Um modelo nulo é uma tentativa de gerar distribuições de valores para uma determinada variável de interesse na ausência do processo causal em estudo, possibilitando, assim como nas ciências experimentais, estipular uma "situação controle" [115, 174]. No nosso caso, o processo causal é a aplicação do algoritmo MOO.

O modelo nulo consiste no processo estatístico pelo qual gera-se uma distribuição de probabilidade por simulação, no qual os valores não se apresentam sob o efeito do processo em questão ("distribuição nula"). O objetivo é mostrar, através do emprego de um modelo nulo, que os resultados encontrados não teriam emergido de dados gerados ao acaso.

6.5 *Spartan*

Para tratar aspectos de *incerteza*, inerentes a sistemas estocásticos, foi utilizada a ferramenta *Spartan* (*Simulation Parameter Analysis R Toolkit Application*) [7], um pacote que disponibiliza técnicas de análise estatística que, juntas, fornecem uma ferramenta para explorar o efeito de incertezas no resultado de simulações/experimentos, auxiliando a identificação de quais resultados experimentais podem ser atribuídos ao problema modelado, mais do que a artefatos de incerteza, parametrização, ou estocasticidade [159].

Spartan foi originalmente desenvolvido para análise em sistemas de simulações biológicas, onde procurava-se identificar quão representativa uma simulação seria em relação ao sistema biológico que buscava representar e como resultados *in silico* poderiam ser interpretados no domínio biológico [6, 188]. Mas seu emprego pode ser estendido, sem perda de generalidade, para a avalição de incertezas em outros contextos, fazendo com que se beneficiem da análise estatística de resultados empíricos [8, 187].

As técnicas de *Spartan* utilizadas são explanadas a seguir. Ressalte-se que, como cada técnica utiliza diferentes métodos de amostragem de parâmetros, não é possível utilizar os resultados gerados em uma técnica para outra.

6.5.1 Técnica 1: Análise da Incerteza Aleatória

Esta técnica é utilizada a fim de definir-se o número de execuções necessárias para mitigar a incerteza associada à aleatoriedade do algoritmo estocástico empregado e atingir-se o nível desejado de acurácia. Encontrar a relação adequada entre tamanho de amostra e efeito da incerteza associada à aleatoriedade é importante para balancear requisitos (e.g., fidelidade desejada) e recursos (e.g., recursos computacionais disponíveis e/ou necessários) [187].

A técnica é aplicada em uma implementação *baseline* do algoritmo, i.e., uma implementação que serve de ponto de partida para efetuar comparações. Um conjunto de valores de parâmetros é definido e fixado para todas as execuções.

Seguindo o protocolo de *Spartan* [7], são definidos tamanhos de amostras 1, 5, 10, 50, 100 e 300. Um tamanho de amostra 5 indica que o experimento será repetido 5 vezes. Para cada tamanho de amostra são gerados 20 conjuntos de experimentos (Figura 6.4).

A consistência da análise é feita contrastando-se a distribuição das soluções encontradas, usando o mesmo conjunto fixo de valores de parâmetros e contendo números idênticos de amostras de experimentos, e.g., para o tamanho de amostra 5, são geradas 20 pastas, cada uma contendo 5 execuções do algoritmo, num total de 100 execuções individuais.

Inicialmente, os tamanhos de amostra são analisados isoladamente. Uma distribuição das medianas das soluções encontradas pelo algoritmo em cada execução é gerada para cada um dos 20 conjuntos da amostra. As distribuições 2 a 20 são contrastadas com a distribuição 1 usando o A-teste de Vargha-Delaney [229], que estabelece uma "significância científica" contrastando dois conjuntos para aquele tamanho de amostra e retornando a probabilidade de que a amostra selecionada aleatoriamente de uma população seja maior do que uma amostra selecionada aleatoriamente de outra população. A significância científica é determinada pela comparação do resultado com medidas estabelecidas por Vargha-Delaney: resultados acima de 0,71 ou abaixo de 0,29 indicam uma diferença cientificamente significativa entre as populações e 0,5 indica que não há diferença [187]. Estas diferenças estatísticas podem ser vistas em gráficos produzidos por *Spartan* para cada tamanho de amostra (Figura 6.5a-g). Um bom tamanho de amostra corresponde àquele em que a diferença estatística é pequena.

Pela comparação de resultados em tamanhos diferentes de amostras, é possível determinar o número de execuções necessárias para obter-se distribuições estatisticamente consistentes (Figura 6.5h). Amostras maiores produzem melhores distribuições, atenuando, assim, o efeito da estocasticidade nos resultados.

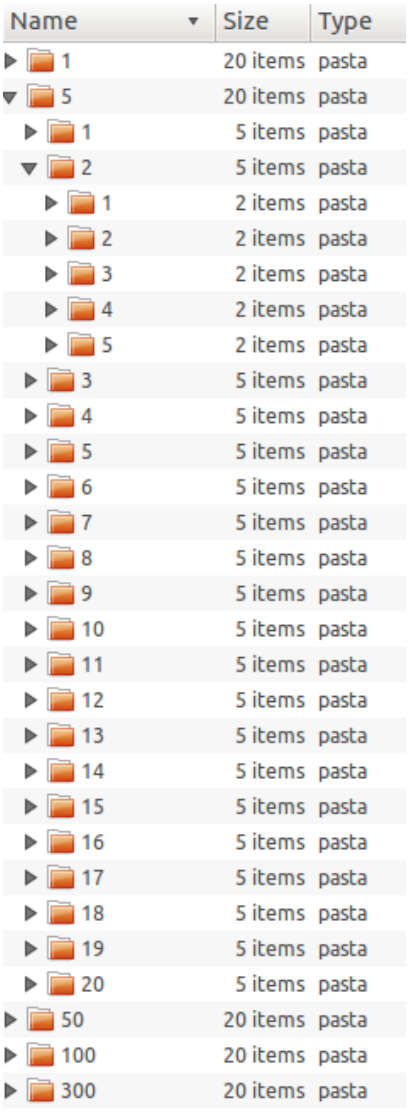


Figura 6.4: Estrutura de arquivos para a Técnica 1 de *Spartan*. O diretório contém as pastas numeradas de acordo com os tamanhos de amostras em análise, no caso, 1, 5, 50, 100 e 300. Cada uma destas pastas, contém 20 subpastas, numeradas de 1 a 20, uma para cada conjuto de resultados. Cada uma das 20 pastas contém o número de execuções do algoritmo para o tamanho de amostra em análise, e.g., se a incerteza de 5 execuções está sendo analisada, cada um das 20 pastas conterá 5 subpastas, numeradas de 1 a 5, cada uma com uma execução completa do algoritmo.

6.5.2 Técnica 2: Análise da Robustez

Quase todos os procedimentos heurísticos envolvem algum ajuste de parâmetros. A tarefa de calibragem dos parâmetros é notadamente desafiadora, pois não se sabe, antecipadamente, qual o impacto dos valores dos parâmetros no desempenho do algoritmo, especialmente quando o algoritmo a ser ajustado é estocástico por natureza [214].

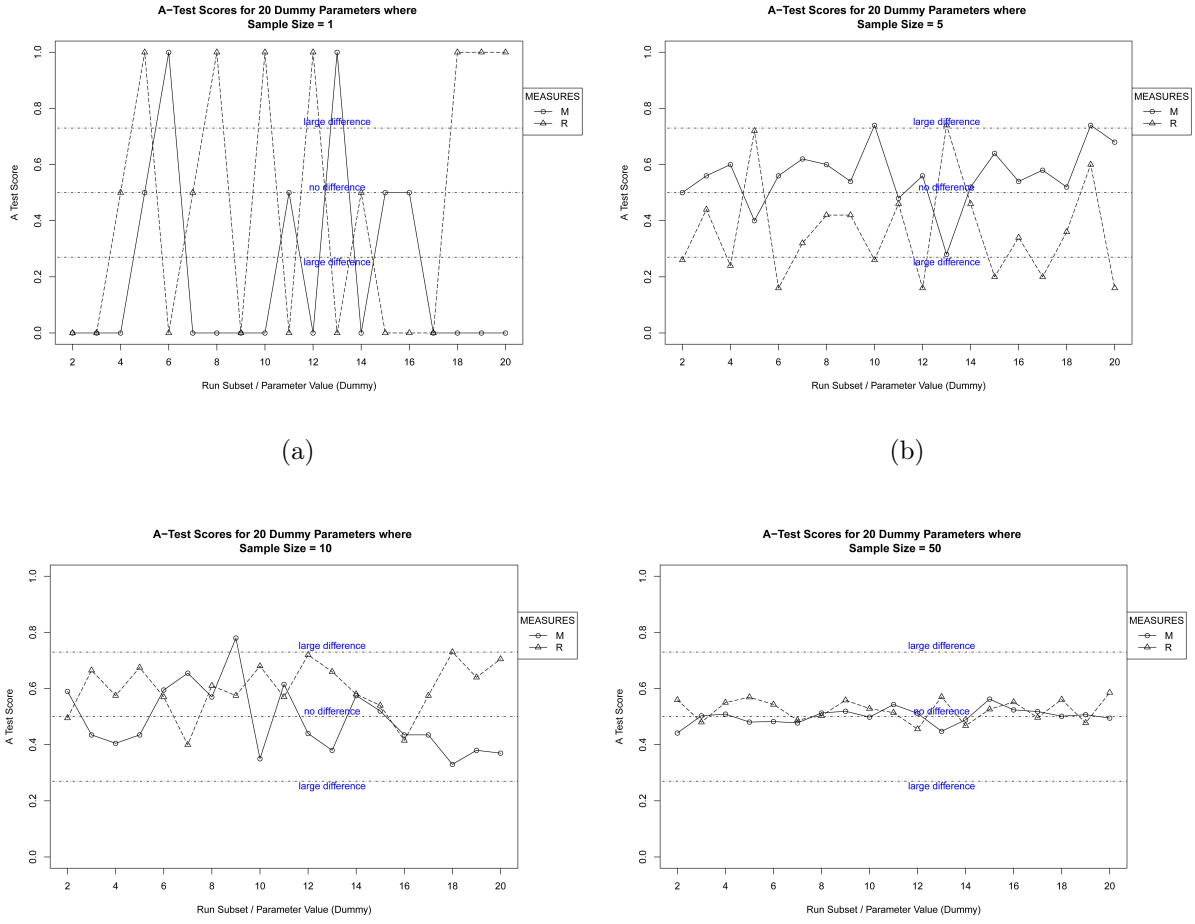
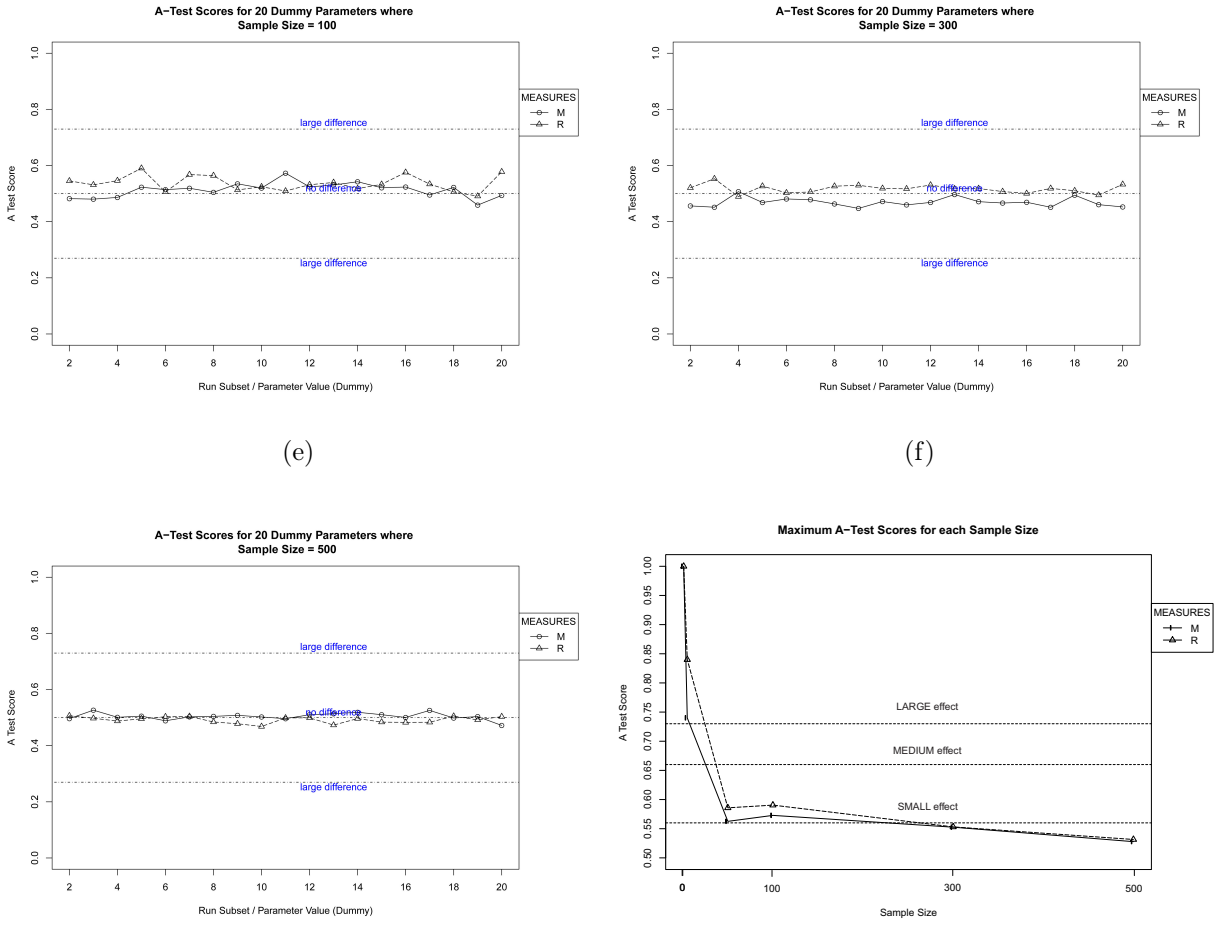


Figura 6.5: Técnica 1 de *Spartan*: análise do número de execuções necessárias para gerar um resultado repesentativo para um experimento estocástico. (a-g) gráfico do resultado do A-Test para tamanhos de amostra 1, 5, 10, 50, 100, 300 e 500, respectivamente. Valores acima de 0,71 ou abaixo de 0,29 indicam uma diferença cientificamente significativa entre resultados; 0,5 indica que não há diferença. (h) A-Test final reunindo o resultado de todos os tamanhos de amostra (considerando-se os 20 conjuntos de resultados para cada tamanho de amostra). Como a direção do efeito não é importante (e sim seu valor), pontuações abaixo de 0,5 são assinaladas nos valores correspondentes acima de 0,5 (como se fosse uma função módulo). A magnitude dos efeitos é indicada: grande (*large*), média (*medium*) ou pequena (*small*). A figura indica que um tamanho de amostra de no mínimo 300 execuções é necessária (e neste caso suficiente) para reduzir a magnitude da incerteza associada à aleatoriedade para que ela atinja um efeito menor do que pequeno, para todas as execuções do algoritmo em estudo.



(e) (f)

(g) (h)

Figura 6.5: Técnica 1 de *Spartan*. (cont.)

A Técnica 2 de *Spartan* determina a robustez do algoritmo a perturbações nos parâmetros, examinando as implicações da incerteza na estimação de seus valores e suas consequências nos resultados dos experimentos.

Quando considerada no contexto da incerteza epistêmica (refletida em parâmetros para os quais os valores mais adequados ainda não foram determinados), a robustez diz respeito ao fato de se produzir resultados mais previsíveis, com um grau relativamente alto de confiança.

A abordagem utilizada é chamada "um de cada vez" (*one at a time*–OAT), i.e., cada parâmetro é ajustado independentemente dos outros, que mantêm seus valores de *baseline* fixos. O método de amostragem começa no valor mais baixo do parâmetro que é aumentado em incrementos regulares até que o limite superior, i.e., o valor mais alto para o parâmetro, seja atingido. Para cada "perturbação" no valor do parâmetro em análise, são realizadas n execuções do algoritmo, sendo que n foi determinado na etapa anterior, empregando a Técnica 1 de *Spartan*. O A-teste de Vargha-Delaney [229] é então aplicado para determinar se a mudança no valor do parâmetro leva a uma mudança cientificamente significativa quando comparada à execução *baseline*. Isso indica quão robusto o algoritmo é à alteração de cada parâmetro, bem como os pontos nos quais a perturbação do parâmetro resulta em mudanças significativas no comportamento do algoritmo.

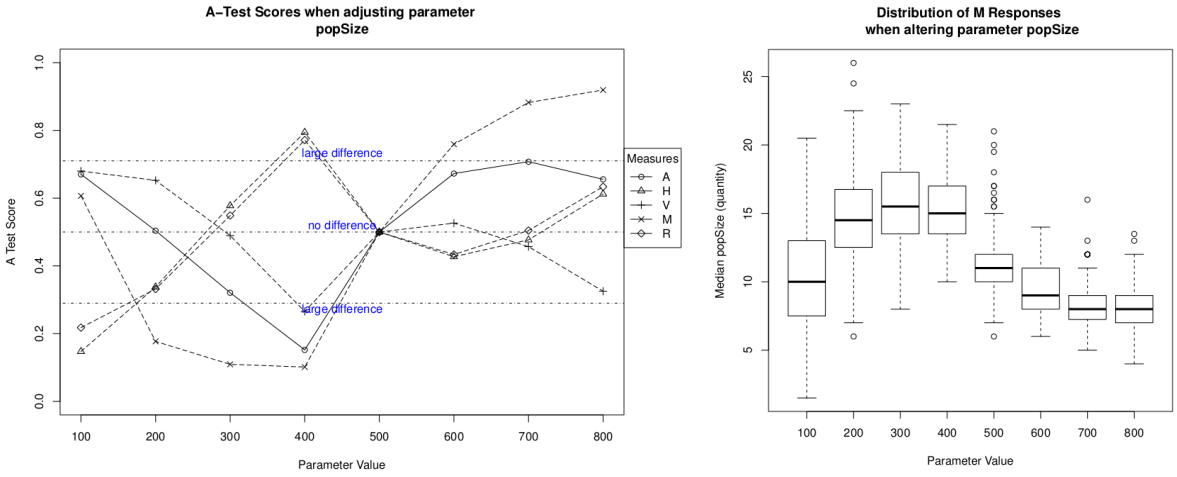
Ao final, os resultados dos A-testes para cada parâmetro são apresentados em um gráfico, permitindo uma fácil identificação dos valores dos parâmetros que causam uma mudança cientificamente significativa nos resultados do algoritmo (Figura 6.6) e aqueles valores mais adequados a serem utilizados (de acordo com o objetivo otimizado).

Quando a execução do algoritmo mostra-se altamente sensível à variação do valor do parâmetro, deve-se tomar cuidado na interpretação dos resultados, pois eles podem ser artefato da parametrização.

6.5.3 Técnica 3: Análise da Sensibilidade Global

Apesar da análise da robustez esclarecer efeitos da perturbação de um parâmetro, não é capaz de revelar o efeito quando dois ou mais parâmetros são ajustados simultaneamente. Não raras vezes, o efeito de um parâmetro é influenciado por outro. A análise da sensibilidade global revela tais efeitos, mostrando como diferentes parâmetros estão relacionados e indicando os que têm maior influência nos resultados do algoritmo [159] .

Spartan fornece uma técnica que perturba todos os parâmetros de interesse simultaneamente, de maneira que os resultados da análise são altamente representativos da dinâmica do algoritmo [7]. A abordagem utilizada emprega amostragem por hipercubo



(a) (b)

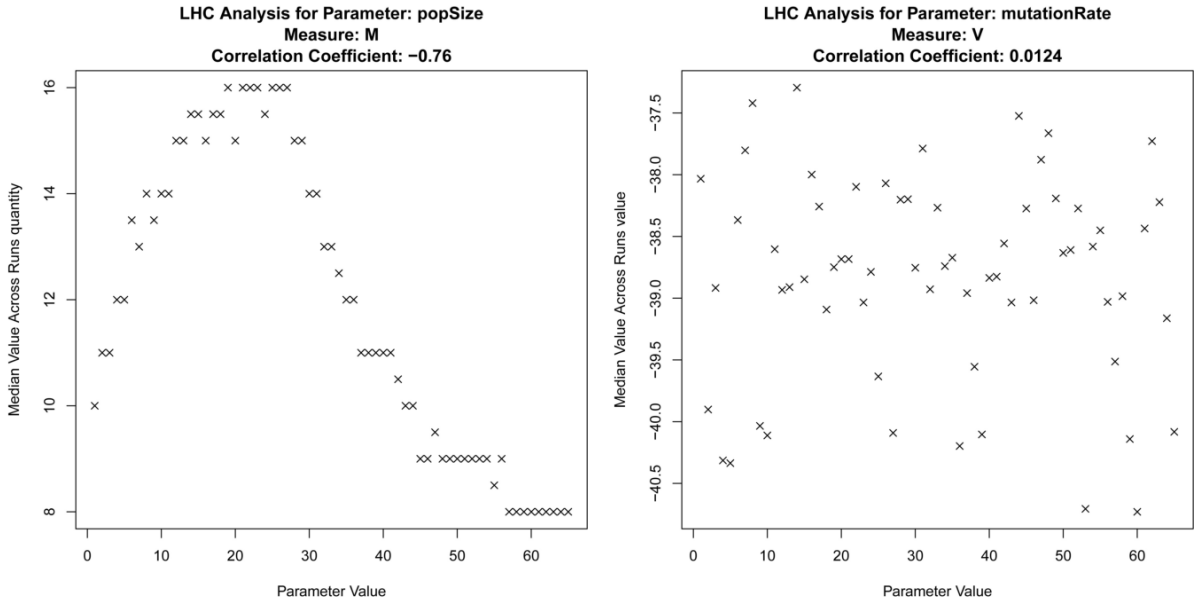
Figura 6.6: Técnica 2 de *Spartan*: análise da robustez. (a) O efeito da magnitude do parâmetro *popSize* sobre diferentes objetivos (*measures*) é mostrado pelas diversas curvas obtidas ao variar os valores assumidos pelo parâmetro (entre 100 a 800). O eixo y mostra os valores para o A-teste de cada resultado sob o parâmetro em questão quando comparado ao valor padrão (*default*), onde todas as linhas convergem. (b) *Boxplot* mostrando a distribuição dos resultados do algoritmo para um objetivo M quando se varia o valor do parâmetro *popSize*. Caso o problema procure minimizar o objetivo M, o valor mais adequado para *popSize* seria 700, fazendo-se a ressalva que a interpretação dos resultados deve ser procedida com cuidado, dada a grande variação no A-Test de M quando variou-se *popSize*.

latino (*latin hypercube sampling–LHS*)[3] para selecionar os conjuntos de valores de parâmetros dentro dos limites mínimo e máximo informados, ao mesmo tempo que minimiza a correlação entre os valores dos parâmetros, garantindo uma cobertura eficiente do espaço de parâmetros, existe apenas uma amostra em cada segmento do domínio de cada parâmetro.

Para cada combinação de parâmetros gerada pelo LHS, são efetuadas n execuções do algoritmo, sendo n determinado empregando-se a Técnica 1 de *Spartan*. A partir do resultado obtido, é gerado um gráfico para cada par parâmetro-objetivo, revelando a correlação entre eles, que é quantificada por meio do PRCC (*Partial Rank Correlation Coefficient*) [187] (Figura 6.7). Segundo Alden et al. [7], o PRCC considera a relação não-linear entre parâmetro-objetivo e corrige o efeito de outros parâmetros na resposta, fornecendo um indicador robusto do efeito do parâmetro na resposta do algoritmo apesar

[3]Para duas variáveis, a LHS consiste em dividir o espaço bidimensional em $n \times n$ casas e escolher n casas de maneira que não haja dois pontos em uma mesma linha ou coluna. O hipercubo latino é a generalização deste método para k dimensões.

dos outros parâmetros também serem perturbados.



(a) (b)

Figura 6.7: Técnica 3 de *Spartan*: análise da sensibilidade global. Cada ponto do gráfico corresponde a uma combinação de parâmetros gerada pelo LHS para a qual foram efetuadas n execuções do algoritmo (n determinada pela Técnica 1 de *Spartan*). No caso em tela, foram criadas 65 combinações de parâmetros. Para cada par parâmetro-objetivo é gerado um gráfico revelando a correlação entre eles, quantificada por meio do PRCC (*Partial Rank Correlation Coefficient*). (a) Gráfico mostrando forte correlação negativa (-0,76) entre o parâmetro *popSize* e o objetivo M. (b) Gráfico mostrando fraca correlação (0,0124) entre o parâmetro *mutationRate* e o objetivo V.

6.6 Métricas

Na otimização monobjetivo, a qualidade de uma solução pode ser medida em termos da função objetivo, em uma minimização, quanto menor o valor, melhor a solução. Entretanto, quando se fala em otimização multiobjetivo, o conceito de dominância de Pareto pode ser utilizada, mas a possibilidade de duas soluções serem incomparáveis, quando nenhuma delas domina a outra, complica a situação. A situação torna-se ainda mais complexa ao se comparar dois conjuntos de soluções, pois algumas soluções podem ser dominadas por soluções do outro conjunto, enquanto outras podem ser incomparáveis. Desta forma, não é imediata a relação entre qualidade e aproximações às frentes de Pareto. Historicamente representações gráficas têm sido utilizadas para comparar resul-

tados obtidos por algoritmos MOO, sem embargo, definir uma métrica adequada não é uma tarefa fácil, sendo objeto de vários estudos [48].

Esta dificuldade é agravada ao se considerar que a maioria dos algoritmos para MOO são estocásticos (e este é o caso dos algoritmos utilizados neste trabalho). Como uma alternativa, o desempenho de algoritmos estocásticos de otimização pode ser avaliado experimentalmente através de múltiplas execuções independentes do algoritmo e análise estatística dos resultados [64]. Lidar com conjuntos de resultados aleatórios introduz dificuldades adicionais na análise.

Medidas de qualidade são necessárias para comparar o resultado obtido por diferentes algoritmos. Certamente, o método mais simples de comparação seria mensurar se a saída de um algoritmo domina completamente a saída de outro. Entretanto, a razão para se usar medidas de qualidade reside no fato delas permitirem conclusões mais consistentes, que são baseadas em certas assunções acerca das preferências dos decisores [242]:

1. Se um algoritmo é melhor do que outro, pode-se quantificar quão melhor ele é?
2. Se não é possível afirmar que um algoritmo é melhor do que o outro, há certos aspectos a respeito dos quais se pode dizer que o primeiro é melhor do que o segundo?

Assim, a questão chave, quando se usam métricas, é: como melhor resumir a aproximação à frente de Pareto através de algumas características numéricas, à semelhança de análises estatísticas, onde a média, o desvio padrão, etc., são usados para descrever, de uma maneira concisa, a probabilidade de distribuição de um resultado? A perda de algum grau de informação é inevitável, mas o ponto crucial é não perder a informação na qual se está interessado [242].

As métricas mais comumente utilizadas são unárias, atribuindo a cada aproximação ao conjunto de Pareto um número que reflita um aspecto qualitativo. Em geral é utilizada uma combinação de métricas a fim de comparar duas aproximações à frente de Pareto. A grande questão é qual afirmação pode ser feita com base na informação fornecida por tais métricas. É possível concluir que um conjunto de aproximações à Frente de Pareto X é melhor do que Z?

Segundo Zitzler, Dev e Thiele [244], uma otimização deveria cumprir as seguintes metas:

Meta 1: minimizar a distância do conjunto de soluções não-dominadas encontrada pelo algoritmo à frente ótima de Pareto;

Meta 2: obter uma boa distribuição de soluções (tanto quanto possível uniforme);

Meta 3: maximizar a extensão da frente não-dominada obtida (i.e., para cada objetivo, uma ampla variedade de valores deve ser coberta pelas soluções não-dominadas).

Em outras palavras, uma otimização deveria buscar a convergência ao ótimo de Pareto e a manutenção da diversidade das soluções [89].

No presente estudo, estas três metas de otimização são avaliadas pelas seguintes métricas:

1. Função C;
2. Empirical Attainment Function (EAF);
3. Hipervolume (H);
4. Espaçamento (S);
5. Extensão (E).

As métricas função C, EAF e H são usadas para avaliar a Meta 1; S a Meta 2 e E a Meta 3 e serão descritas a seguir.

6.6.1 Função C [244]

Sejam $\overrightarrow{x}'$ e $\overrightarrow{x}''$ dois conjuntos de vetores de decisão. A função C mapeia o par ordenado ($\overrightarrow{x}'$, $\overrightarrow{x}''$) no intervalo [0,1]:

$$C(\overrightarrow{x}', \overrightarrow{x}'') = \frac{|\{x' \in \overrightarrow{x}'; \exists x'' \in \overrightarrow{x}'' : x' \preceq_{cd} x''\}|}{|\overrightarrow{x}''|} \tag{6.8}$$

Usando a função C, é possível determinar se a saída de um algoritmo domina a saída do outro, i.e., um par de conjuntos não-dominados é comparado calculando-se a fração de cada conjunto que é *coberto* pelo outro. $C(\overrightarrow{x}', \overrightarrow{x}'') = 1$ significa que todas as soluções em $\overrightarrow{x}''$ são dominadas pelas soluções em $\overrightarrow{x}'$, enquanto que $C(\overrightarrow{x}', \overrightarrow{x}'') = 0$, indica que nenhuma das soluções em $\overrightarrow{x}''$ é coberta pelo conjunto $\overrightarrow{x}'$. Ambos $C(\overrightarrow{x}', \overrightarrow{x}'')$ e $C(\overrightarrow{x}'', \overrightarrow{x}')$ devem ser considerados, uma vez que $C(\overrightarrow{x}', \overrightarrow{x}'')$ não é necessariamente igual a $1 - C(\overrightarrow{x}'', \overrightarrow{x}')$.

6.6.2 Função de Aproveitamento Empírica (*Empirical Attainment Function–EAF*) [48, 64, 105, 119]

EAF é um indicador de qualidade usado para avaliar algoritmos estocásticos. É computado a partir da combinação de conjuntos de aproximação. Sejam $b_1(z), ..., b_n(z)$ n

execuções do algoritmo de otimização, então a EAF é definida como $EAF : \mathbb{R}^d \longmapsto [0, 1]$ com:

$$EAF = \frac{1}{n} \sum_{i=1}^{n} b_i(z) \tag{6.9}$$

Ela oferece uma descrição útil da distribuição das soluções. Diferenças na frequência com que certos objetivos são alcançados pelo algoritmo são representadas graficamente. A intensidade de coloração corresponde à frequência da solução.

6.6.3 Hipervolume (H) [144, 248]

Para o conjunto A de vetores não dominados $\overrightarrow{w_i}$, e um vetor referência $\overrightarrow{w}_{ref}$, que é dominado por todos os membros de A e cujos componentes correspondem ao valor máximo de cada objetivo[4], o *hipervolume* é a soma de todas as áreas retangulares, limitadas por A e $\overrightarrow{w}^{ref}$ [5] de acordo com:

$$H(A, \overrightarrow{w}_{ref}) \triangleq \bigcup_{i \in 1..|A|} H(\overrightarrow{w_i}, \overrightarrow{w}_{ref}) \tag{6.10}$$

Valores maiores de H correspondem a melhores soluções, pois indicam frentes mais próximas à verdadeira frente de Pareto (PF_{true}) e, portanto, mais distantes do ponto de referência $\overrightarrow{w}_{ref}$.

O uso do hipervolume traz duas principais vantagens [243]:

1. é sensível a qualquer tipo de melhora, i.e., sempre que um conjunto de aproximação à frente de Pareto X domina outro Z, a medida apresenta um valor necessariamente melhor para X do que para Z;

2. como resultado da primeira propriedade, o hipervolume garante que qualquer conjunto de aproximação à frente de Pareto X que alcance o maior valor possível para um problema particular contém todos os pontos ótimos de Pareto [104].

O hipervolume de um conjunto de vetores não-dominados A de uma otimização com k objetivos pode ser calculado, na prática, projetando-se o conjunto de vetores progressivamente em dimensões menores e calculando a sua integral (o somatório das áreas sucessivas delimitadas pelos pontos). Para calcular H em apenas duas dimensões (otimização de dois

[4]Considerando-se que a otimização é uma minimização.

[5]No problema SCP em questão, apresentado na Seção 6.3, tem-se que $\overrightarrow{x}^{ref} = \{max(f_1(\overrightarrow{x})), max(f_2(\overrightarrow{x})), max(f_3(\overrightarrow{x}))\} = \{642, 55, 0\}$.

objetivos, $k = 2$) os pontos são colocados em ordem decrescente dos valores do objetivo 2, e em seguida computa-se a seguinte expressão:

$$\sum_{i \in 1...|A|} |w_i^{k-1} - w_{ref}^{k-1}| * |w_i^k - w_{i-1}^k| \tag{6.11}$$

Onde inicialmente $\overrightarrow{w}_0 = \overrightarrow{w}_{ref}$. Para dimensões maiores, o cálculo é generalizado pela função recursiva $calculaHipervolume(A, \overrightarrow{w}_{ref}, k)$ (Algoritmo 17), proposta por Knowles e Corne [144].

Algoritmo 17: calculaHipervolume $(A, \overrightarrow{w}_{ref}, k)$

Dados:
Seja A um conjunto de vetores não-dominados.
Seja k o número de dimensões no espaço de busca dos vetores contidos em A, i.e., o número de funções objetivo otimizadas e, consequentemente, o número de dimensões para o cálculo do hipervolume.
Seja $\overrightarrow{w}_{ref}$ o vetor de referência dominado por todos os vetores de A, cujos componentes correspondem aos máximos valores possíveis em cada objetivo (considerando-se que a otimização é uma minimização).
Seja $\overrightarrow{w}_{high}$ o vetor com o maior valor no objetivo k dentre os vetores constantes do conjunto A atualizado.

```
1  begin
2      H ← 0
3      w^k_prev ← w^k_ref
4      while A is not ∅ do
           /* Ordena A em ordem decrescente com relação ao objetivo k   */
5          A ← sort(A, k, descend)
           /* Se A tem apenas 2 dimensões                               */
6          if k < 3 then
7              H_{k-1} ← w^1_high
8          else
9              H_{k-1} ← calculaHipervolume(A, w_ref, k − 1)
           /* Multiplica a aresta/área/volume encontrada(o) pela próxima
              dimensão                                                  */
10         H ← H + H_{k-1} * |w^1_high − w^1_prev|
11         w^1_prev ← w^1_high
           /* Remove os pontos ja utilizados de A                       */
12         A ← A \ {w_i | w^1_i ≥ w^1_high, w_i ∈ A}
13     return S
```

A Figura 6.8 mostra graficamente, passo-a-passo, o cálculo de H para o conjunto $A = \{\overrightarrow{w}_1, \overrightarrow{w}_2, \overrightarrow{w}_3\}$ representado na Figura 6.8a, utilizando-se a chamada à função $calculaHipervolume(A, \overrightarrow{w}_{ref}, 3)$. Na primeira instância de $calculaHipervolume$, a li-

nha 5 do Algoritmo 17 ordena os pontos/vetores de A em ordem decrescente dos valores da 3^a dimensão (eixo z) (linha 5). Em seguida, é feita a chamada recursiva a $calculaHipervolume(A, \overrightarrow{w}_{ref}, 2)$ (linha 9), quando os vetores de A são colocados em ordem decrescente de valores da 2^a dimensão (eixo y) (linha 5), a partir de então a área é calculada incrementalmente, por meio da atualização dos valores de $\overrightarrow{w}_{prev}$ (Figuras 6.8b-e) (linhas 5-12), a primeira área "parcial total" é multiplicada pela altura h_1. Em seguida, é calculada a área seguinte (Figuras 6.8f-h); a segunda área "parcial total" é multiplicada por h_2. Por fim, é calculada a terceira e última área "parcial total" (Figuras 6.8i-j) que é, por sua vez, multiplicada por h_3. A soma dos três volumes calculados sucessivamente é o valor final para o hipervolume do conjunto A.

Mais informações sobre hipervolume podem ser encontradas nas referências [23, 25, 26, 99, 176, 247, 243].

6.6.4 Espaçamento (S) [211]

Esta métrica é usada para avaliar numericamente a distribuição das soluções ao longo da frente de Pareto conhecida (PF_{known}, i.e., o conjunto de soluções não-dominadas obtido até então). S mede a variância da distância de cada elemento de PF_{known} em relação a seu vizinho mais próximo, assim, busca medir a dispersão (o "espalhamento") das soluções no espaço de objetivos:

$$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(\overline{d} - d_i)^2} \tag{6.12}$$

Onde $\|\cdot\|$ é a norma/distância entre dois pontos, $d_i = min_j(\|f_1^i - f_1^j\| + ... + \|f_k^i - f_k^j\|)$, i.e., a norma/distância ao vizinho mais próximo de i; $i, j = \{1, ..., n\}$; $\overline{d}$ é a média de todas as distâncias d_i; $f_1, ..., f_k$ são as funções objetivo; n é o número de vetores em PF_{known}.

Se $S = 0$, o algoritmo encontrou a distribuição ideal de vetores não-dominados (todas as soluções estão uniformemente espaçadas).

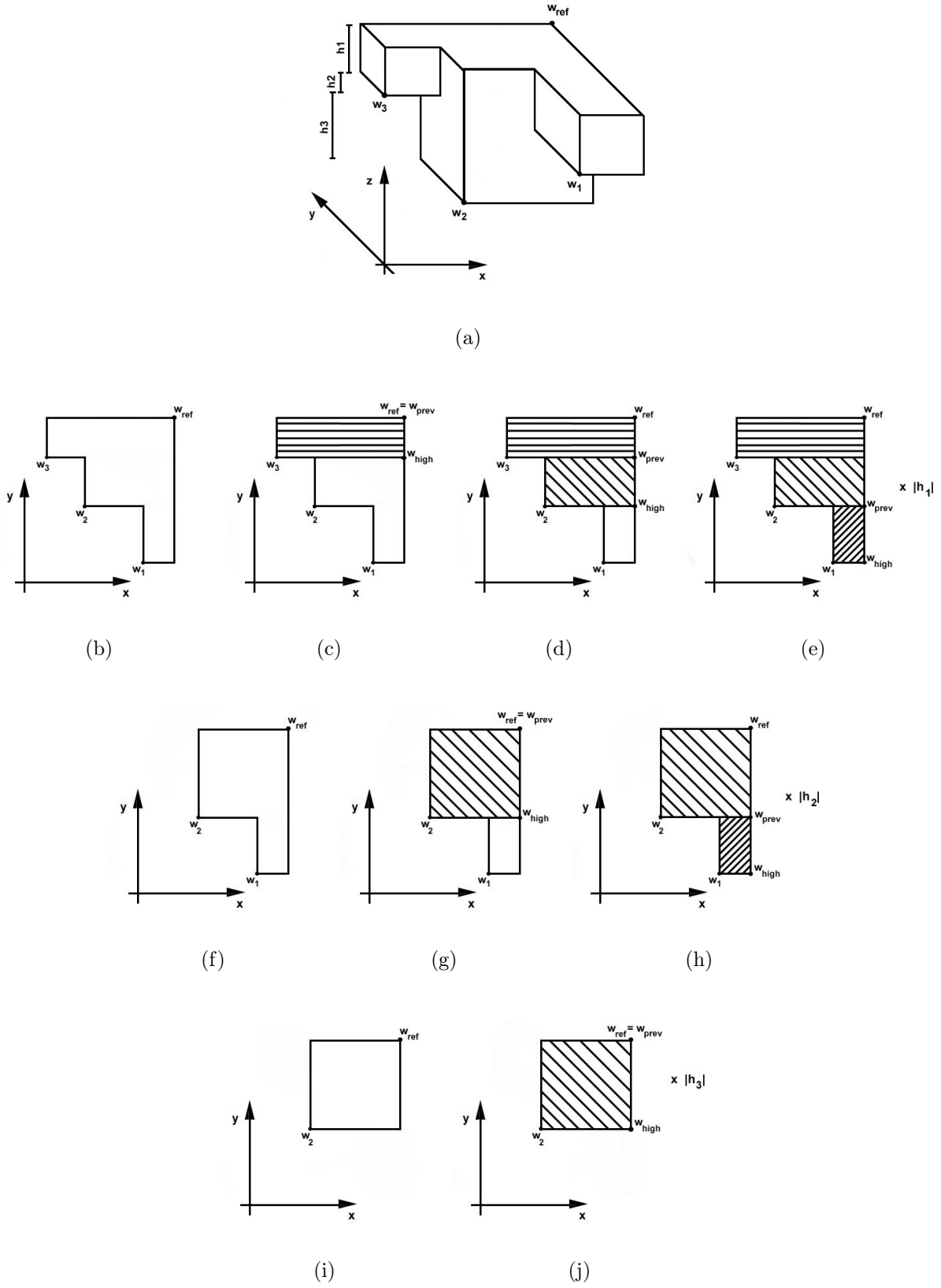


Figura 6.8: Representação gráfica, passo-a-passo, do cálculo do hipervolume, usando a função recursiva $calculaHipervolume(A, \overrightarrow{w}_{ref}, k)$.

6.6.5 Extensão (E) [244]

Seja X um conjunto de vetores no espaço de objetivos. E usa a maior extensão dos vetores de decisão $\overrightarrow{a}, \overrightarrow{b} \in X$ em cada dimensão k para estimar a amplitude pela qual a frente se estende:

$$E = \sqrt{\sum_{i=1}^{k} max\|a_i - b_i\|} \tag{6.13}$$

Onde $\|\cdot\|$ é a norma/distância entre dois pontos. Quanto maior o valor de E, melhor.

Capítulo 7

Calibragem de Parâmetros e Avaliação de Desempenho do MAIS

Para avaliar MAIS, os algoritmos NSGA-II e SPEA2 foram utilizados como *baseline*. Uma vez que MAIS é capaz de lidar com restrições, foi implementada a versão de NSGA-II com restrições (Seção 4.1.3). Há que se destacar que SPEA2 não lida com restrições. Os algoritmos foram implementados em Matlab®.

Infraestrutura Computacional. Os experimentos foram executados em um *cluster* composto por 49 computadores equipados com Intel® Core™ i5-2500 CPU 3.30GHz, 8Gb RAM; e 9 computadores Intel® Core™ Duo CPU E7500 2.93GHz, 2Gb RAM.

Escolha Inicial de Parâmetros. Antes de executar os experimentos propriamente ditos, foram executadas baterias de testes para estimar, empiricamente, os valores iniciais mais adequados para os parâmetros, que são mostrados na Tabela 7.1 (para MAIS) e na Tabela 7.2 (para NSGA-II e SPEA2).

Tabela 7.1: Valor dos parâmetros iniciais para MAIS.

Parâmetro	*Valor*
Tamanho da população (P) (popSize)	500
Tamanho da população secundária/memória (Pm) (popMem)	500
Porcentagem de indivíduos selecionados para clonagem (d)	10%
Quantidade inicial de clones produzidos (cloneNum)	$4 \times P$
Quantidades de pontos mutados k (mutPoints)	1
Probabilidade de mutação (mutRate)	0.5
Taxa de mutação uniforme (pm) (uniformMut)	decrescente iniciando em 60% e terminando em $\frac{1}{k}$, $k = 642$
Quantidade de gerações para a criação de novos Ab's (X)	20
Quantidade de novos Ab's criados a cada X gerações	20% de P

Tabela 7.2: Valor dos parâmetros iniciais para NSGA-II e SPEA.

	Valor	
Parâmetro	*NSGA-II*	*SPEA2*
Tamanho da população (P)	500	500
Probabilidade de *crossover*	0.9	0.9
Operador de *crossover*	*single point crossover (SPX)*	SPX
Seleção	torneio binário	torneio binário
Probabilidade de mutação	0.5	0.5
Taxa de mutação	1/L	1/L

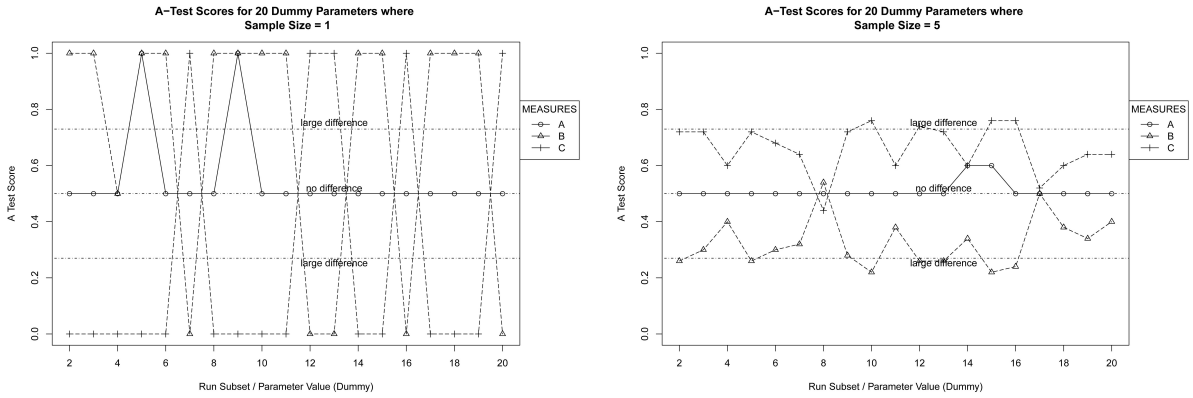
Para assegurar tanto quanto possível uma comparação adequada, adotou-se o critério sugerido por Coello et al. [48] de que todos os algoritmos executassem o mesmo número de avaliações dos objetivos (estabelecido em 500.000 avaliações), garantindo um esforço computacional aproximadamente equivalente. Tamanhos de populações e valores de parâmetros (tanto quanto possível) também foram os mesmos.

7.1 Técnica 1 de *Spartan*

Seguindo o protocolo de *Spartan*, foram analisados 20 conjuntos com amostras de tamanho 1, 5, 10, 50, 100 e 300 execuções, requerendo, portanto, 9.320 execuções individuais de cada um dos três algoritmos (27.960 execuções, no total).

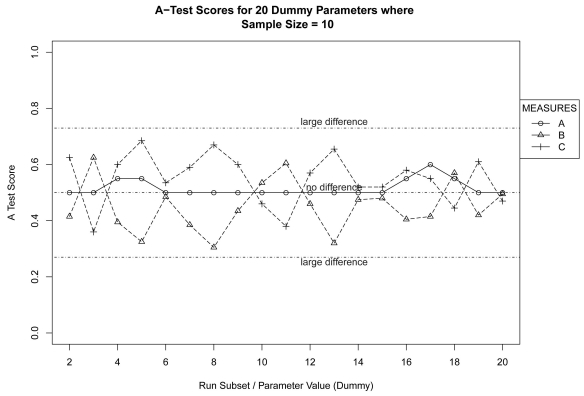
Os resultados são apresentados nas Figuras 7.1, 7.2 e 7.3, para MAIS, NSGA-II e SPEA2, respectivamente, e mostram que para todos os três algoritmos, 300 execuções são suficientes para reduzir a magnitude do efeito da incerteza associada à aleatoriedade nos resultados para menos do que *small* (o nível desejado).

Com isso, todos os demais experimentos foram executados para um único conjunto de 300 experimentos, conferindo uma garantia estatística de que os resultados obtidos não são resultado de fatores relacionados à aleatoriedade da heurística empregada.

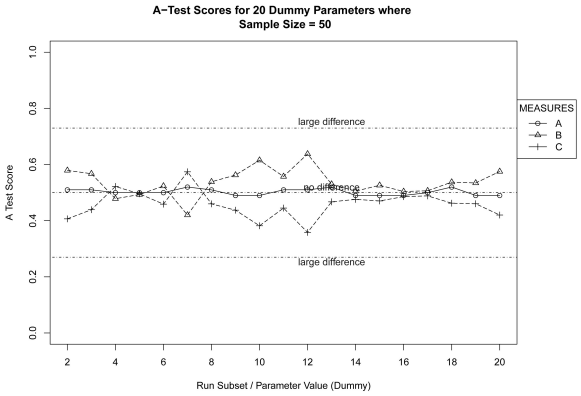


(a)

(b)

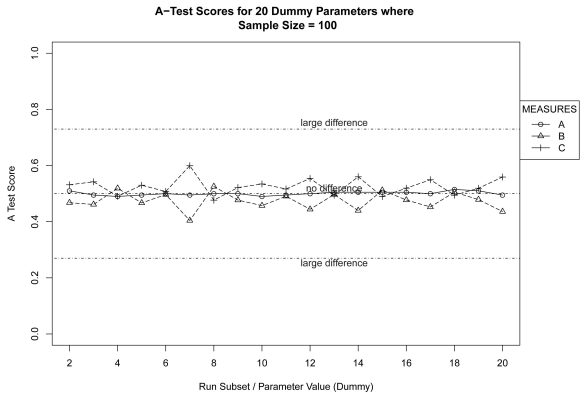


(c)

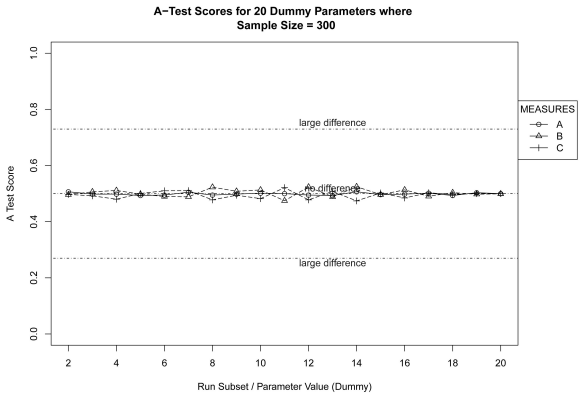


(d)

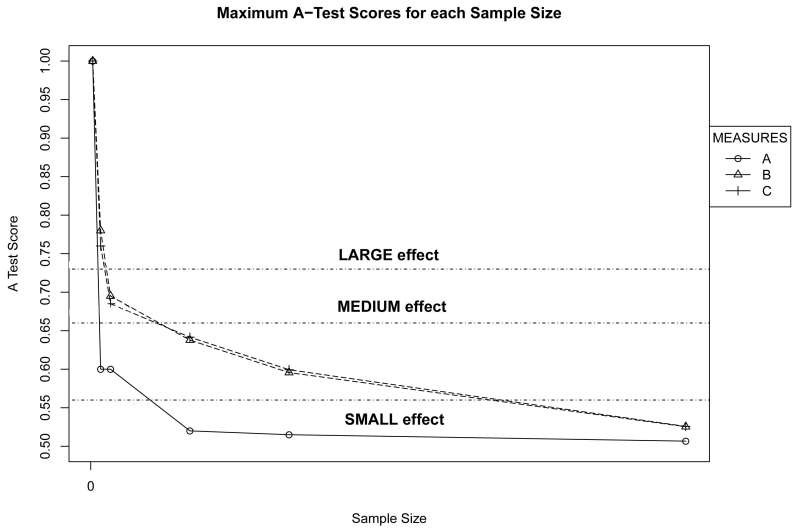
Figura 7.1: Técnica 1 de *Spartan* aplicada a MAIS. (a-f) A-Test para tamanhos de amostra 1, 5, 10, 50, 100 e 300, respectivamente. (g) A-Test final reunindo o resultado de todos os tamanhos de amostra. Com 300 execuções, a estocasticidade sobre os objetivos A (alelos faltantes), B (número de *accessions* selecionados), and C (heterozigose) atinge um efeito desejado (abaixo de *small*).



(e)

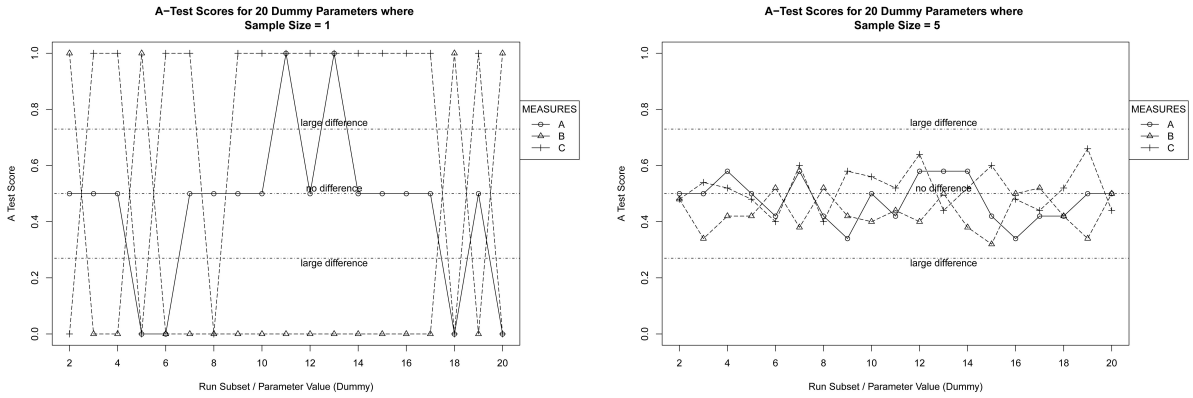


(f)



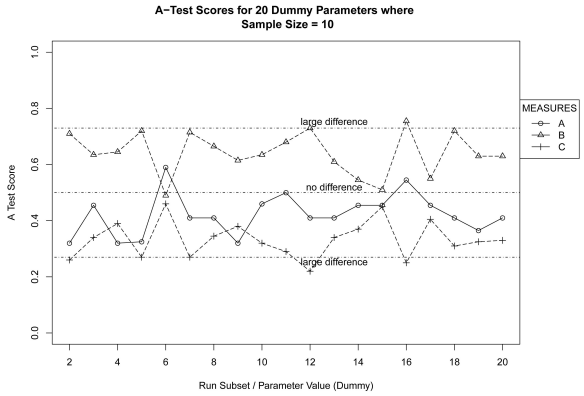
(g)

Figura 7.1: Técnica 1 de *Spartan* aplicada a MAIS. (cont.)

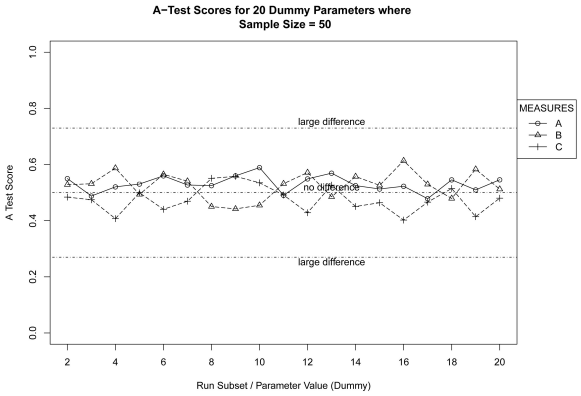


(a)

(b)

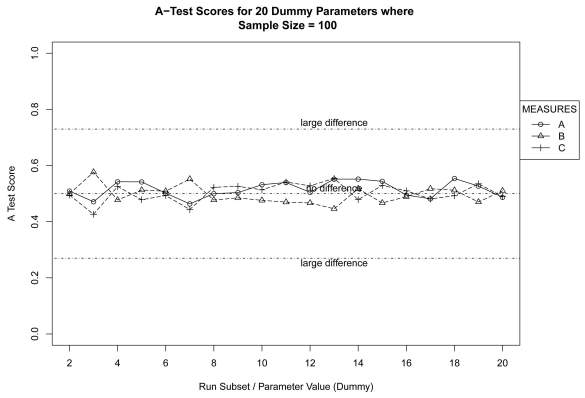


(c)

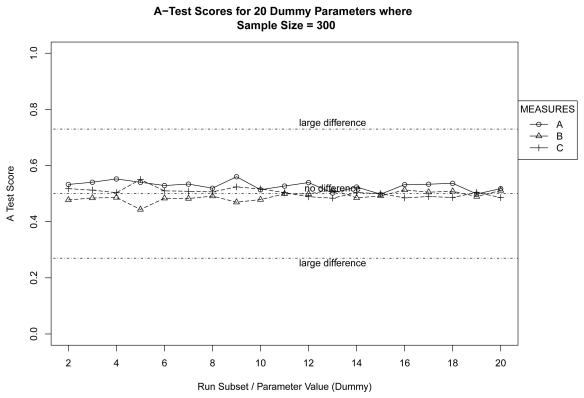


(d)

Figura 7.2: Técnica 1 de *Spartan* aplicada a NSGA-II. (a-f) A-Test para tamanhos de amostra 1, 5, 10, 50, 100 e 300, respectivamente. (g) A-Test final reunindo o resultado de todos os tamanhos de amostra. Com 300 execuções, a estocasticidade sobre os objetivos A (alelos faltantes), B (número de amostras selecionadas), and C (heterozigose) atinge um efeito desejado (abaixo de *small*).



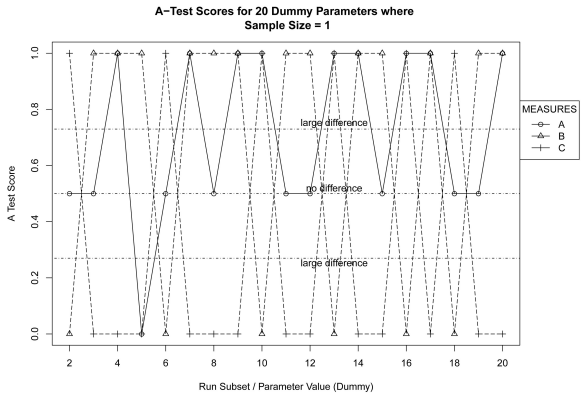
(e)



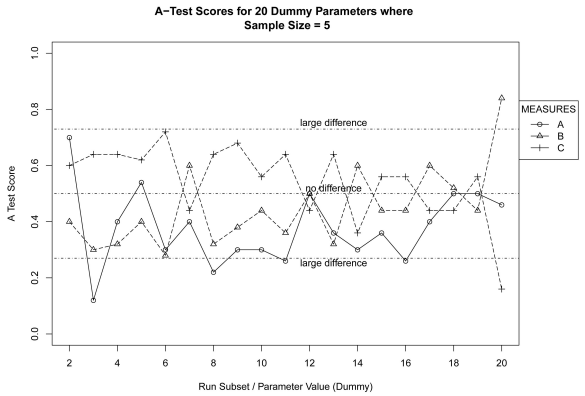
(f)

(g)

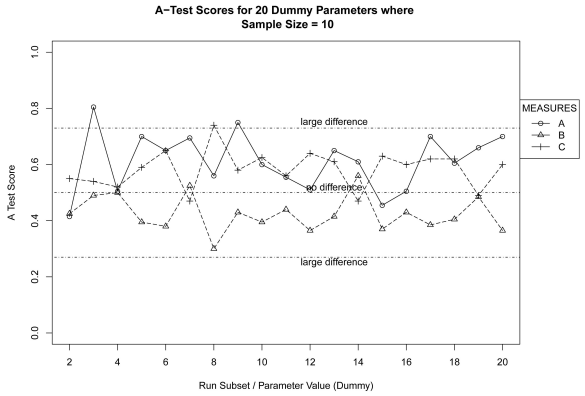
Figura 7.2: Técnica 1 de *Spartan* aplicada a NSGA-II. (cont.)



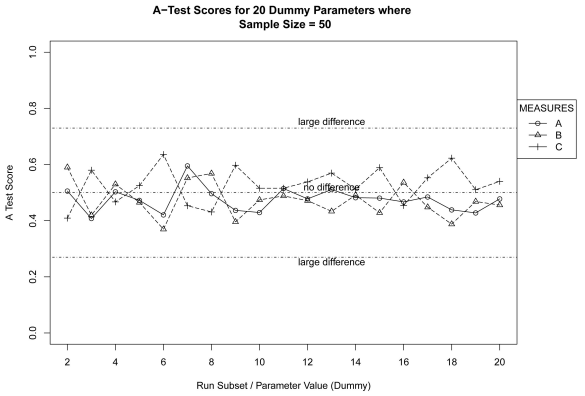
(a)



(b)

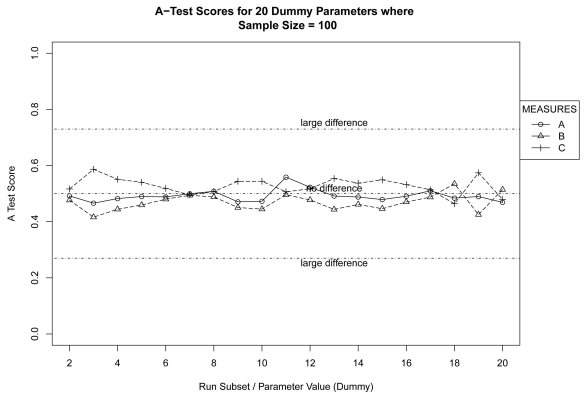


(c)

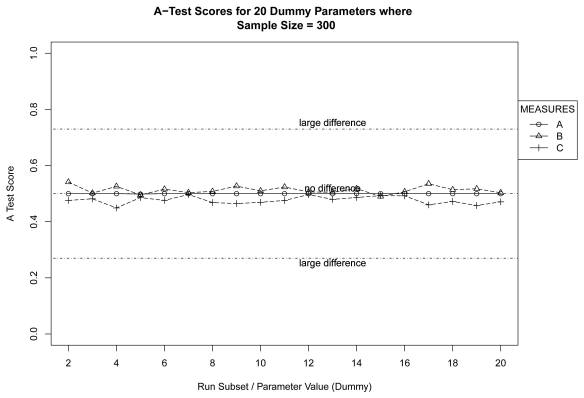


(d)

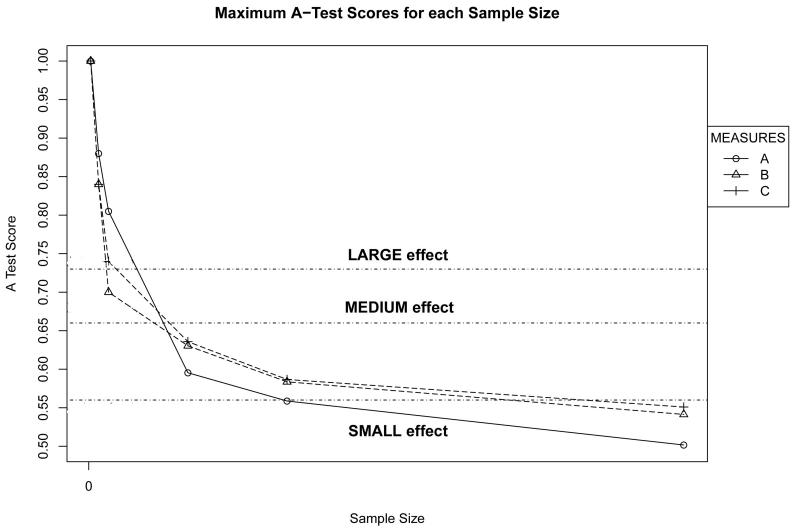
Figura 7.3: Técnica 1 de *Spartan* aplicada a SPEA2. (a-f) A-Test para tamanhos de amostra 1, 5, 10, 50, 100 e 300, respectivamente. (g) A-Test final reunindo o resultado de todos os tamanhos de amostra. Com 300 execuções, a estocasticidade sobre os objetivos A (alelos faltantes), B (número de amostras selecionadas), and C (heterozigose) atinge um efeito desejado (abaixo de *small*).



(e)



(f)



(g)

Figura 7.3: Técnica 1 de *Spartan* aplicada a SPEA2. (cont.)

7.2 Métricas

Para avaliar o desempenho de MAIS efetuou-se a comparação do mesmo com NSGA-II e SPEA2 utilizando-se as métricas definidas na Seção 6.6.

Em que pese saber que 300 execuções seriam suficientes para minimizar a aleatoriedade dos algoritmos estocásticos empregados, como tinha-se um total de 9.320 execuções individuais de cada algoritmo como resultado da aplicação da Técnica 1 de *Spartan*, optou-se por utilizar todo o conjunto de 20 pastas de 300 execuções, num total de 6.000 execuções individuais de cada algoritmo para o cálculo das métricas.

Cumpre destacar que, como SPEA2 não lida com restrições, para possibilitar uma comparação adequada com os demais algoritmos, as métricas foram calculadas considerando-se apenas as soluções contidas no espaço de soluções delimitado pelas restrições expressas nas Equações 6.6 e 6.7.

Função C

Para o cálculo da função C, para cada algoritmo, o resultado de 6.000 execuções individuais foram unificadas, as soluções dominadas foram removidas do conjunto união formando uma frente global de soluções não-dominadas para cada algoritmo ($\overrightarrow{X}$) [244]. A partir de tal frente, os valores da função C foram calculados considerando os algoritmos dois a dois, os resultados são apresentados na Tabela 7.3.

Tabela 7.3: Resultados para função C calculada em pares, onde $\overrightarrow{X}'$ corresponde à frente global de soluções não-dominadas para o primeiro algoritmo e $\overrightarrow{X}''$ para o segundo algoritmo. Melhores resultados são indicados em negrito.

	$C(MAIS, \overrightarrow{X}'')$	$C(NSGA-II, \overrightarrow{X}'')$	$C(SPEA2, \overrightarrow{X}'')$
$C(\overrightarrow{X}', MAIS)$	X	0,774075	0,795735
$C(\overrightarrow{X}', NSGA-II)$	**0,835213**	X	**0,971955**
$C(\overrightarrow{X}', SPEA2)$	**0,956174**	0,890803	X

Verificou-se que ***C(MAIS, NSGA-II)= 0,835213*** e *C(NSGA-II,MAIS) = 0,774075*, significando que apesar de discretamente, soluções encontradas por MAIS "cobriram" NSGA-II, i.e., há mais soluções de MAIS que dominam soluções de NSGA-II do que o contrário. O mesmo se pode dizer com relação a MAIS e SPEA2 (***C(MAIS, SPEA2)= 0,956174*** e *C(SPEA2,MAIS) = 0,795735*). Por fim, SPEA2 teve melhor resultado do que NSGA-II (***C(SPEA2, NSGA-II)= 0,971955*** e *C(NSGA-II,SPEA2) = 0,890803*). Os dados apresentados acima suportam a hipótese que MAIS é capaz de encontrar melhores aproximações à verdadeira frente de Pareto (PF_{true}) – ainda que, neste problema do mundo real, PF_{true} não seja conhecida.

O restante das métricas foi calculado considerando-se as 6.000 execuções individuais dos algoritmos (sem consolidá-las/unificá-las).

EAF

As superfícies EAF obtidas para cada algoritmo são apresentadas na Figura 7.4.

Constata-se que MAIS foi capaz de encontrar menores *core collections* e mais próximas à origem dos eixos (i.e., provavelmente mais próximas de PF_{true}). Além disso, MAIS tem soluções mais regulares e suavemente distribuídas no espaço de busca delimitado pelas restrições, sendo capaz de melhor explorá-lo.

Em que pese SPEA2 ter sido superado por MAIS, cumpre ressaltar que, ao não lidar com restrições, com o mesmo número de avaliações das funções objetivo, ele explorou uma área mais extensa do espaço de soluções. O efeito colateral disso pode ser visto na Figura 7.4g, comparada às Figuras 7.4b e 7.4d. A frequência com que soluções são encontradas por SPEA2 diminui drasticamente (praticamente todas têm frequência abaixo de 0,1), diminuindo ainda mais a probabilidade de se encontrar uma "melhor solução", dado que os algoritmos trabalhados têm natureza estocástica.

Hipervolume, espaçamento e extensão

As Tabelas 7.4, 7.5 e 7.6 apresentam, respectivamente, os valores das métricas hipervolume, extensão e espaçamento, obtidos pelos algoritmos e a representação gráfica de tais valores é feita na Figura 7.5.

Tabela 7.4: Hipervolume para MAIS, NSGA-II e SPEA2 calculado para 6.000 execuções independentes. Melhores resultados são indicados em negrito.

	Hipervolume (H)		
	MAIS	NSGA-II	SPEA2
Média	$\mathbf{7{,}75216}\times10^7$	$7{,}12362\times10^7$	$7{,}48034\times10^7$
(Desvio Padrão)	$\mathbf{(0{,}34316}\times10^6\mathbf{)}$	$(1{,}58110\times10^6)$	$(2{,}22650\times10^6)$
Melhor	$7{,}87650\times10^7$	$7{,}55250\times10^7$	$\mathbf{7{,}96500}\times10^7$
Mediana	$\mathbf{7{,}75350}\times10^7$	$7{,}13380\times10^7$	$7{,}50175\times10^7$
Pior	$\mathbf{7{,}58140}\times10^7$	$6{,}39660\times10^7$	$6{,}75400\times10^7$

Para a métrica E, a fim de minimizar o impacto de *outliers*, procedeu-se à exclusão dos mesmos utilizando o *z-score*[1] com uma confiança de 99%.

[1] O *z-score* nada mais é do que uma maneira de padronizar os dados exprimindo seus valores em termos de uma distribuição com um média de valor 0 e um desvio padrão de valor 1. Tomando o valor absoluto de *z-score* (i.e., ignorando se ele é positivo ou negativo), em uma distribuição normal, espera-se que cerca de 5% dos valores absolutos sejam maiores que 1,96 e 1% tenham valor absoluto maior do que 2.58 e nenhum tenha valor maior do que 3,29.

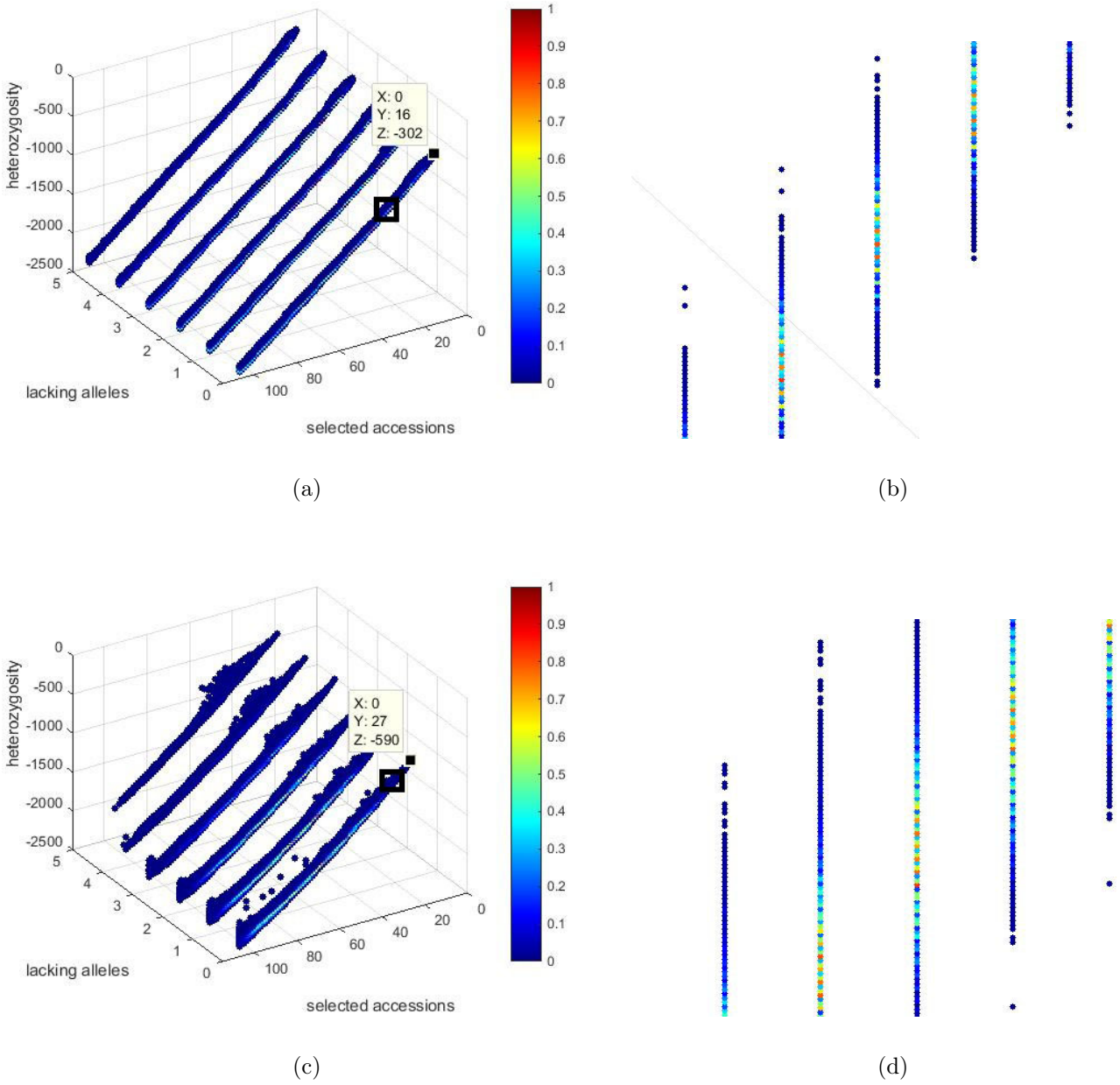
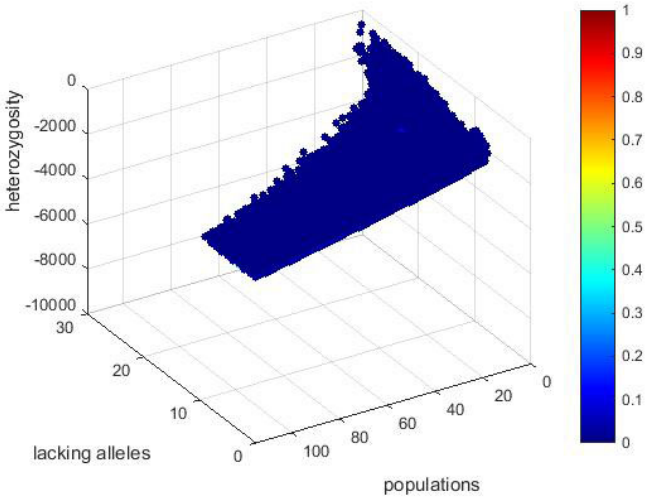
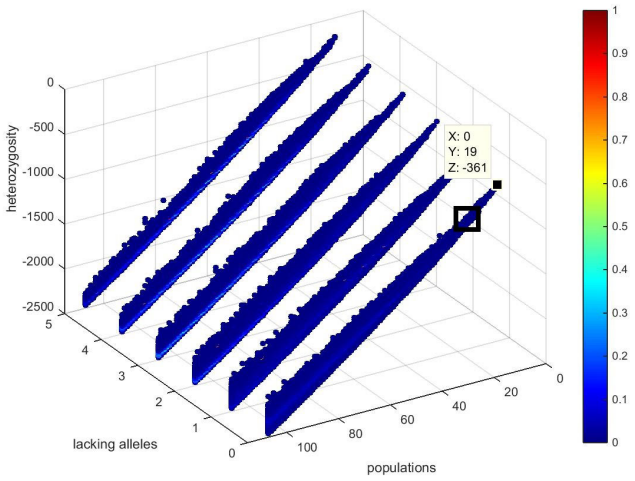


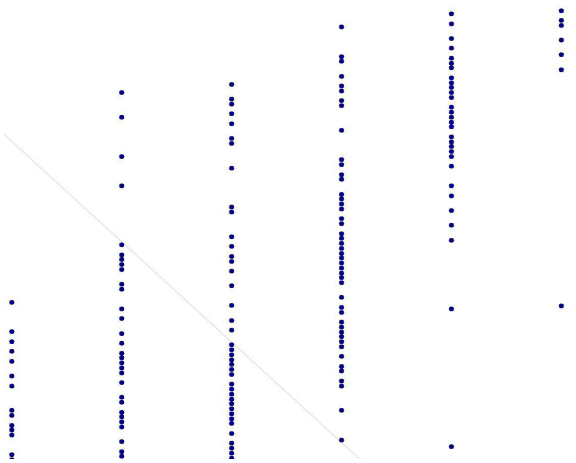
Figura 7.4: Superfícies EAF obtidas para os algoritmos: (a) MAIS; (b) aproximação de 6× da região marcada por um quadrado em *a*; (c) NSGA-II; (d) aproximação de 6× da região marcada por um quadrado em *c*; (e) SPEA2 – toda a superfície obtida pelo algoritmo; (f) SPEA2 – região delimitada pelas restrições expressa pelas Equações 6.6 e 6.7; (g) aproximação de 6× da região marcada por um quadrado em *f*. A frequência com que as soluções são encontradas é indicada pela gradação de cores da escala de referência, variando de 0 a 1, azul escuro e vermelho escuro, respectivamente. Cabe destacar que todos os pontos *plotados* são ótimos no sentido que nenhum é melhor no contexto considerado e na ausência de preferências expressas pelo decisor. Em (a), (c) e (f), as legendas indicam as soluções com nenhum alelo faltante que correspondem aos menores conjuntos de *accessions* selecionados. X - alelos faltantes; Y - número de *accessions* selecionados; Z - heterozigose.



(e)

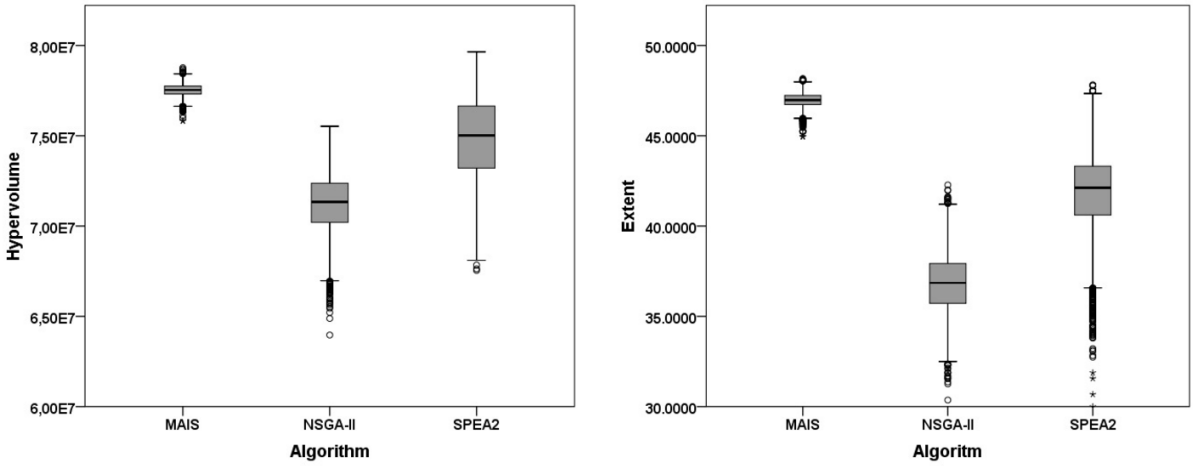


(f)



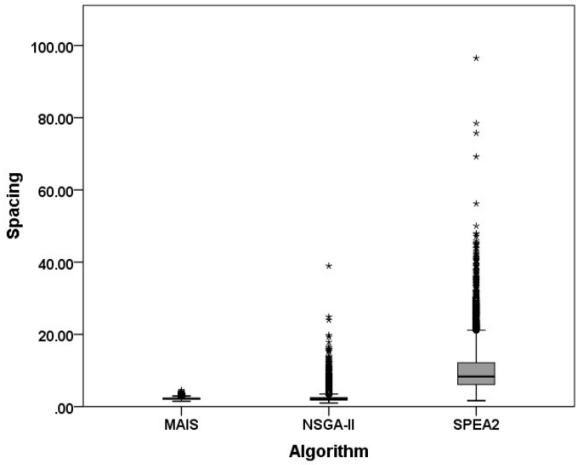
(g)

Figura 7.4: Superfícies EAF (cont.)

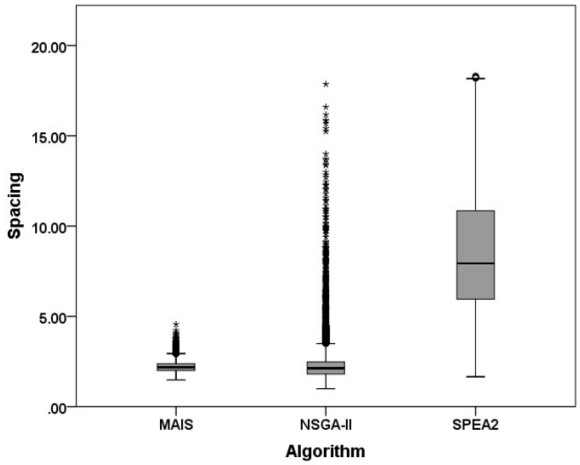


(a)

(b)



(c)



(d)

Figura 7.5: *Boxplot* das métricas para MAIS, NSGA-II e SPEA2. (a) Hipervolume (H); (b) Extensão (E); (c) Espaçamento (incluídos *outliers*); (d) Espaçamento com *z-score* (excluídos *outliers* considerada uma confiança de 99%). Para H e E, quanto maior o valor da métrica, melhor; para S, quanto menor o valor (mais próximo de zero), melhor.

Tabela 7.5: Extensão para MAIS, NSGA-II e SPEA2 calculado para 6.000 execuções independentes. Melhores resultados são indicados em negrito.

	Extensão (E)		
	MAIS	NSGA-II	SPEA2
Média	**46,9660**	36,8337	41,8190
(Desvio Padrão)	**(0,3982)**	(1,6244)	(2,1593)
Melhor	**48,1660**	42,2730	47,8120
Mediana	**46,9790**	36,8510	42,1190
Pior	**44,9440**	30,3640	30,0000

Tabela 7.6: Espaçamento para MAIS, NSGA-II e SPEA2 calculado para 6.000 execuções independentes. Melhores resultados são indicados em negrito. N corresponde ao tamanho da amostra.

	Espaçamento			Espaçamento com *z-score*		
	MAIS	NSGA-II	SPEA2	MAIS	NSGA-II	SPEA2
N	6.000	6.000	6.000	6.000	5.994	5.428
Média	**2,2196**	2,4452	10,0492	**2,2196**	2,4235	8.5030
(Desvio Padrão)	**(0,3183)**	(1,5347)	(6,2475)	**(0,3183)**	(1,3551)	(3,3182)
Melhor	1,4800	**0,9900**	1,6500	1,4800	**0,9900**	1,6500
Mediana	2,1771	**2,1209**	8,3164	2,1771	**2,1204**	7,8767
Pior	**4,5600**	38,9400	96,4800	**4,5600**	16,5900	17,2900

Da observação da linha N da Tabela 7.6, na coluna referente aos valores para a métrica E com *z-score*, verifica-se que MAIS produziu resultados estatisticamente mais regulares, não gerando *outliers*, enquanto que NSGA-II e SPEA2 os produziram.

MAIS superou NSGA-II e SPEA2 em todas as métricas testadas.

A análise gráfica das superfícies EAF (Figura 7.4), associada aos valores dos resultados das funções C (Tabela 7.3) e da métrica H (Tabela 7.4), corroboram a hipótese que MAIS foi capaz de encontrar soluções mais próximas da PF_{true}.

Além disso, MAIS foi capaz de encontrar *core collections* mais regularmente distribuídas ao longo da PF_{known}, como pode ser visto para a métrica S (Tabela 7.6 e Figura 7.5d).

A métrica E (Tabela 7.5 e Figura 7.5b) mostra que MAIS foi mais efetivo em estender a frente não-dominada, explorando uma região mais ampla do espaço de objetivos.

Dados estes pontos, pode afirmar-se que para o problema de encontrar uma *core collection*, MAIS apresentou melhor desempenho consideradas as métricas utilizadas quando comparado a NSGA-II e SPEA2, sendo capaz de encontrar *core collections* menores e melhor distribuídas, bem como de explorar uma região mais extensa do espaço de objetivos.

Neste ponto, cabe uma ressalva: em que pese o cumprimento das metas apresentadas na Seção 6.6 ser um "objetivo" a ser buscado em termos de otimização (*computacionalmente falando*), ainda não está claro, em termos biológicos, se o cumprimento de tais

metas de otimização são realmente desejáveis quando se fala em SCP, i.e., se *biologicamente falando* o perseguimento de tais metas se reflete em resultados melhores, é o caso, por exemplo, da Meta 2. Este é um tema de pesquisa em aberto, cabendo análise e interpretação dos resultados obtidos com vistas à construção conjunta de conhecimento.

Tempo de execução

Apresenta-se, nesta seção, a título de curiosidade, os tempos médios de execução individual de MAIS, NSGA-II e SPEA2 observados empiricamente.

Ressalte-se que, ainda que não se tivesse por objetivo avaliar a complexidade, o desempenho, custo e eficência dos algoritmos em termos de quantidade de recursos utilizados/tempo de execução (mas sim em relação aos resultados obtidos, estes avaliados em termos das métricas definidas na Seção 6.6), considera-se interessante apresentar os dados relativos ao tempo de execução.

A execução dos três algoritmos, utilizando a infraestrutura computacional retro-mencionada e medindo diretamente o tempo das execuções individuais de cada algoritmo quando da aplicação da Técnica 1 de *Spartan*, revelou os tempos médios de execução apresentados na Tabela 7.7.

Tabela 7.7: Tempo médio de execução individual dos algoritmos quando aplicada a Técnica 1 de *Spartan*. Melhores resultados são indicados em negrito. Tempo apresentado como hh:mm:ss (horas:minutos:segundos).

	Tempo médio de execução	
	Média	(Desvio Padrão)
MAIS	**01:21:55**	**(00:01:30)**
NSGA-II	01:46:15	(00:14:59)
SPEA2	08:37:55	(01:13:59)

As medidas de tempo obtidas desta forma são empíricas e, sem um aprofundamento, os resultados não devem ser generalizados[2], mas servem como uma estimativa interessante.

7.3 Modelo Nulo

Foram geradas aleatoriamente 20 populações de 500 indivíduos a fim de verificar se não seriam obtidos resultados melhores do que os alcançados ao final da execução dos algoritmos MOO.

[2]Os principais motivos trazidos pela literatura para a não-generalização seriam: (i) os resultados são dependentes do compilador que pode favorecer algumas construções em detrimento de outras; (ii) os resultados dependem do hardware; (iii) quando grandes quantidades de memória são utilizadas, as medidas de tempo podem depender deste aspecto[249].

O valor da função C para os algoritmos utilizados neste estudo comparados ao modelo nulo são mostrados na Tabela 7.8 e servem como *tournament performance* mencionado na Seção 3.2.

Tabela 7.8: Resultados para função C calculada em pares considerando o modelo nulo. Melhores resultados são indicados em negrito.

Função C	Valor	Função C	Valor
$C(MAIS, NULO)$	**0,592292**	$C(NULO, MAIS)$	0,000000
$C(NSGA-II, NULO)$	**0,578623**	$C(NULO, NSGA-II)$	0,000000
$C(SPEA2, NULO)$	**0,598664**	$C(NULO, SPEA)$	0,000000

O modelo nulo encontrou apenas 45 soluções com nenhum alelo faltante, e dentre estas soluções, aquela com menor número de *accessions* selecionou 187. Ademais, considerando-se as restrições expressas nas Equações 6.6 e 6.7, das 10.000 soluções geradas, apenas 3.969 apresentaram até 5 alelos faltantes.

7.4 Calibragem de parâmetros para MAIS

7.4.1 Técnica 2 de *Spartan*

Nesta etapa, MAIS foi testado quanto à robustez frente a perturbações nos parâmetros constantes da Tabela 7.9.

Tabela 7.9: Limites mínimo e máximo, bem como incrementos para os valores de parâmetros.

parâmetro		*baseline*	mínimo	máximo	incremento
tamanho da população	popSize	500	100	800	100
tamanho da memória	popMem	100	500	800	100
mutação uniforme	uniformMut	0,6	0,2	1	0,2
número de clones a serem criados	cloneNum	4	1	6	1
número de mutações (hipermutação)	mutPoints	3	1	5	1
taxa da hipermutação	mutRate	0,5	0,1	0,9	0,2

O *baseline* usado assumiu os valores de referência na literatura, bem como valores empiricamente estabelecidos (utilizando resultados de experimentos realizados previamente [210]). Mas de acordo com Read et al. [187], poderiam ter sido usados quaisquer pontos de interesse no espaço de parâmetros (dentro dos limites máximo e mínimo definidos para o parâmetro).

Foram analisados **37 conjuntos** (resultado da combinação dos diferentes valores atribuídos a cada parâmetro) com 300 execuções cada (de acordo com os resultados obtidos

com a Técnica 1 de *Spartan*), num total de **11.100 execuções individuais do algoritmo**. Os resultados são apresentados nas Figuras 7.6 a 7.10.

Este método essencialmente indica se uma alteração no valor do parâmetro leva a uma mudança *cientificamente significativa* quando comparada à execução *baseline*, o que revela quão robusto o algoritmo é à alteração daquele parâmetro e os pontos nos quais a perturbação do parâmetro resulta em mudanças significativas do algoritmo. Os resultados dos A-testes para cada parâmetro são apresentados nos gráficos (d) das Figuras 7.6 a 7.10.

Para a instância do problema tratada[3], considerando o tamanho da população de Ab's, é possível observar que o algoritmo se mostra altamente sensível à variação do tamanho da população na faixa entre 100 a 500 Ab's (Figura 7.6d), logo, os resultados obtidos com tais parâmetros devem ser considerados com cautela, pois podem ser artefato da parametrização. Isso faz sentido, pois para populações pequenas, o tamanho da população não é adequado para garantir a manuteção de indivíduos bons em quantidade suficiente para uma evolução adequada dos Ab's ao longo das gerações (no passo 11 descrito na Seção 5.2 eles seriam eliminados). Para o valor 600, a variação é média e a partir de 700, o tamanho da população passa a ter pequena influência nos resultados do algoritmo (o que é desejado).

Esta técnica também fornece indicações quanto aos valores mais adequados a serem utilizados (podendo ser usada em associação à Técnica 3 tratada na seção seguinte).

Quanto ao objetivo de minimizar alelos falantes, como o valor oscilou entre quantidades muito próximas (0-2) os gráficos não são tão informativos. Já para o número de *accessions* selecionados, como busca-se o menor valor, pela inspeção visual, é possível ter indicações dos melhores valores a serem empregados, o mesmo valendo para o valor de heterozigose (procura-se o menor valor – o valor mais negativo), e.g., para o tamanho da população, a Figura 7.6b indica que valores a partir de 500 são mais adequados quando se considera a minimização de *accessions*. Tal análise, é corroborada pela Figura 7.6d.

Para o tamanho da população secundária (memória), a Figura 7.7 indica que de 100 a 400 tem-se uma mudança cientificamente significativa nos resultados e somente valores a partir 500 seriam adequados.

Já para a mutação uniforme, valores extremos (muito baixos, 0,2 ou muito altos, 0,8-1,0) representam impacto entre médio e alto no algoritmo (Figura 7.8). O comportamento mais estável é encontrado para valores entre 0,2-0,8. Se a mutação ocorre a uma taxa muito baixa, o algoritmo quase não se beneficia de tal operador, por outro lado, se ela é muito alta, são introduzidas "muitas" alterações de maneira que uma boa solução pode ser perdida. O mesmo comportamento com valores localizados nos extremos dos limites

[3]Quantidade de alelos distintos = 55; quantidade de *accessions* = 642; número de objetivos a serem otimizados = 3.

máximos e mínimos dos parâmetros foi observado para taxa de hipermutação (Figura 7.10) e para o número de mutações (hipermutação) (Figura 7.11).

A variação no número de clones gerados não demonstrou ter um impacto significativo no funcionamento do algoritmo e nos resultados obtidos (Figura 7.9). Isto é interessante do ponto de vista de desenvolvimento do algoritmo, pois tem-se uma indicação clara de que esse parâmetro não é crítico, podendo-se escolher um valor considerado adequado[4] e excluí-lo da calibragem.

7.4.2 Técnica 3 de *Spartan*

Essa técnica revela o efeito no algoritmo quando dois ou mais parâmetros são ajustados simultaneamente, pois, não raras vezes, um parâmetros pode sofrer influência de outro. A análise da sensibilidade global revela tais efeitos, mostrando como diferentes parâmetros estão relacionados e indicando os que têm maior influência nos resultados do algoritmo [159] .

Com base nos valores mínimo e máximo estabelecidos para cada parâmetro (Tabela 7.9), foram definidos **50 conjuntos** de valores de parâmetros pela amostragem por LHS na Técnica 3 de *Spartan*, tais conjuntos são listados na Tabela 7.10.

Para cada uma das 50 combinação de parâmetros gerada pelo LHS, foram efetuadas 300 execuções do algoritmo (conforme determinado pela Técnica 1 de *Spartan*), num total de **15.000 execuções individuais do algoritmo**. O gráfico para cada par parâmetro-objetivo, revelando a correlação entre eles quantificada por meio do PRCC, é mostrada na Figura 7.12.

Segundo Alden et al. [7], o PRCC considera a relação não-linear entre parâmetro-objetivo e corrige o efeito de outros parâmetros na resposta, fornecendo um indicador do efeito do parâmetro na resposta do algoritmo apesar dos outros parâmetros também serem perturbados.

Mais uma vez, assim como para a Técnica 2, como o valor de alelos faltantes oscilou entre quantidades muito próximas (0-2), os gráficos para esse objetivo (A) não são tão informativos, à exceção do parâmetro tamanho da população (popSize) (Figura 7.12a), para o qual o PRCC é de -0,523, indicando uma correlação negativa média entre o tamanho da população e o número de alelos faltantes, no sentido que, ao aumentar-se o tamanho da população, o número de alelos faltantes diminuiu (um efeito interessante do ponto de vista da minimização almejada).

[4]Em que pese não ter impacto significativo na dinâmica do algoritmo, o número de clones tem impacto no tempo de execução do mesmo. Quantidades grandes de clones levam a um tempo de execução aumentado.

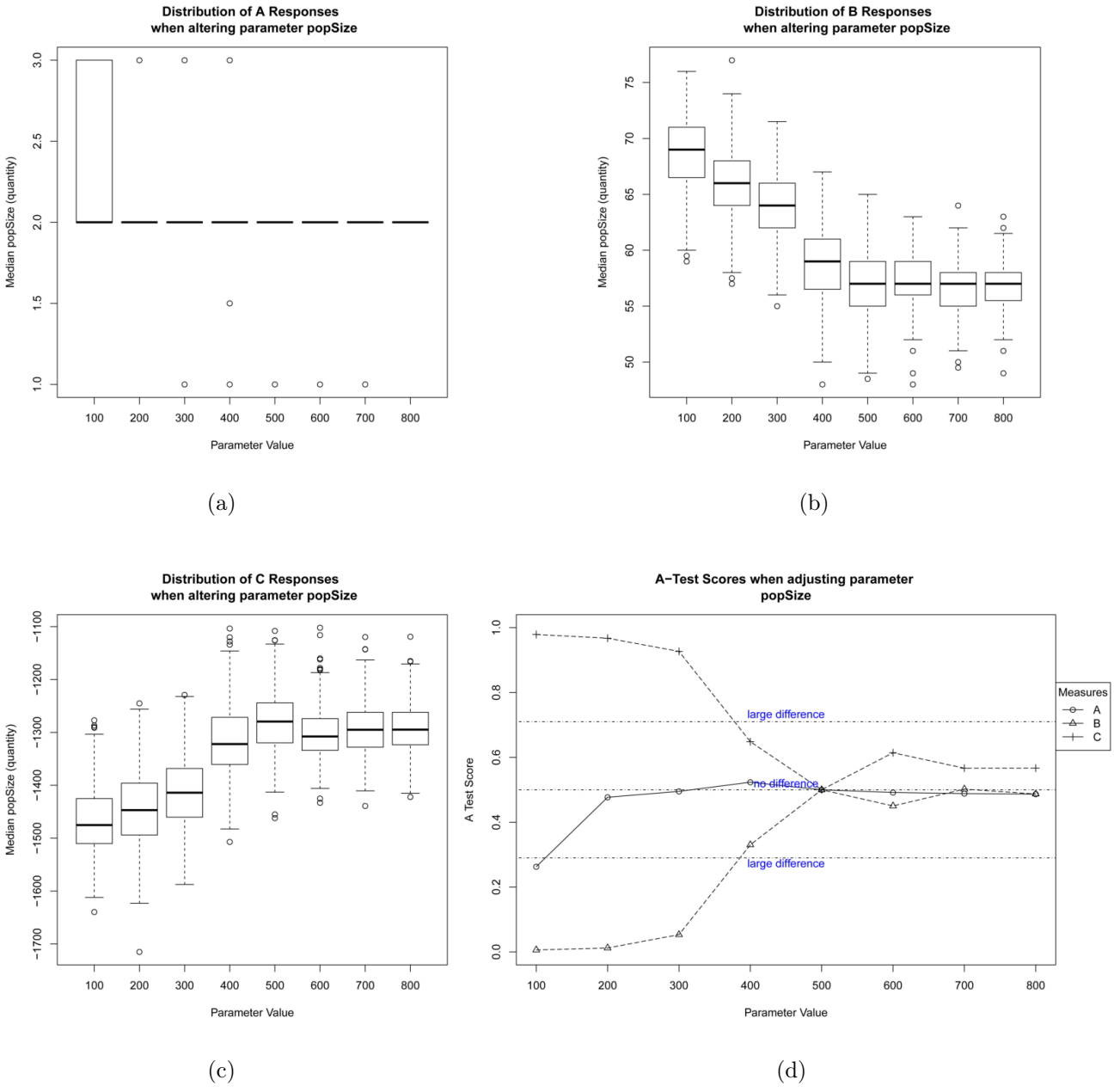


Figura 7.6: Técnica 2 de *Spartan* aplicada a MAIS para o parâmetro tamanho da população (popSize), indicando o impacto da variação do parâmetro sobre: (a) alelos faltantes; (b) número de *accessions* selecionados; (c) heterozigose. (d) Efeito da magnitude do parâmetro sobre os diferentes objetivos. A – alelos faltantes; B – número de *accessions* selecionados; e C – heterozigose.

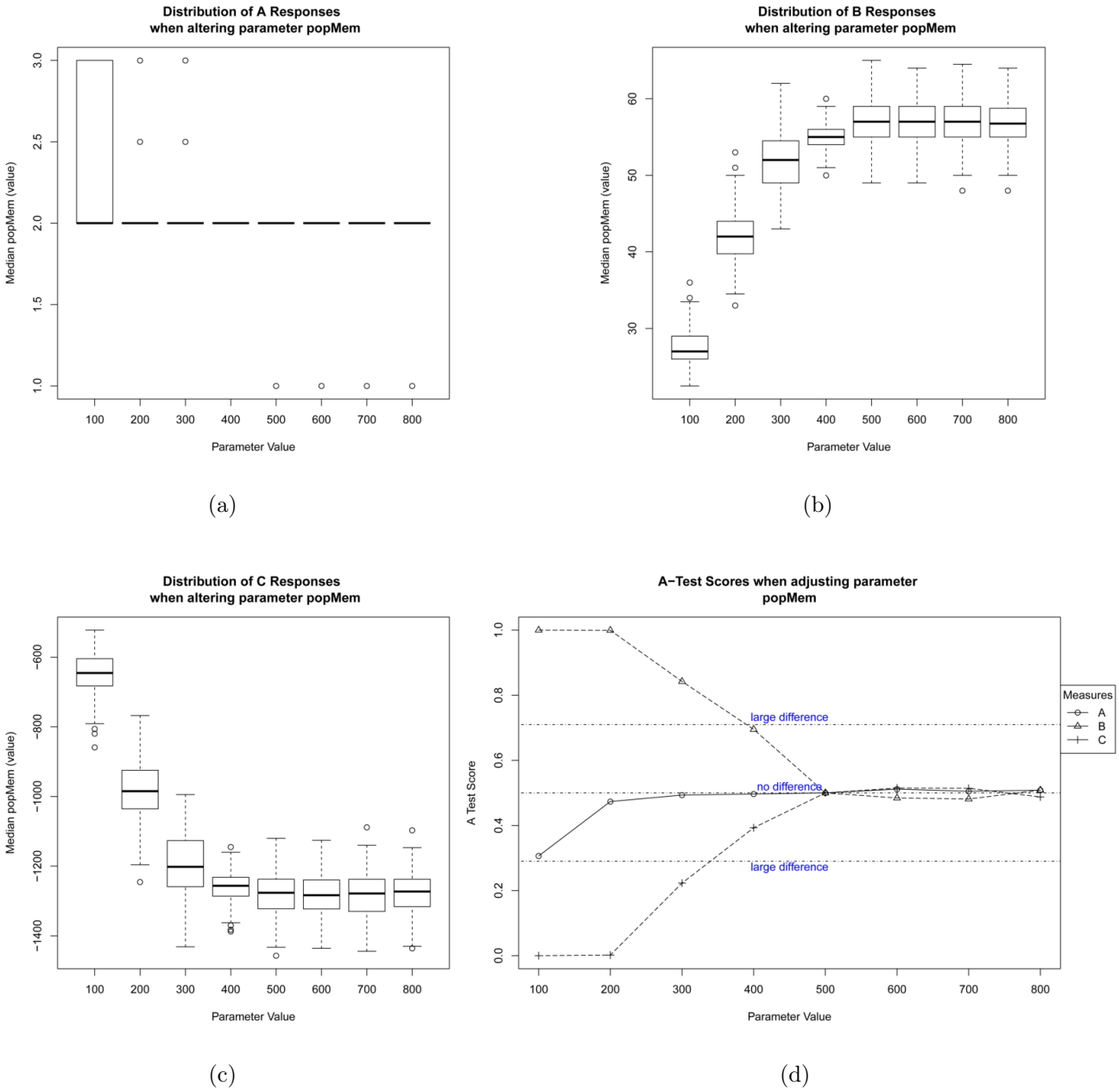


Figura 7.7: Técnica 2 de *Spartan* aplicada a MAIS para o parâmetro tamanho da população secundária (popMem), indicando o impacto da variação do parâmetro sobre: (a) alelos faltantes; (b) número de *accessions* selecionados; (c) heterozigose. (d) Efeito da magnitude do parâmetro sobre os diferentes objetivos. A – alelos faltantes; B – número de *accessions* selecionados; e C – heterozigose.

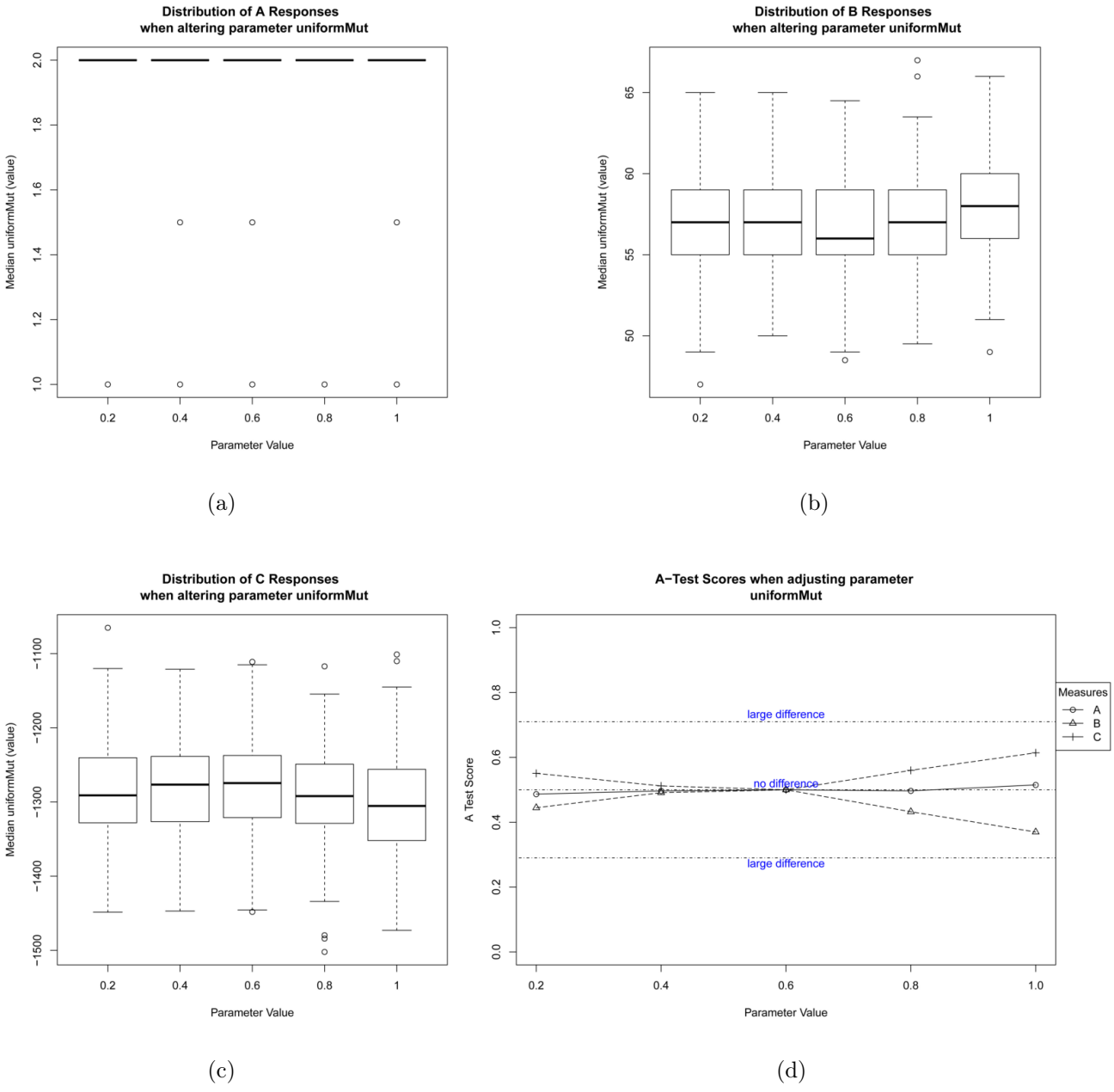


Figura 7.8: Técnica 2 de *Spartan* aplicada a MAIS para o parâmetro mutação uniforme (uniformMut), indicando o impacto da variação do parâmetro sobre: (a) alelos faltantes; (b) número de *accessions* selecionados; (c) heterozigose. (d) Efeito da magnitude do parâmetro sobre os diferentes objetivos. A – alelos faltantes; B – número de *accessions* selecionados; e C – heterozigose.

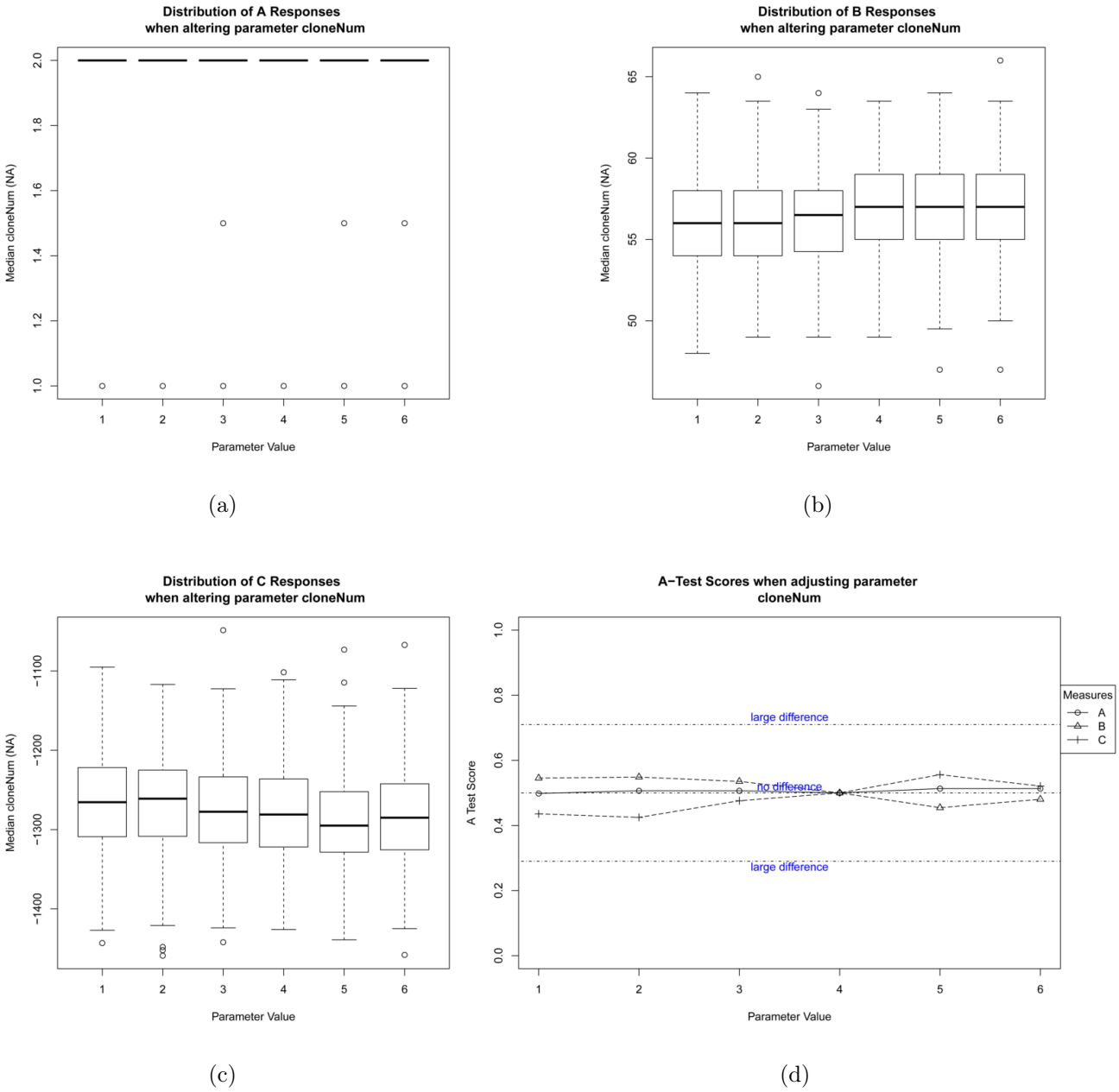


Figura 7.9: A Técnica 2 de *Spartan* aplicada a MAIS para o parâmetro número de clones a serem criados (cloneNum), indicando o impacto da variação do parâmetro sobre: (a) alelos faltantes; (b) número de *accessions* selecionados; (c) heterozigose. (d) Efeito da magnitude do parâmetro sobre os diferentes objetivos. A – alelos faltantes; B – número de *accessions* selecionados; e C – heterozigose.

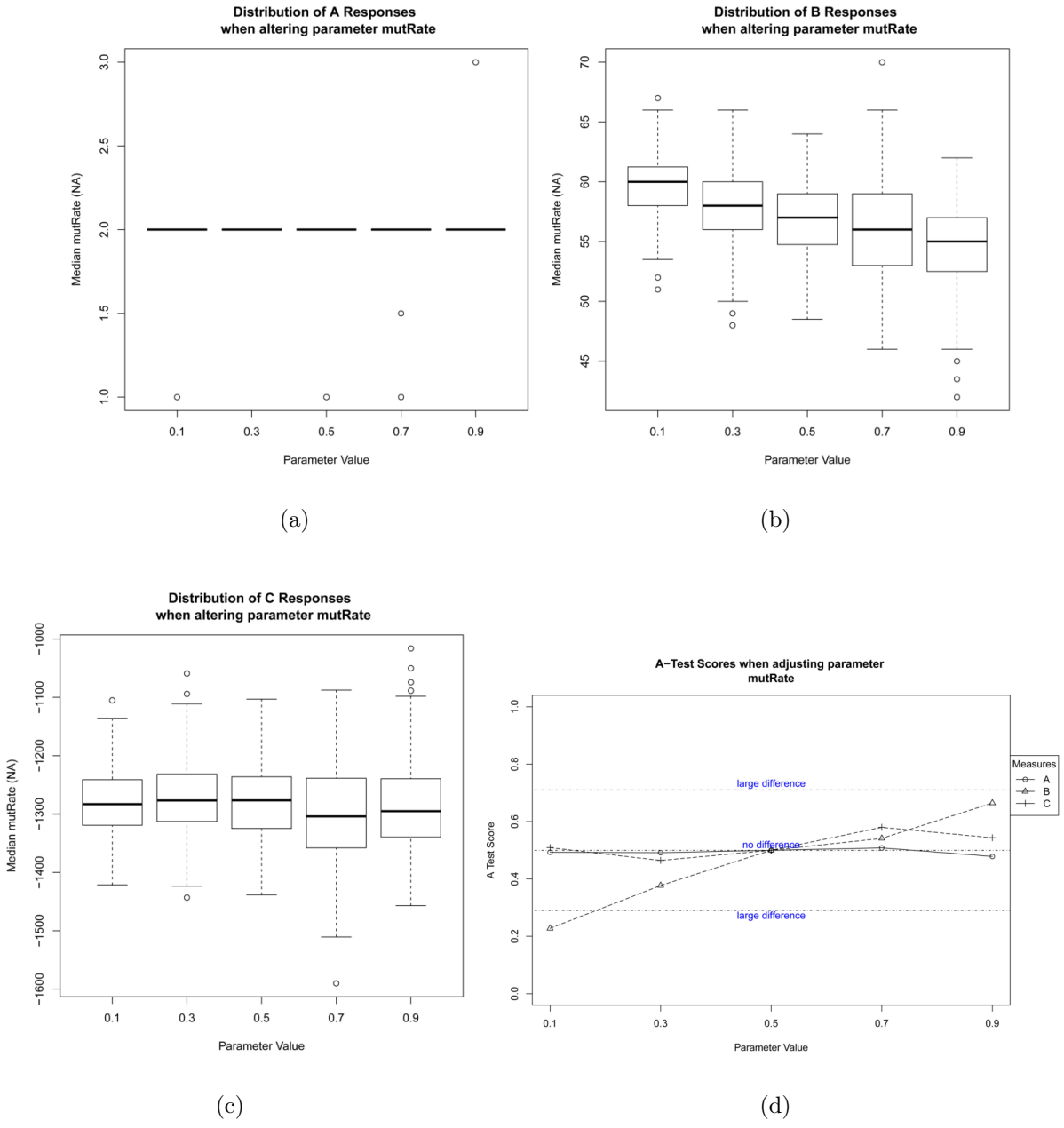


Figura 7.10: Técnica 2 de *Spartan* aplicada a MAIS para o parâmetro taxa de hipermutação (mutRate), indicando o impacto da variação do parâmetro sobre: (a) alelos faltantes; (b) número de *accessions* selecionados; (c) heterozigose. (d) Efeito da magnitude do parâmetro sobre os diferentes objetivos. A – alelos faltantes; B – número de *accessions* selecionados; e C – heterozigose.

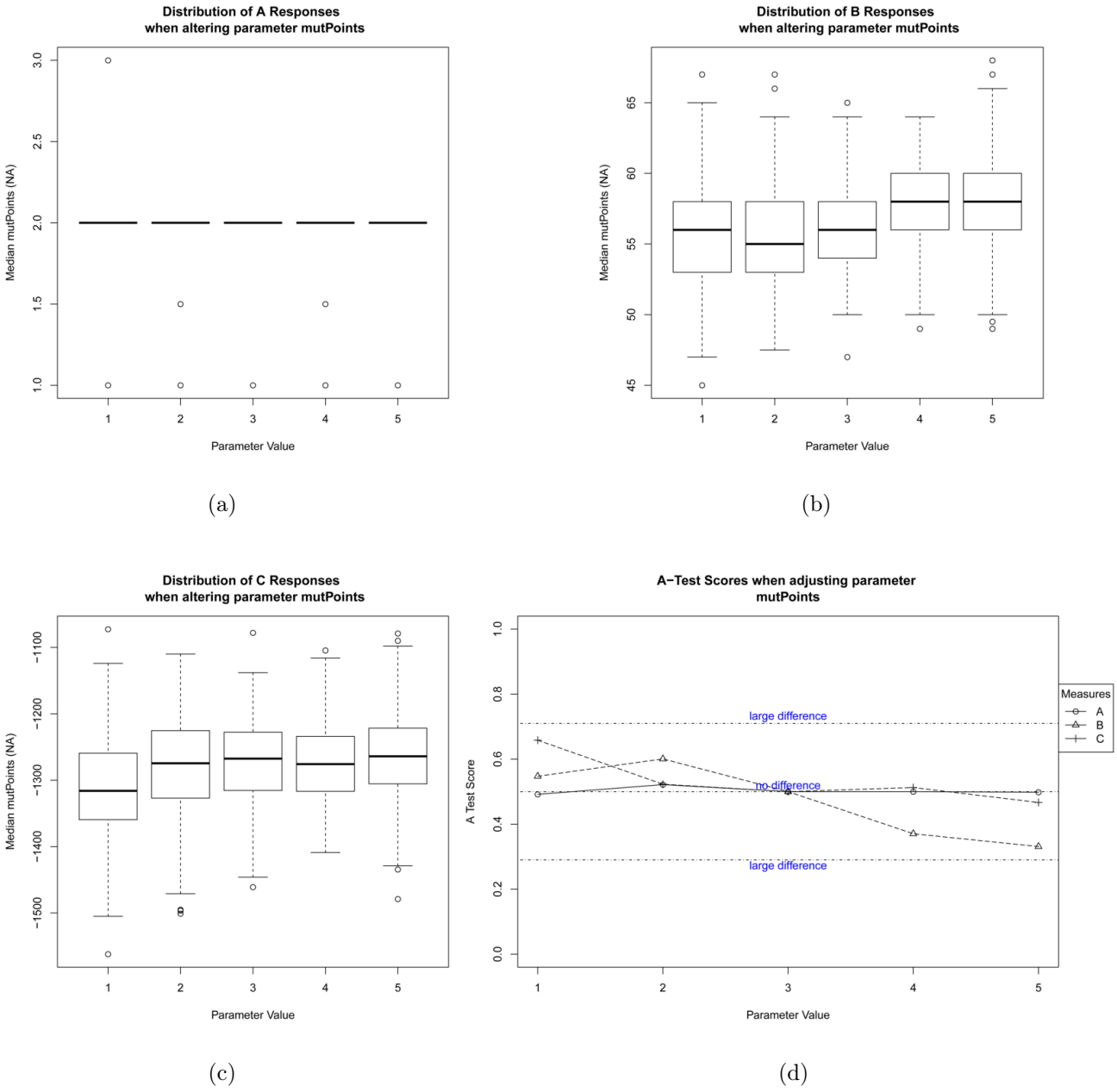


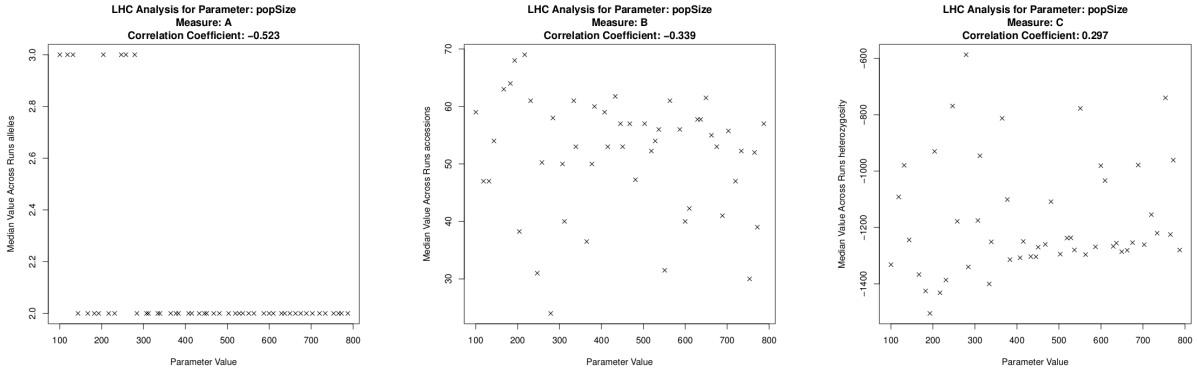
Figura 7.11: Técnica 2 de *Spartan* aplicada a MAIS para o parâmetro número de mutações(hipermutação) (mutPoints), , indicando o impacto da variação do parâmetro sobre: (a) alelos faltantes; (b) número de *accessions* selecionados; (c) heterozigose. (d) Efeito da magnitude do parâmetro sobre os diferentes objetivos. A – alelos faltantes; B – número de *accessions* selecionados; e C – heterozigose.

Tabela 7.10: Conjunto de valores de parâmetros definidos pela amostragem por LHS na Técnica 3 de *Spartan* aplicada a MAIS. O valores para popSize, popMem e mutPoints foram arredondados para o valor inteiro correspondente.

Amostra (FolderIndex)	popSize	popMem	uniformMut	cloneNum	mutRate	mutPoints
1	481,10	266,29	0,46	3,58	0,33	2,04
2	536,87	778,49	0,64	2,81	0,64	3,63
3	192,86	363,81	0,22	2,95	0,38	1,52
4	445,34	472,91	0,72	5,41	0,69	3,50
5	563,39	630,02	0,89	5,50	0,26	3,89
6	629,26	654,93	0,40	2,66	0,44	4,42
7	182,55	481,38	0,49	5,67	0,77	4,70
8	143,44	611,80	0,60	4,85	0,53	1,02
9	467,52	689,71	0,80	1,91	0,21	2,23
10	688,68	230,19	0,66	3,48	0,47	1,60
11	753,52	138,26	0,87	3,17	0,93	1,31
12	311,74	215,48	0,70	4,43	0,57	3,42
13	450,57	412,57	0,75	2,78	0,81	3,83
14	432,96	337,51	0,25	3,21	0,24	4,52
15	599,68	277,70	0,59	1,75	0,91	4,12
16	787,44	404,61	0,91	3,01	0,45	3,20
17	377,20	170,02	0,99	5,71	0,27	3,07
18	719,98	436,47	0,48	3,32	0,85	1,22
19	257,99	311,57	0,28	2,14	0,95	3,66
20	550,96	166,48	0,45	1,29	0,75	2,34
21	246,81	152,13	0,85	5,90	0,90	2,13
22	231,02	747,22	0,84	5,25	0,55	1,73
23	100,11	588,77	0,62	4,35	0,61	1,83
24	702,92	731,32	0,25	2,36	0,35	1,70
25	118,48	303,93	0,32	5,38	0,86	3,10
26	166,88	283,87	0,95	2,49	0,68	4,48
27	278,96	108,60	0,37	2,29	0,40	1,12
28	284,26	768,28	0,22	1,34	0,88	4,27
29	649,42	548,20	0,92	3,97	0,23	4,89
30	519,14	431,26	0,31	1,54	0,70	2,68
31	216,84	729,53	0,76	4,16	0,36	4,99
32	338,97	452,42	0,67	2,52	0,65	1,47
33	609,69	244,10	0,82	3,80	0,74	2,76
34	733,70	796,16	0,74	1,16	0,97	2,85
35	415,06	526,54	0,35	4,91	0,96	4,65
36	307,05	371,30	0,58	4,56	0,82	2,44
37	528,29	347,80	0,69	4,06	0,49	3,25
38	587,02	573,69	0,53	2,10	0,50	2,69
39	333,93	540,68	0,51	4,72	0,52	2,03
40	772,14	201,03	0,96	5,09	0,59	1,92
41	204,02	193,50	0,98	4,24	0,79	2,48
42	407,60	662,63	0,56	5,84	0,63	4,83
43	503,09	713,62	0,33	5,18	0,55	2,98
44	636,75	382,71	0,43	4,65	0,29	3,98
45	675,16	502,37	0,54	1,66	0,80	4,33
46	383,42	641,03	0,86	3,77	0,31	3,35
47	765,35	518,08	0,38	1,02	0,33	1,40
48	364,80	120,49	0,27	3,63	0,41	4,14
49	131,56	683,93	0,78	1,44	1,00	2,55
50	662,52	590,60	0,41	1,86	0,73	3,75

Os resultados corroboraram os achados da Seção 7.4.1. O tamanho da população secundária (memória) tem grande efeito na resposta do algoritmo. O tamanho da população, o número de mutações (hipermutação) e a taxa de hipermutação apresentaram efeito moderado. Já a mutação uniforme e o número de clones mostraram efeito pequeno.

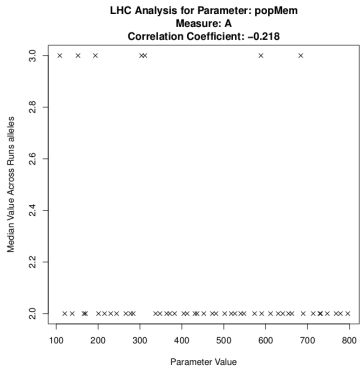
Os valores das métricas H, E e S calculadas para os conjuntos de valores de parâmetros definidos pela amostragem por LHS constam, respectivamente, das Tabelas 7.11, 7.12 e 7.13. A Figura 7.13 mostra os *boxplots* para as referidas métricas. Pela análise dos dados, é possível identificar conjuntos de valores mais adequados para execução do algoritmo. Por exemplo, as amostras de LHS #2, #4, #24, #38, #43 e #50 aparecem entre os 10 melhores resultados para as três métricas consideradas e podem indicar bons valores a serem utilizados pelo algoritmo na instância tratada. Na Tabela 7.14 é apresentado o *ranking* das amostras mais "bem posicionadas" levando em consideração as três métricas (hipervolume, extensão e espaçamento).



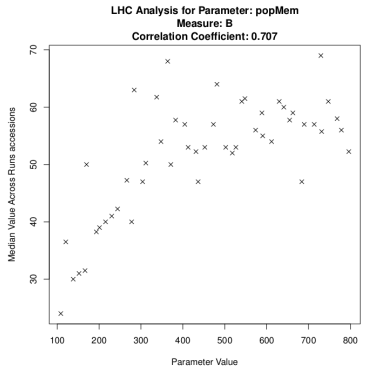
(a)

(b)

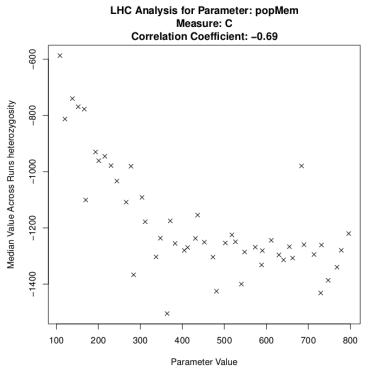
(c)



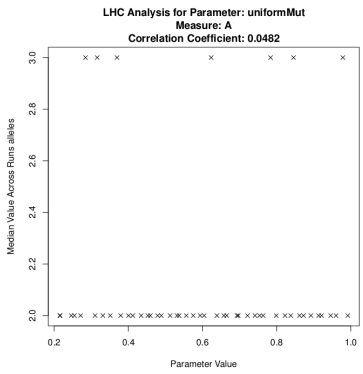
(d)



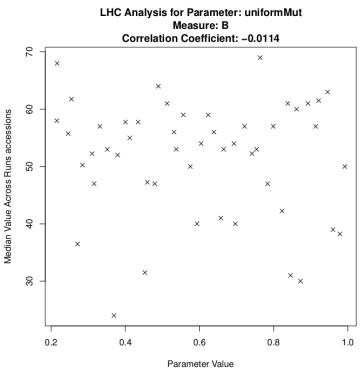
(e)



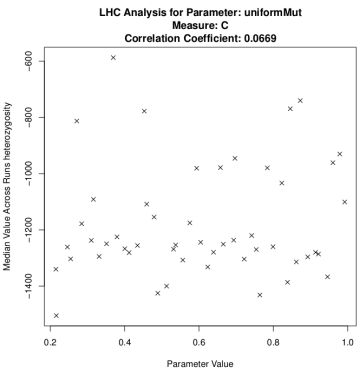
(f)



(g)

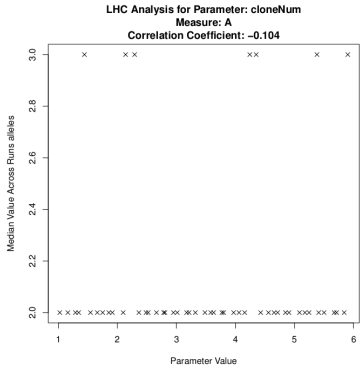


(h)

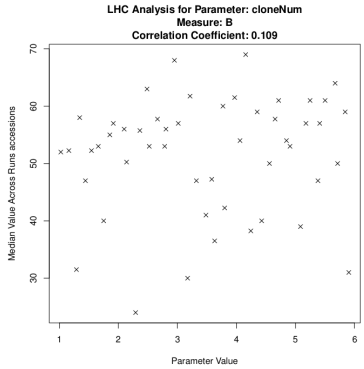


(i)

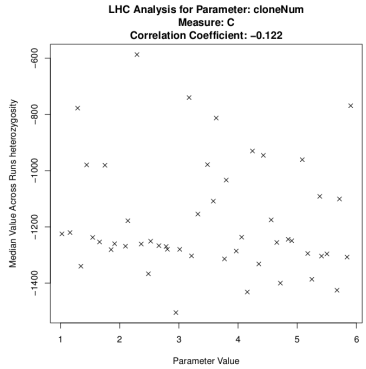
Figura 7.12: A Técnica 3 de *Spartan* aplicada a MAIS. A – alelos faltantes; B – número de *accessions* selecionados; e C – heterozigose.



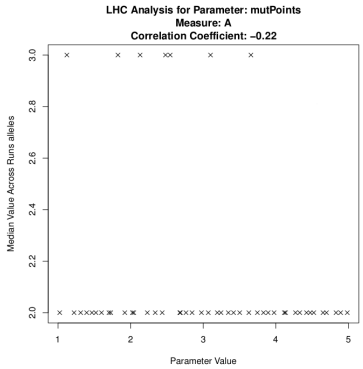
(j)



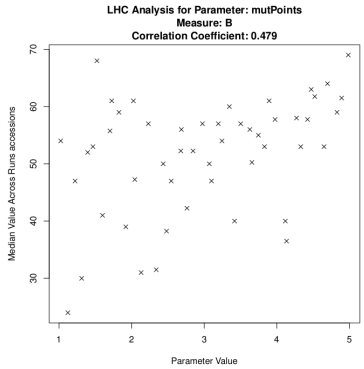
(k)



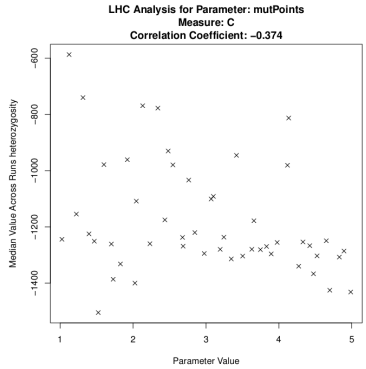
(l)



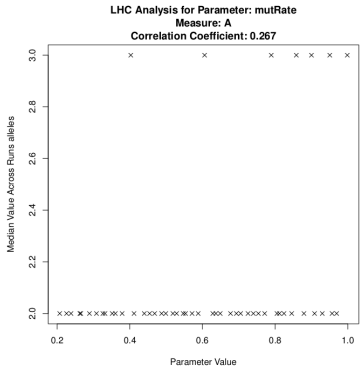
(m)



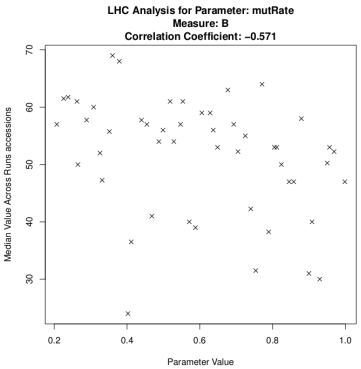
(n)



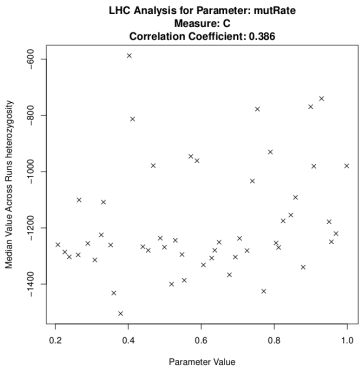
(o)



(p)



(q)



(r)

Figura 7.12: Técnica 3 de *Spartan* aplicada a MAIS. (cont.)

Tabela 7.11: Hipervolume calculado para as amostras de parâmetros LHS. As amostras estão dispostas em ordem decrescente das médias dos valores de H (quanto maior o valor de H, melhor).

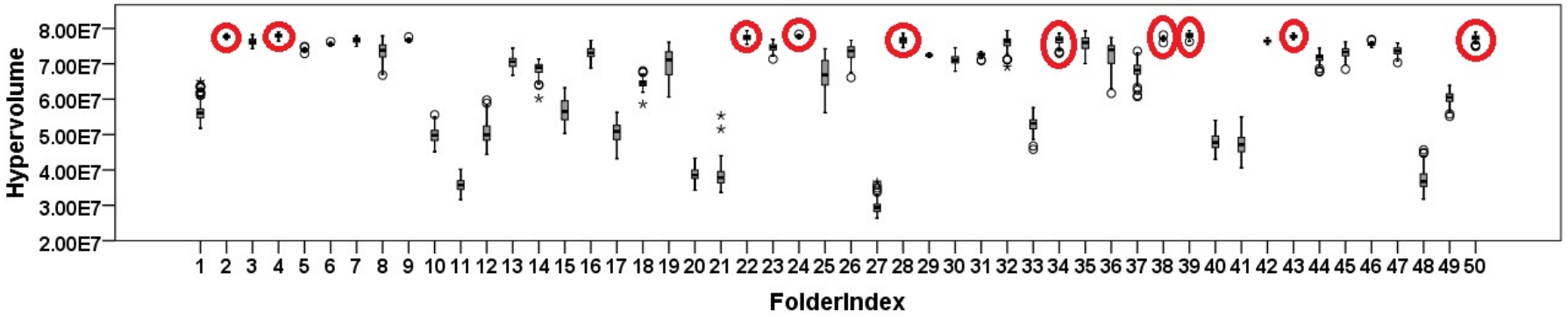
Amostra (FolderIndex)	Média	Desvio Padrão
39	78.130.790	645.287,334
4	77.913.050	570.285,599
24	77.731.520	250.477,583
43	77.702.820	383.845,018
2	77.661.170	312.005,441
22	77.425.010	833.676,735
50	77.309.490	834.810,684
38	77.170.130	339.633,244
34	76.690.870	1.226.934,067
28	76.675.000	1.009.084,494
7	76.640.110	652.425,125
9	76.597.210	29.790,651
42	76.441.810	423.659,519
3	76.222.730	851.994,983
32	75.994.670	1.869.800,638
35	75.707.000	2.035.877,549
46	75.547.950	394.467,865
6	75.508.110	248.345,877
23	74.597.830	1.085.195,500
5	73.887.760	329.836,265
47	73.607.590	1.213.288,112
8	73.388.900	2.572.360,911
45	73.168.450	1.449.548,968
16	73.103.160	1.520.263,432
26	73.046.650	2.311.507,193
36	72.675.410	3.447.596,420
31	72.630.070	514.909,647
29	72.412.620	265.768,365
44	71.668.090	1.334.144,663
30	71.104.250	1.274.333,427
13	70.500.160	1.484.407,767
19	70.183.380	3.843.233,468
14	68.504.900	1.892.713,900
37	68.109.340	2.189.928,439
25	66.880.450	4.243.827,907
18	64.508.260	1.311.877,509
49	60.391.160	1.806.062,326
15	56.870.490	3.349.564,466
1	56.638.420	2.869.171,626
33	52.863.610	2.105.870,510
17	50.638.320	2.804.714,411
12	50.511.820	3.057.901,490
10	49.916.780	2.113.088,541
40	47.903.700	2.426.571,183
41	47.249.140	2.856.913,200
20	38.656.400	1.869.448,668
21	38.337.280	3.270.549,122
48	37.249.540	2.784.340,121
11	35.760.000	1.708.669,983
27	29.667.950	1.892.950,790

Tabela 7.12: Extensão calculada para as amostras de parâmetros LHS. As amostras estão dispostas em ordem decrescente das médias dos valores de E (quanto maior o valor de E, melhor).

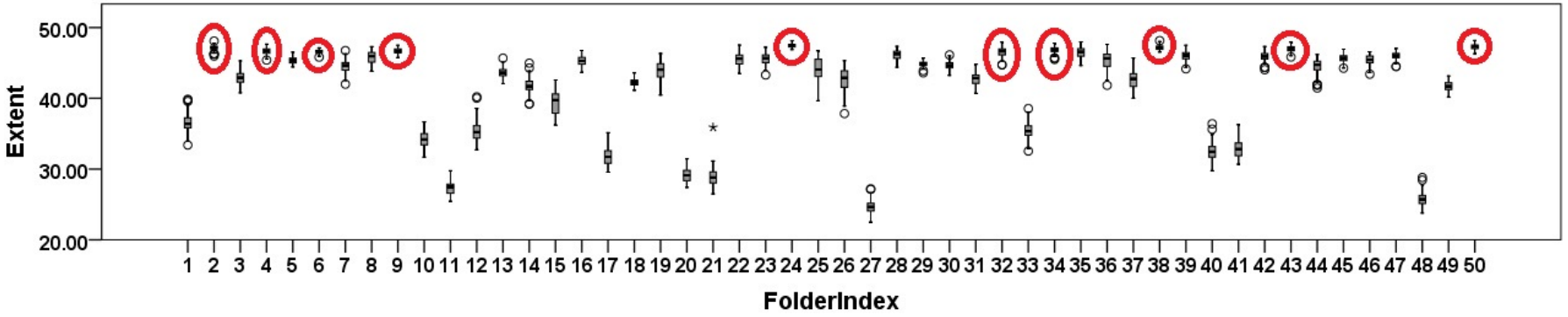
Amostra (FolderIndex)	Média	Desvio Padrão
24	47,47285	0,295958
50	47,27840	0,380059
38	47,16751	0,297713
2	47,13692	0,367862
43	46,95279	0,399853
34	46,84817	0,411892
9	46,67655	0,367986
4	46,67046	0,456712
6	46,59680	0,277993
32	46,55552	0,614877
35	46,48657	0,728439
28	46,14092	0,714146
39	46,03318	0,690682
47	45,99201	0,485938
42	45,91152	0,583334
8	45,72516	0,920403
45	45,66601	0,563734
23	45,56709	0,780015
22	45,49127	0,867643
46	45,43690	0,670469
5	45,37785	0,448524
16	45,29552	0,615009
36	45,27131	1,270060
29	44,87673	0,388042
30	44,66486	0,559842
7	44,60575	0,835038
44	44,50444	1,035821
25	44,11470	1,586661
19	43,90065	1,293173
13	43,66288	0,740271
3	42,89456	0,861119
31	42,71489	0,794153
26	42,70190	1,568179
37	42,62206	1,242088
18	42,22894	0,513116
14	41,80647	1,075232
49	41,71952	0,644529
15	39,39578	1,615305
1	36,51928	1,228877
33	35,49804	1,180000
12	35,35341	1,377754
10	34,21354	0,992963
41	32,87671	1,218501
40	32,49831	1,175560
17	31,77133	1,163773
20	29,16466	0,938537
21	28,90632	1,232344
11	27,32040	0,914049
48	25,74513	0,915264
27	24,71433	0,921693

Tabela 7.13: Espaçamento calculado para as amostras de parâmetros LHS. As amostras estão dispostas em ordem crescente das médias dos valores de S (quanto mais próximo de zero, melhor).

Amostra (FolderIndex)	Média	Desvio Padrão
47	1,987586	0,3695717
24	2,001169	0,2918622
50	2,015216	0,3142218
39	2,062480	0,3134922
30	2,093892	0,4704429
45	2,100912	0,3349659
2	2,131976	0,2496668
38	2,148816	0,3060283
4	2,156669	0,3396246
43	2,162743	0,3118612
28	2,179884	0,3553203
32	2,189248	0,4059809
6	2,218053	0,3655956
35	2,226985	0,4375231
34	2,292336	0,3463494
42	2,298063	0,3433768
22	2,323460	0,5424241
18	2,361784	0,6742941
16	2,376552	0,5881257
9	2,422720	0,4440032
20	2,532657	0,9610591
46	2,542385	0,4579959
1	2,563923	1,0489350
13	2,627846	0,6807156
7	2,695044	0,5620670
23	2,711502	0,4045869
44	2,725675	0,9124822
37	2,754952	0,9630909
36	2,805731	0,5443303
14	2,840180	1,1270379
8	2,901074	0,5599814
5	2,919191	0,8196957
33	2,966206	1,2854156
3	2,975729	0,7945033
15	3,003213	0,9326356
27	3,096703	1,0388489
11	3,131956	1,1661412
29	3,133650	0,8799347
19	3,212225	0,7288167
10	3,214372	1,3586570
40	3,231730	1,3969377
31	3,278820	0,9590321
48	3,283235	1,2174392
26	3,337792	0,7896966
17	3,493749	1,1263485
41	3,525759	1,0235570
25	3,526894	0,8665469
12	3,540187	1,0626901
21	3,641491	1,1379357
49	4,250288	0,6250316

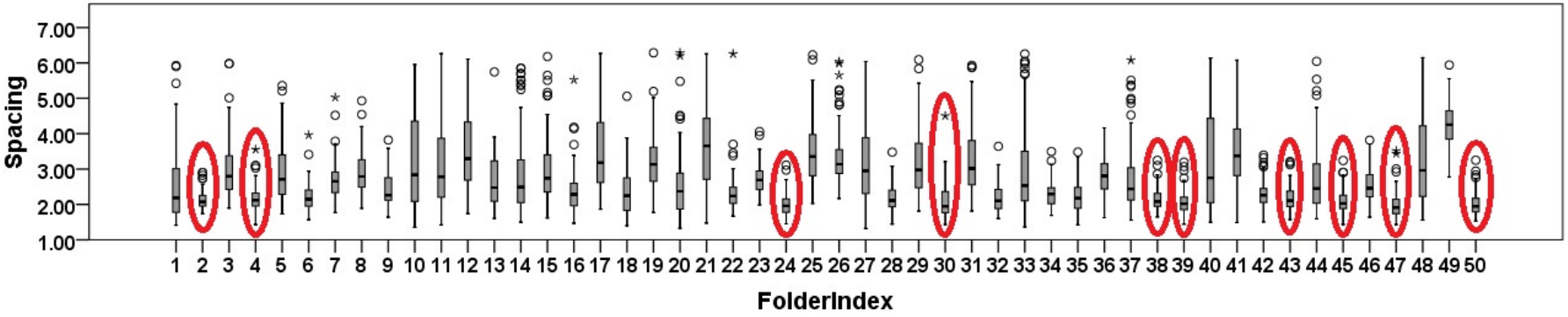


(a)



(b)

Figura 7.13: *Boxplot* de métricas de desempenho calculadas para as 50 amostras LHS de MAIS: (a) H; (b) E; e, (c) S com *z-score* (excluídos *outliers* considerada uma confiança de 99%). Os 10 melhores valores para cada métrica estão destacados por elipses em vermelho. Destaque-se que as amostrada de LHS #2, #4, #24, #38, #43 e #50 aparecem entre os 10 melhores resultados para as três métricas consideradas.



(d)

Figura 7.13: *Boxplot* de métricas de desempenho calculadas para as 50 amostras LHS de MAIS. (cont.)

Tabela 7.14: *Ranking* das amostras de parâmetros LHS "melhor posicionadas" levando em consideração as três métricas (hipervolume, extensão e espaçamento).

Posição no Ranking	Amostra (FolderIndex)
1	39
2	31
3	9
4	18
5	42
6	32
7	30
8	24
9	35
10	36
11	38
12	20
13	7
14	19
15	25
16	26
17	47
18	12
19	41
20	11
21	22
22	48
23	21
24	33
25	5
26	45
27	40
28	34
29	2
30	8
31	49
32	46
33	23
34	4
35	14
36	16
37	13
38	28
39	29
40	43
41	6
42	10
43	50
44	44
45	15
46	37
47	1
48	3
49	17
50	27

Capítulo 8

Trabalhos Publicados

Esta tese é composta de sete *papers* escritos em co-autoria nos quais a doutoranda é a primeira autora. Os *papers* são resultado da pesquisa desenvolvida ao longo do Doutorado, cinco foram aceitos em conferências e dois em periódicos. Pretende-se ainda submeter a publicação os resultados inéditos apresentados no Capítulo 7.

A presente tese contribui para o estabelecimento e a consolidação da Ecoinformática, área emergente de manifestada relevância, sobretudo no contexto atual em que iniciativas de Ecologia e Sustentabilidade têm se mostrado tão significativas.

Quando a autora iniciou seu Doutorado, a área específica de Ecoinformática era incipiente, sobretudo no Brasil. Quase não havia produção e mesmo chamadas em conferências de Ciência da Computação para trabalhos na área, tampouco havia periódicos específicos, motivo pelo qual *papers* que integram esta tese foram submetidos a revistas em **estratos superiores do Qualis Capes com enfoque em Biodiversidade e Interdisciplinaridade**. Em que pese a Ecoinformática ainda não ter atingido o patamar que áreas multidisciplinares (e.g., Bioinformática) já alcançaram, ao longo do desenvolvimento desta tese, as primeiras chamadas específicas para Ecologia em conferências de Ciência da Computação começaram a ser feitas. É o caso, no Brasil, do *Workshop de Computação Aplicada à Gestão do Meio Ambiente e Recursos (WCAMA)*, evento satélite do *Congresso da Sociedade Brasileira de Computação (CSBC)*, com primeira edição em 2009 e VI edição em 2015[1].

A Seção 8.1 apresenta a relação dos trabalhos publicados ou aceitos para publicação. Os *papers* estão listados na ordem que preserva o progresso lógico da pesquisa e não necessariamente na ordem de publicação. As Seções 8.2 a 8.8 trazem a íntegra dos *papers*.

Os *Papers 4, 5 e 6* foram escritos em colaboração com o Prof. Dr. Jon Timmis, da *University of York*, resultado do período de Doutorado Sanduíche na referida Universidade, por meio do Programa Ciência Sem Fronteiras do CNPq/MCTI.

[1] https://wcama.wordpress.com/historia-do-wcama-history-of-wcama/

Cumpre notar que os problemas apresentados na Seção 1.2 foram tratados nos seguintes *papers*:

Problema 1: *Papers* 1 e 2 (dados *D. alata*, baru), *Paper* 7 (dados *E. dysenterica*, cagaita);

Problema 2: *Papers* 3 e 6 (dados *D. alata*).

Problema 3: *Papers* 4 e 5 (dados de 96 espécies de plantas do Cerrado).

8.1 Sumário dos Trabalhos Publicados

8.1.1 Experimentos exploratórios: MOO para SCP

Paper 1

Short paper aceito para publicação e apresentação oral na 8^{th} *International Conference on Ecological Informatics - ISEI2012*, realizada de 3 a 7 de dezembro de 2012, em Brasília-DF. As edição da ISEI[2] são, até o momento, as únicas conferências dedicadas exclusivamente ao tema Ecoinformática de que se tem notícia.

Paper 2

Estende o trabalho apresentado no *Paper 1*. O *paper* foi aceito para publicação no *Journal Genetics and Molecular Research (GMR)*; ISSN 1676-5680. De acordo com o Qualis Capes[3], este periórico é classificado no **estrato A2** da área de avaliação **Interdisciplinar** e **estrato B2** da área de avaliação **Ciência da Computação**.

Nos *Papers 1 e 2*, utilizando dados de *Dipteryx alata* (baru), buscou-se o menor conjunto de populações locais que deveriam ser conservadas para representar a diversidade genética desta espécie do Cerrado brasileiro, utilizando informação relativa à frequência alélica, heterozigose e equilíbrio de Hardy-Weinberg (HWE). Foram trabalhadas 3 variações de problema. Inicialmente foi reproduzido o experimento de Diniz-Filho et al. (2012), mas utilizando a abordagem multiobjetivo, com isso os achados daquele estudo foram ratificados, pois o menor conjunto capaz de representar todos os alelos em estudo continha 7 populações, mas o método proposto foi capaz de identificar uma maior diversidade de soluções com a quantidade mínima de populações. Nas 2^a e 3^a variações do problema, foram otimizados simultaneamente 4 e 5 objetivos, respectivamente. Foram encontrados resultados semelhantes aos encontrados na 1^a variação do problema, mas a utilização de

[2]Desde 1998 ela já foi realizada na França, Autrália, Itália, Coréia do Sul, Estados Unidos, México, Bélgica, Brasil e China.

[3]Sistema brasileiro de avaliação de periódicos, mantido pela CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) e usado para estratificação da qualidade da produção intelectual dos programas de pós-graduação. Disponível em <http://qualis.capes.gov.br/webqualis>.

mais objetivos permitiu identificar um portfólio maior de soluções, com resultados variando entre 8 e 22 populações. Ao mesmo tempo, os resultados para 7 populações foram refinados, sendo indicando dentre eles, aquelas combinações com maior diversidade intra-específica e maior possibilidade de persistência ao longo do tempo. Esta foi a primeira vez que a abordagem multiobjetivo foi aplicada ao problema SCP, assim como a primeira vez que informação referente a frequência alélica, heterozigose e HWE foram utilizadas para auxiliar na definição de estratégias de conservação de biodiversidade.

Paper 3

Complementa os *Papers 1* e *2* apresentando uma aplicação com potencial de inovação tecnológica ao identificar com exatidão amostras que complementam coleções de germoplasma pré-existentes, bem como ao estabelecer *core collections* capazes de preservar a diversidade biológica da espécie em estudo.

Publicado no *Journal Tree Genetics & Genomes (TGG)*; ISSN 1614-2942 (versão impressa)/ISSN 1614-2950 (versão eletrônica). Este periórico tem registrado historicamente impacto crescente na comunidade científica, tendo atingido em 2013 **Fator de Impacto 2.435**. De acordo com o Qualis Capes é, classificado no **estrato A1** da área de avaliação **Biodiversidade** e no **estrato A2** da área de avaliação **Interdisciplinar**.

Novamente utilizando dados de *D. alata*, buscou-se encontrar *core collections* para coleções de germoplasma a fim de representar a diversidade da coleção ao mesmo tempo em que se minimizam custos de conservação, mantendo-se a máxima variabilidade genética. Neste caso, havia dois problemas a serem resolvidos: (i) selecionar um conjunto de amostras geneticamente complementar a uma coleção de germoplasma já existente; (ii) definir uma *core collection* para uma coleção de germoplasma. Neste caso, utilizou-se informação alélica de duas fontes, uma *ex situ* correspondente à coleção de gemoplasma da Escola de Agronomia da Universidade Federal de Goiás (UFG-AS) e outra *in situ* correspondendo a amostras coletadas de 642 árvores em seu *habitat* natural. Como resultado, foi possível identificar dentre as 642 amostras, as árvores exatas que deveriam ser amostradas para complementar a diversidade genética já existente na coleção de germoplasma da UFG-AS. Além disso, com o protocolo proposto, foi possível lidar com grande volume de amostras e definir uma *core collection* para a UFG-AS considerando o material *ex situ* e *in situ* juntos. O método proposto pode ser utilizado para ajudar a construir coleções com máxima riqueza alélica assim como ser estendido para conservação *in situ*. Esta foi a primeira vez que uma abordagem multiobjetivo foi aplicada aos problemas de complementar uma coleção de germoplasma, bem como definir uma *core collection* para a mesma.

8.1.2 MOO, SCP e mudança climática

Paper 4

Short Paper aceito para publicação e apresentação oral na *ACM Genetic and Evolutionary Computation Conference* (GECCO'2014), realizada de 12 a 16 de julho de 2014, em Vancouver, Canadá. Esta conferência é muito competitiva, com taxa de aceitação de cerca de 33% sendo a mais importante na área de Computação Evolutiva/Bioinspirada, patrocinada pelo *SIGEVO ACM Special Interest Group on Genetic and Evolutionary Computation*. É classificada, segundo o Qualis-CC Conferências[4], no **estrato A1**, tendo um **índice-H 66**.

Paper 5

Estende o trabalho apresentado no *Paper 4*. Este *full paper* foi aceito para publicação e apresentação oral na *International Conference on Evolutionary Multi-Criterion Optimization* (EMO'2015), realizada de 29 de março a 1 de abril de 2015, em Guimarães, Portugal. Esta conferência é a mais importante na área de Otimização em Ciência da Computação. Segundo o Qualis-CC Conferências, está classificada no estrato **A2**, com **índice-H 40**.

O trabalho foi publicado no vol. 9019, *Evolutionary Multi-Criterion Optimization*, ISBN 978-3-319-15891-4, do *Lecture Notes in Computer Science*, classificado no **estrato A1** na área de **Biodiversidade**.

Nos *Papers 4 e 5*, com o objetivo de definir prioridades de conservação, aplicou-se a abordagem MOO a dados de ocorrência de 96 espécies de plantas existente atualmente no Cerrado e a dados correspondentes à estimativa da ocorrência de tais espécies projetada para 2080 com base em simulações climáticas para o futuro. Como informação adicional foram utilizados dados relativos ao interesse econômico da região, compreendendo a ocupação humana, o remanescente de vegetação e a evapotranspiração anual. Como resultado, o método proposto foi capaz de identificar locais com: (i) alta prioridade para conservação; (ii) significativo risco de investimento; e, (iii) que podem tornar-se atrativos no futuro. Esta foi a primeira vez que MOO associado a previsão climática foi aplicado ao problema SCP em uma análise de priorização dinâmica para conservação da biodiversidade.

[4]Classificação de conferências da área de Ciência da Computação, estabelecida pela CAPES, com última atualização dada por meio do Comunicado n. 004/2012, de 31 de agosto de 2012. Disponível em <http://www.capes.gov.br/component/content/article?id=4656:ciencia-da-computacao>.

8.1.3 MOO, SCP... e MAIS

Paper 6

Full paper aceito para publicação e apresentação oral na *ACM Genetic and Evolutionary Computation Conference* (GECCO'2015), a ser realizada de 11 a 15 de julho de 2015, em Madri, Espanha. Conforme citado para o *Paper 4*, esta conferência é classificada, segundo o Qualis-CC Conferências, no **estrato A1**, tendo um **índice-H 66**.

Neste *paper*, para o problema de se encontrar uma *core collection* para um banco de germoplasma, foi proposta a utilização do algoritmo *Multi-Objective Artificial Immune Algorithm System (MAIS)*, utilizando princípios de planejamento sistemático de conservação e incorporando informação de heterozigose. Procedendo desta forma, a otimização levou em consideração padrões de diversidade genotípica. Como estudo de caso foram utilizados dados de marcadores moleculares de *D. alata*. Utilizando o método proposto, foi possível identificar dentre vários *accesssions* disponíveis, as entradas exatas que deveriam ser selecionadas para compor a *core collection* a fim de se preservar a diversidade da espécie. Quando comparado a NSGA-II, um algortimo estado-da-arte em otimização multiobjetivo, MAIS apresentou melhores resultados dadas as métricas utilizadas. A abordagem proposta pode ser utilizada no auxílio à construção de *cores* como máxima riqueza genética bem como ser estendido para casos de conservação *in situ*. Até onde se sabe, esta foi a primeira vez que um algoritmo AIS foi aplicado ao problema de se encontrar um *core* para um banco de germoplasma usando informação de heterozigose.

Paper 7

Full paper aceito para publicação e apresentação oral no *Workshop on Evolutionary Multi-Objective Optimization* do *IEEE 2015 International Congress on Evolutionary Computation* (CEC2015), realizado de 25 a 28 de maio de 2015, em Sendai, Japão.

O CEC é uma conferência do *Institute of Electrical and Electronics Engineers–IEEE*, classificada, segundo o Qualis-CC Conferências, no **estrato A2**, tendo um **índice-H 34**. Trabalhos apresentados no âmbito do CEC se beneficiam da grande visibilidade proporcionada pela expressiva inserção do evento na comunidade de pesquisa em Computação Evolutiva.

O *Papers 7* tratou de políticas para o desenvolvimento de sustentabilidade relacionada à conservação de biodiversidade. A partir da informação de diversidade alélica em 23 populações de *Eugenia dysenterica* (cagaita), trabalhou-se o problema de compor uma *core collection* a fim de representar a diversidade genética da espécie, tomando por base informações de marcadores genéticos. Novamente propôs-se a utilização do algoritmo

Multi-Objective Artificial Immune System (MAIS), incorporando informação alélica e de *habitat*. Foi possível identificar os melhores conjuntos de populações que deveriam ser protegidas com o propósito de preservar a diversidade da espécie. Esta foi a primeira vez que um algoritmo inspirado em Sistemas Imunológicos Artificiais foi aplicado ao problema SCP usando informação genética e considerando variáveis ambientais, a saber, estabilidade da região e perda de *habitat*.

8.2 *Paper 1*

Schlottfeldt, S.; Walter, M.E.M.T.; Diniz-Filho, J.A.F.; Telles, M.P.C. Multiobjective Optimization in Systematic Conservation Planning to Represent Genetic Variability within Species. In Proceedings of the 8th International Conference on Ecological Informatics, ISEI'2012, Brasília, Brazil, 2012.

MULTIOBJECTIVE OPTIMIZATION IN SYSTEMATIC CONSERVATION PLANNING TO REPRESENT GENETIC VARIABILITY WITHIN SPECIES

Santos, S.S.1 (shanass@unb.br)
Walter, M.E.M.T.2 (mia@cic.unb.br)
Diniz-Filho, J.A.F.3
Loyola, R.D.3 (diniz, loyola@icb.ufg.br)
Telles, M.P.C.4 (tellesmpc@gmail.com)

1 Departamento de Ciência da Computação, Universidade de Brasília. 70910-900, Brasília, DF, Brasil.
2 Departamento de Ciência da Computação, Instituto de Ciências Exatas, Universidade de Brasília, 70910-900, Brasília, DF, Brazil.
3 Departamento de Ecologia, ICB, Universidade Federal de Goiás. CxP 131, 74001-970, Goiânia, GO, Brazil.
4 Departamento de Biologia Geral, ICB, Universidade Federal de Goiás. CxP 131, 74001-970 Goiânia, GO, Brazil

03-07 Dec 2012
Brasilia, Brazil

Session: Land resources monitoring and conservation (Chair: Miguel Marini, University of Brasilia, Brazil)
4th December 2012, room 2, 17h40-18h00 (ISEI-10)

Abstract

1. Introduction:

The current biodiversity crisis is conducting scientists to develop systematic strategies to effectively achieve conservation goals. The overall underlying principle of this strategy lies in the framework usually called “Systematic Conservation Planning” (SCP).

In the core of SCP is the set covering problem, a classical question in computer science and complexity theory that was shown to be NP-complete (Cormen et al. 2001). In SCP, there are mainly two ways to state the problem: 1) to select a set of sites (among several available ones) so as to minimize the overall cost of conservation action but maximizing the natural features representation (i.e. the minimum set coverage problem); or, 2) given a limited budget, to select the combination of sites that maximizes the representation of natural features (i.e. the maximal set coverage problem). Clearly, in both ways, there are two conflicting objectives, which makes the problem a candidate for multiobjective optimization.

Several real world problems involve simultaneous optimization of multiple conflicting objectives. In these cases, there is no single optimal solution, but rather a set of solutions that should be considered equivalent in the absence of information about the relevance of each objective relative to the others (Fonseca and Fleming, 1995; Valenzuela-Rendón et al., 1998). Such solutions (non-dominated solutions) are optimal in the sense that no solution in the search space is superior to them when all the objectives are considered. They are called Pareto Optimal Solutions (Zitzler and Thiele, 1998).

Although SCP has been usually applied at species level (or hierarchically higher) (Brooks et al. 2004; Grelle et al. 2010; Pressey 2004), it is possible to solve a series of problems at much lower hierarchical levels, such as to use alleles from molecular analysis at population level as basic units for analysis, in the context of the new field of conservation genetics (Diniz-Filho et al., 2012, Diniz-Filho and Telles 2006). However, these previous attempts used only the presence or absence of alleles, which is not as informative as using directly the allele frequencies, which reflect in a much appropriate way the ecological and evolutionary processes driving genetic diversity in local populations.

Here we propose a solution based on Evolutionary Algorithms (EA) (Bäck, 1996) that allows us to cope with an instance of the SCP problem with more than two dimensions. This gives us more flexibility by including additional decision variables, also adding more complexity and increasing the power of decision. In more detail, we used NSGA-II (Deb et al., 2000), a state of art Multiobjective Evolutionary Algorithm (MOEA), in order to find optimal solutions. NSGA-II finds the smallest set of local populations of Dipteryx alata that should be conserved to represent the known genetic diversity of this Brazilian Cerrado species, based on allele frequency information associated to heterozygosity and Hardy-Weinberg equilibrium. As long as we know, this is the first attempt to apply multiobjective algorithms to a SCP problem with more than two dimensions and using alleles from molecular analysis at population level as basic units for analysis.

2. Methods

2.1. Data

We used data from Dipteryx alata (a Fabaceae tree species widely distributed and endemics to Brazilian Cerrado) consisting of 55 alleles from 9 microsatellite loci coded for a total of 642 individual trees sampled in 25 local populations throughout species’ geographic range.

2.2. Modeling

2.2.1. Problem

Fixing the smallest population set that represents all the collected genetic diversity (all the 55 alleles) to 7 (our lower bound) (Diniz-Filho et al., 2012), we relaxed the restriction that all the alleles in the populations might be represented, as well as added information about allele frequency.

Our instance of the SCP problem can be stated as follows: we aim to select a minimum population set, maximizing the frequency of alleles among the 55 identified, i.e. we seek to maximize the frequency of the maximum number of alleles and simultaneously minimize the number of selected populations. To achieve this, we also associate a third dimension to be optimized: A) maximization of the heterozygosity of the local population in the solution; B) maximization of the Hardy-Weinberg equilibrium level in the populations in the solutions; C) maximization of populations with maximum levels of combined heterozygosity and Hardy-Weinberg equilibrium. Each dimension is considered an objective function.

3. Results

For problem instance A, described in Section 2.2.1, we obtained as best solution 1 missing allele and 7 populations (5, 15, 17, 19, 20, 21 and 25). Problem instance B has as best solution no missing alleles and 10 populations (1, 4, 5, 6, 7, 15, 17, 19, 24 and 25). Problem instance C

outputs an apparently worst solution represented by 4 missing alleles in 10 populations (1, 2, 4, 5, 6, 7, 15, 17, 20 and 24).

We also reproduced the experiment developed by Diniz-Filho et al. (2012) based on presence-absence of the alleles (and not allele frequencies), having found 2 common solutions among 4 that were previously identified using the simulated annealing method, and produced 4 new solutions for the problem of minimizing the number of populations when all the 7 alleles are present (a 2-dimensional problem).

4. Conclusion and Perspectives

In this work, we used one Evolutionary Algorithms (EA) to solve an instance of the SCP problem with more than two dimensions, so adding more complexity as well as increasing the power of decision. We used NSGA-II (a Multiobjective Evolutionary Algorithm) to find optimal solutions for the problem of identifying the smallest set of local populations of the Brazilian Cerrado Dipteryx alata species that should be conserved to represent genetic diversity based on allele frequency information associated to heterozygosity and Hardy-Weinberg equilibrium. As far as we know, this is the first attempt to apply multiobjective algorithms to a SCP problem with more than two dimensions, and also using alleles from molecular analysis at population level as basic units for analysis.

For 1 missing allele, we found very similar solutions (there is only one different solution), which suggests that heterozygosity and Hardy-Weinberg equilibrium guided the set of solutions in the same direction. This is noteworthy when compared to the results of problem instance C, which produced apparently worst solutions (4 missing alleles in 10 populations). Therefore, the Hardy-Weinberg equilibrium was the only to find a result with any missing allele, which can be happened due to not accurate choice of the weights when combining both criteria.

More experiments will be done to determine in NSGA-II more precise and controlled objective function. It is interesting to propose a more specialized multiobjective evolutionary algorithm to the SCP problems.

Keywords: multiobjective optimization, conservation planning, genetic variability.

References

Bäck, T., 1996. Evolutionary algorithms in theory and practice: evolution strategies, evolutionary programming, genetic algorithms. Oxford University Press, Oxford.

Brooks, T.M., da Fonseca G.A.B. and Rodrigues A.S.L., 2004. Species, data, and conservation planning. Conserv Biol, 18, 1682-1688. doi: 10.1111/j.1523-1739.2004.00457.x

Cormen, T.H., Stein, C., Rivest, R.L. and Leiserson, C.H., 2001. Introduction to Algorithms. 2nd ed. McGraw-Hill Higher Education.

Deb, K., Agarwal, S., Pratap, Am. and T. Meyarivan, 2000. A fast elitist non-dominated sorting genetic algorithm for multiobjective optimization: NSGA-II. In Proceedings of the Parallel Problem Solving from Nature VI Conference, 849–858.

Diniz-Filho, J.A.F., Loyola, R.D., Melo, D.B., Oliveira, G., Collevatti R.G., Soares, T.N., Nabout, J.C., Lima, J.S., Dobrovolski, R., Chaves, L.J. and Naves, R.V.,

2012. Planning for optimal Conservation Geographical Genetic Variability within Species. Conservation Genetics, 13, 1085-1093.

Diniz-Filho, J.A.F. and Telles, M.P.C., 2006. Optimization procedures for establishing reserve networks for biodiversity conservation taking into account population genetic structure. Genet Mol Biol, 29, 207-214.

Fonseca, C.M. and Fleming, P.J., 1995. An overview of evolutionary algorithms in multiobjective optimization. Evolutionary Computation, 3, 1–16.

Grelle, C.E.V., Lorini, M.L. and Pinto, M.P., 2010. Reserve selection based on vegetation in the Brazilian Atlantic Forest. Nat Conservacao, 8, 46-53. doi: 10.4322/natcon.00801007.

Pressey, R.L., 2004. Conservation planning and biodiversity: Assembling the best data for the job. Conserv Biol 18, 1677-1681. doi: 10.1111/j.1523-1739.2004.00434.x

Valenzuela-Rendón, M., 1998. Reinforcement learning in the fuzzy classifier system. Expert Systems with Applications, 14, 237-247.

Zitzler. E. and Thiele, L., 1998. An Evolutionary Approach for Multiobjective Optimization: The Strength Pareto Approach. TIK Report 43, Computer Engineering and Networks Laboratory (TIK), ETH Zurich.

8.3 *Paper 2*

Schlottfeldt, S.; Walter, M.E.M.T.; de Carvalho, A.C.P.L.F; Soares, T.N.; Telles, M.P.C.; Loyola, R.D.; Diniz-Filho, J.A.F. Multiobjective Optimization for Conservation of Genetic Resources Based on Molecular Variability. 2015. Genet. Mol. Res. 14 (2): 6744-6761. ISSN 1676-5680.

A publicação final está disponível em GMR via
http://dx.doi.org/10.4238/2015.June.18.18

GMR Genetics and Molecular Research
Online Journal
ISSN:1676-5680

Multi-objective optimization in systematic conservation planning and the representation of genetic variability among populations

S. Schlottfeldt[1], M.E.M.T. Walter[1], A.C.P. L.F. Carvalho[2], T.N. Soares[3], M.P.C. Telles[3], R.D. Loyola[4] and J.A.F. Diniz-Filho[4]

[1]Departamento de Ciência da Computação,
Instituto de Ciências Exatas, Universidade de Brasília, Brasília, DF, Brasil
[2]Departamento de Ciência da Computação, Universidade de São Paulo,
São Carlos, SP, Brasil
[3]Departamento de Biologia Geral, Instituto de Ciências Biológicas,
Universidade Federal de Goiás, Goiânia, GO, Brasil
[4]Departamento de Ecologia, Instituto de Ciências Biológicas,
Universidade Federal de Goiás, Goiânia, GO, Brasil

Corresponding author: S. Schlottfeldt
E-mail: shanass@unb.br

Genet. Mol. Res. 14 (2): 6744-6761 (2015)
Received August 21, 2014
Accepted February 13, 2015
Published June 18, 2015
DOI http://dx.doi.org/10.4238/2015.June.18.18

ABSTRACT. Biodiversity crises have led scientists to develop strategies for achieving conservation goals. The underlying principle of these strategies lies in systematic conservation planning (SCP), in which there are at least 2 conflicting objectives, making it a good candidate for multi-objective optimization. Although SCP is typically applied at the species level (or hierarchically higher), it can be used at lower hierarchical levels, such as using alleles as basic units for analysis, for conservation genetics. Here, we propose a method of SCP using a multi-objective approach. We used non-dominated sorting genetic algorithm II in order to identify the smallest set of local populations of *Dipteryx alata* (baru) (a Brazilian Cerrado species) for conservation, representing

the known genetic diversity and using allele frequency information associated with heterozygosity and Hardy-Weinberg equilibrium. We worked in 3 variations for the problem. First, we reproduced a previous experiment, but using a multi-objective approach. We found that the smallest set of populations needed to represent all alleles under study was 7, corroborating the results of the previous study, but with more distinct solutions. In the 2nd and 3rd variations, we performed simultaneous optimization of 4 and 5 objectives, respectively. We found similar but refined results for 7 populations, and a larger portfolio considering intra-specific diversity and persistence with populations ranging from 8-22. This is the first study to apply multi-objective algorithms to an SCP problem using alleles at the population level as basic units for analysis.

Key words: Conservation planning; Multi-objective optimization; Metaheuristics; Genetic variability; Biodiversity

INTRODUCTION

The current biodiversity crisis is forcing scientists to develop systematic strategies for effectively achieving conservation goals. The overall underlying principle of this strategy lies in systematic conservation planning (SCP), which involves a series of decisions in order to determine the most cost-effective method of investing in conservation measures (Margules and Pressey, 2000). The overall reasoning is to develop a protocol in which a set of conservation targets and goals are defined and achieved in the most objective and rational manner possible.

Although SCP is typically applied to species or even hierarchically higher levels (Brooks et al., 2004; Pressey, 2004), it is possible to solve a series of problems at much lower hierarchical level. Alleles from molecular analysis at the population level can be used as basic units in the context of conservation genetics (Diniz-Filho and Telles, 2002; Diniz-Filho et al., 2012) based on the idea of intra-specific conservation prioritization. This idea has its roots in the early 1990's in the debate of evolutionary significant units or management units for conservation (Fraser and Bernatchez, 2001). Although the definition of evolutionary significant units or management units provides the concept of intra-specific units (e.g., subspecies or local varieties) that can be used as conservation targets, it is not able to deal with continuous genetic variation at the species level (Diniz-Filho and Telles, 2002) and, more importantly, does not provide a method for identifying intra-specific variation units or components that should be prioritized in the context of SCP.

The complementarity concept is in the foundation of SCP and is mathematically modeled by the set covering problem, a classical problem in algorithm complexity theory that was shown to be NP-complete (Cormen et al., 2001). Understanding that a problem is NP-complete provides an indication of the difficulty of the problem. In general, it is not difficult to verify whether an answer to an NP-complete problem is correct. However, whether the solution is efficient must be determined by testing all possible options until finding one that solves the problem correctly. NP-complete problems arise in several real-world applications and, in practice, knowing that a problem is NP-complete can prevent the spending of time attempting to determine a polynomial-time algorithm to solve it exactly when such an algorithm likely does not exist.

Independently of the hierarchical level, the SCP problem can be stated as follows: to select a set of sites (among several available sites) to minimize the cost of conservation and, at the same time, maximize the natural feature representation, which can be modeled by the minimum set covering problem. This approach presents at least 2 conflicting objectives, as one is usually interested in the conservation of some biological features (e.g., alleles, species, or any other biological unit) subjected to conflicting interests in land use (e.g., human population living in the region, price of land, agricultural potential of the area, or probability of habitat loss). In this context, the conflicting objectives make the problem a perfect candidate for multi-objective optimization. Furthermore, other parameters can be considered by adding socio-economical costs to the areas or by minimizing their spatial aggregation; this adds dimensions to SCP, which is already multi-objective in its origin. Indeed, several real world problems involve simultaneous optimization of multiple conflicting objectives, which should be analyzed as independent dimensions rather than combined into a single weighted function.

These optimization problems with more than one objective are referred to as vector optimization or multi-objective problems (Zitzler et al., 2002; Coello-Coello et al., 2007). In these cases, there is no single optimal solution, but rather a set of solutions that should be considered to be equivalent in the absence of information regarding the relevance of each objective relative to the others (Fonseca and Fleming, 1995). Such solutions, known as non-dominated, are optimal in the sense that none can be declared the best when all objectives are considered (Zitzler et al., 2002). These solutions are called Pareto optimal solutions, and the graph of the solutions form the Pareto front (Coello-Coello et al., 2007).

Multi-objective evolutionary algorithms (MOEA) have been successfully applied for multi-objective problems. The positive aspects of MOEA include efficient solution space exploration, parallelism, ability to escape of local optima, capacity to handle complex problems for which it is not possible (or at least it is difficult) to obtain a detailed description, and they are less susceptible to the shape or continuity of functions (Coello-Coello et al., 2007).

In SCP, the development of algorithms and tools for decision support began in the early 1980s (Sarkar, 2012) and became an important element in conservation biology research. Several approaches have been suggested over the past decades, ranging from a simple scoring system to more complex optimization techniques (Table 1).

In their origin, such algorithms sequentially select complementary sites until all species are represented in typical greedy algorithm behavior. However, greedy algorithms are not guaranteed to identify optimal solutions (Possingham et al., 2000). Nevertheless, this class of algorithms has distinguished importance in the development of algorithms and tools for SCP. In contrast, the exact approach (which ensures the production of optimal solutions) was initially discussed by Cocks and Baird in 1989 (Sarkar, 2012). However, as SCP is an NP-complete problem, even the available software packages computing exact algorithms cannot solve some large data sets (Pressey et al., 1996), which is a common limitation in the SCP context. Because of these characteristics, another approach used for SCP involves metaheuristics, a method for solving an optimization problem using a combination of random choices and historical knowledge of previous results computed by the method, such that the heuristic explores the solution space. However, it is worth noting that this technique does not guarantee optimal solutions. The mainly metaheuristics used for SCP include simulated annealing [SPEXAN (Ball, 2000), SITES (Possingham et al., 2000), and Marxan (Ardron et al., 2010)], as well as the tabu search [ConsNet (Ciarleglio, 2010)].

Table 1. Methods and associated strategies proposed to solve the SCP problem.

Methods	Solution strategies						
	Scoring	Greedy	LP/ILP*	Rules	Metaheuristics		
					Simulated annealing	Tabu search	Evolutionary
Kirkpatrick's Algorithm (Kirkpatrick, 1983)	X	X					
Ackery and Vane-Wright's Algorithm (Ackery and Vane-Wright, 1984)		X					
Margules and Nicholls' Algorithm (Margules and Nicholls, 1987)[1]		X		X			
Margules, Nicholls and Pressey's Algorithms (Margules et al. 1988)[2]		X		X			
Nicholls and Margules' Algorithm (Nicholls and Margules, 1993)[3]		X		X			
Rebelo and Siegfried's Algorithm (Rebelo and Siegfried, 1990)[4]		X		X			
Exact algorithms (software packages as LINDO, CPLEX, OSL) (Underhill, 1994; Arthur et al., 1997)			X				
C-Plan (Pressey et al., 1996)[5]		X		X			
SPEXAN (Ball, 2000)					X		
SITES (Possingham et al., 2000)					X		
Marxan (Possingham et al., 2000; Ardron et al., 2010)[7]					X		
ResNet (Garson et al., 2002)[6]		X					
Target (Barton et al. 2004)		X					
Zonation (Moilanen, 2005)[8]		X					
MultCSync (Moffett et al., 2005)[9]				X			
ConsNet (Ciarleglio et al., 2010)[10]				X			
SCP-NSGA-II (Schlottfeldt et al., 2012)						X	X

Heuristics to improve results: [1]tiebreak criterion (area). [2]Tiebreak criterion (rarity, frequency). [3]Tiebreak criterion (rarity, distance, area). [4]Tiebreak criterion (richness). [5]Tiebreak criterion. [6]Tiebreak criterion (complementarity, proximity). [7]Maximization of compactness, minimization of perimeter. [8]Connectivity, BLQ (boundary quality penalty), BLP (boundary length penalty). [9]Multicriteria analysis. [10]MASTS (modular abstract self-learning tabu search), DNS (dynamic neighborhood selection), RBO (rule-based objectives). *LP/ILP: Linear programming and integer linear programming.

In this study, we propose a more sophisticated and general solution for the SCP problem based on multi-objective optimization, which allowed us to cope with more than one objective. This provided more flexibility by including additional objectives, adding more complexity, and increasing the power of decision. In particular, we used non-dominated sorting genetic algorithm II (NSGA-II, a state-of-the-art MOEA) to search for optimal solutions. Our hypothesis was that NSGA-II could identify the smallest set of local populations of *Dipteryx alata* (also known as baru) that should be conserved to represent the known genetic diversity of this Brazilian Cerrado species, thus focusing on the *in situ* strategy. However, rather than simply representing the known alleles, the proposed approach begins with allele frequency information and incorporates information about heterozygosity in the local population and Hardy-Weinberg equilibrium. By including these 2 characteristics, local populations can be better represented in terms of their genetic diversity, allowing identification of sets of populations with a higher probability of persistence overtime. This is the first study to apply multi-objective optimization algorithms to an SCP problem with more than 2 objectives using alleles from molecular analysis at the population level as basic units.

MATERIAL AND METHODS

Data

We used data from *D. alata* (a Fabaceae tree species widely distributed and endemic to Brazilian Cerrado) consisting of 55 alleles from 9 microsatellite loci (Table 2) coding for a total of 642 individual trees sampled in 25 local populations distributed throughout species' geographical range (Figure 1), with sample sizes within the local populations ranging from 12-32 (Tables 3 and 4) (Diniz-Filho et al., 2012; Soares et al., 2012).

Based on the sampled data, we produced 4 matrices used as input for our MOEA:

1) Matrix A: an allele-by-site presence-absence matrix. Here, the population can be understood as a site, as each sampled tree was GPS-georeferenced. In matrix A_{pxa}, p = 25 (populations), a = 55 (alleles), and a_{ij} represents the occurrence of allele *j* in population *i*.

2) Matrix B: allele frequencies within local populations, where allele frequency over population was normalized in order to minimize possible distortions due to different numbers of individuals sampled among populations (i.e. differences in the sample sizes). In matrix B_{pxa}, p = 25 (populations), a = 55 (alleles), and b_{ij} is the normalized frequency of allele *j* in population *i*.

3) Matrix C: heterozygosity per population per locus. In matrix C_{pxl}, p = 25 (populations), l = 9 (loci), and c_{ij} represents the amount of heterozygosity of locus *j* in population *i*.

4) Matrix D: we used the Hardy-Weinberg equilibrium (HWE) concept which states that in the absence of evolutionary pressure, allele and genotype frequencies will remain the same along generations. The matrix D is composed of the probabilities of chi-square tests for the HWE of each locus in each local population, such that higher P values indicate a population closer to equilibrium. In matrix D_{pxl}, p = 25 (populations), l = 9 (loci), and d_{ij} represents the expected HWE of locus *j* in population *i*.

Modeling

Our problem was to identify solutions with the smallest set of *D. alata* local populations representing the genetic diversity of the species for its conservation and persistence. We considered 3 variations for the problem as follows.

Table 2. Alleles identified from 9 sequenced microsatellite from *Dipteryx alata*.

	Bm164	DaE06	DaE12	DaE20	DaE34	DaE41	DaE63	DaE67	DaE46	
	156	212	216	146	118	208	170	104	244	
	158	216	218	154	120	210	176	106	247	
	165	220	219	156	122	214		108	250	
	168		220	158	124			110	253	
	170		222		126			112		
	174				128			114		
	176				130			116		
	178				132			118		
					134			120		
					136			122		
					138			124		
					142					
					146					
					148					
					150					Total
Total	8	3	5	4	15	3	2	11	4	55

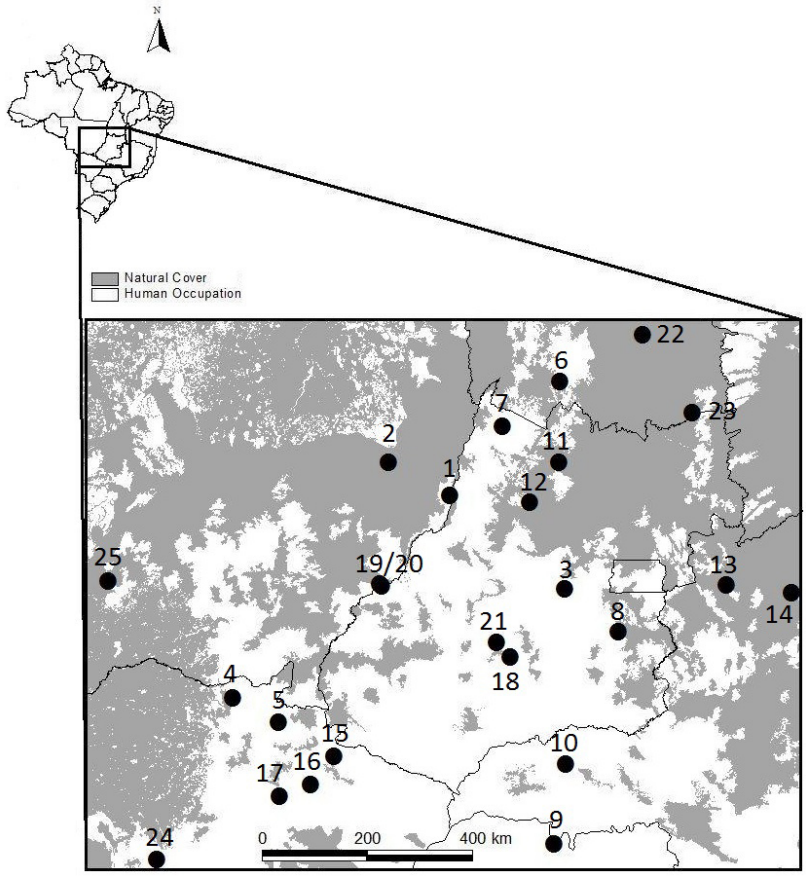


Figure 1. Geographic location of the 25 local populations of *Dipteryx alata* in Central Brazil analyzed using SCP methods based on microsatellite allelic variation. The region shown in dark tone remains covered by natural remnants of Cerrado vegetation.

Table 3. Populations and number of sampled individuals.

Population No.	Population name	No. of sampled individuals
1	CMT	32
2	ABMT	32
3	PGO	32
4	SMS	31
5	AMS	32
6	ATO	32
7	SMGO	32
8	LGO	32
9	ISP	31
10	MAMG	32
11	ENGO	12
12	STGO	12
13	AMG	32
14	PMG	32
15	PMS	13
16	PCMS	13
17	CMS	13
18	IGO	13
19	RAMT	27
20	RAGO	37
21	JGO	32
22	NTO	12
23	ARTO	15
24	AQMS	31
25	CAMT	30
Total	-	642

Table 4. Partial and schematic representation of collected data from the 642 individual trees sampled in 25 local populations throughout *Dipteryx alata*'s geographical range.

Sampled tree	Allele																	
	Bm164		DaE06		DaE12		DaE20		DaE34		DaE41		DaE63		DaE67		DaE46	
1CMT	158	158	216	216	220	220	154	154	110	110	126	126	208	208	176	170	253	253
2CMT	170	158	216	216	220	220	154	154	116	114	126	126	210	210	176	170	253	253
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
29CAMT	176	174	220	220	220	218	154	154	114	110	132	132	208	208	176	176	250	250
30CAMT	156	156	220	220	220	218	154	154	114	110	124	124	208	208	176	176	250	250

Variation 1: *Two objective optimization* (*2-D optimization*)

In this first variation, we reproduced the experiment of Diniz-Filho et al. (2012), but using a multi-objective optimization approach. We applied NSGA-II rather than simulated annealing. The latter was originally used by Diniz-Filho et al. (2012) and employs a monobjective optimization approach while dealing with SCP.

The purpose of this first problem variation was to determine the smallest set of *D. alata* local populations (Equation 1) that should be preserved in order to represent the genetic diversity of the species, targeting its *in situ* conservation, i.e. each 1 of the 55 alleles should be represented at least once (Equation 2).

Using matrix A, described in Subsection Data, a candidate solution for the problem is a vector $x = \{x_1, x_2, \ldots, x_{24}, x_{25}\}$, where $x_i \in \{0, 1\}$, such that $x_i = 1$, if population *i* is selected to compose the solution; or 0, otherwise.

The aim was to obtain:

$$min\left(\sum_{i=1}^{p} x_i\right) \quad \text{(Equation 1)}$$

Subject to:

$$\forall j \in \{1,2,...,n\},\ \sum_{i=1}^{p} a_{ij}x_i \geq 1 \quad \text{(Equation 2)}$$

where p = 25 (populations) and n = 55 (alleles).

Regarding the multi-objective optimization aproach, there are 2 objectives to be optimized:

1) Minimize the number of selected populations and

2) Maximize the number of alleles.

For simplicity, in the 2nd objective function, we used the number of missing alleles (those not present in the solution); therefore we worked with 2 minimization functions (Equations 3 and 4):

$$min(f_1(\vec{x})) = min(number\_of\_populations(\vec{x})) \quad \text{(Equation 3)}$$

$$min(f_2(\vec{x})) = min(missing\_alleles(\vec{x})) = min(55 - represented\_alleles(\vec{x})) \quad \text{(Equation 4)}$$

Variation 2: 4 objective optimization (4-D optimization)

Variation 1 solutions ensure that all 55 alleles would be represented, but not their persistence over time. One attempt to cope with this limitation would be to maximize the allele's frequency, while simultaneously prioritizing the presence of heterozygosity or HWE in the set of selected populations for conservation.

Therefore, we obtained 3 more objectives (Equations 5-7) that were combined and added to Equations 3 and 4 in order to obtain a more consistent and refined solution to predict *D. alata* persistence:

1) Using matrix B to maximize the total allele frequency (Equation 5).

$$max(f_3(\vec{x})) = max(allele\_frequency(\vec{x})) \quad \text{(Equation 5)}$$

2) Using matrix C to maximize the heterozygosity of local populations:

$$max(f_4(\overrightarrow{x})) = max(heterozygozity(\overrightarrow{x})) \quad \text{(Equation 6)}$$

2) Using matrix D to maximize the HWE level in populations (Equation 7):

$$max(f_5(\overrightarrow{x})) = max(HWE(\overrightarrow{x})) \quad \text{(Equation 7)}$$

Each equation (each dimension) is an objective function that can be optimized. We executed 10,000 independently runs of NSGA-II (see Subsection Implementation) for each combination of the 4 objectives for optimization as follows:

1) Equations 3, 4, 5, and 6;
2) Equations 3, 4, 5, and 7;

Variation 3: *5 objective optimization* (*5-D optimization*)

In this variation, we performed the optimization considering all the previous stated objectives. Five objectives were optimized simultaneously, including: number of populations, number of missing alleles, allele frequency, heterozygosity, and HWE (Equations 3-7).

NSGA-II

Evolutionary algorithms (EA) are inspired by biological evolution and use operators based on mutation, recombination, and natural selection (Bäck, 1996). Candidate solutions in EA play the role of individuals in a population. The results of previous studies suggest that EA are particularly appropriated for finding Pareto optimal solutions, particularly because they can efficiently process a set of solutions in parallel. Fonseca and Fleming (1995) as well as Valenzuela-Rendón (1998) suggested that multi-objective optimization is a research area in which EA can produce better results than traditional optimization techniques.

We used the NSGA-II (Deb et al., 2002), a state-of-the-art in MOEA (López-Jaimes and Coello-Coello, 2009). NSGA-II is a fast and elitist-based algorithm, in which individuals are classified based on a rank order, built on a dominance relationship, and a crowding operator. The best individuals are selected and evolutionary operators (crossover and mutation) are applied.

Briefly, the population is randomly initialized; prior to selection, the population is separated into categories (ranks) constructed based on dominance. For each non-dominated individual, a rank value of 1 is assigned, meaning that it belongs to the 1st Pareto front, which allows individuals to have the same potential to be selected. These individuals are removed from the population and the process of classifying the remaining individuals in their respective front ranks continues, e.g., individuals in the 2nd front receive a rank value of 2, and so on.

After assigning a rank to each individual, a value of crowding (an agglomeration comparison operator that allows prioritizing less crowded regions of solution space), is calculated. Individuals are ranked based in ascending order of their rank values and in descending order of their crowding values. The best individuals are selected and crossover and mutation operators are applied. The process continues until a stop condition is reached, e.g., a specified number of generations or a specified number of objective function evaluations.

Implementation

Candidate solutions were encoded as a binary vector of length L, where L is the number of populations (in this case, 25). Each element of the vector stores a 0 or 1 value, where 1 indicates that the corresponding population was selected to integrate a candidate solution; otherwise, a value of 0 is used.

At each execution, 500 initial solutions were randomly generated. These solutions were then evolved using NSGA-II, implemented in Matlab®.

Before running the experiments, we used a sample set to empirically estimate the most suitable parameter values, which were set to: population size = 500 individuals; crossover operator = single point crossover; crossover probability = 0.90; mutation probability = 1/L (where L is the number of populations); selection by binary tournament; mutation rate = 0.5; and number of objective functions evaluation = 100,000.

After obtaining the configuration parameters, 10,000 runs of NSGA-II for each problem variation (described in 2.2.1) were carried out. The tests were performed on 2 servers, a Hewlett-PackardProLiant DL585 G7, 4xAMD 2.8Ghz 16-cores, 512 GB RAM and a Hewlett-Packard ProLiant DL385p Gen8, 2xAMD 2.8Ghz 16-cores, 256 GB RAM.

Null model

A null model is an attempt to generate value distributions for a given variable of interest in the absence of the process under study. In experimental sciences, this allows for a "controlled situation" (Paes and Blinder, 1995; Gotelli and Graves, 1996). The main goal of using a null model is to show that the experimental results would not have emerged from randomly generated data. In this study, 10,000 populations of 500 individuals (in a total of 5,000,000 individuals) were randomly generated in order to determine whether the same results would be found without the execution of NSGA-II.

RESULTS

Variation 1: 2-D optimization

We found that the smallest population set needed to represent all 55 alleles had size 7 (Figure 2), corroborating results of Diniz-Filho et al. (2012). However, it must be highlighted that while the previous study found only 2 distinct solutions, we found 6 different solutions for the investigated problem using multi-objective optimization. Four were new solutions (S3, S4, S5, S6) and 2 were the same as those determined by Diniz-Filho et al. (2012) (S1* and S2*) (Table 5).

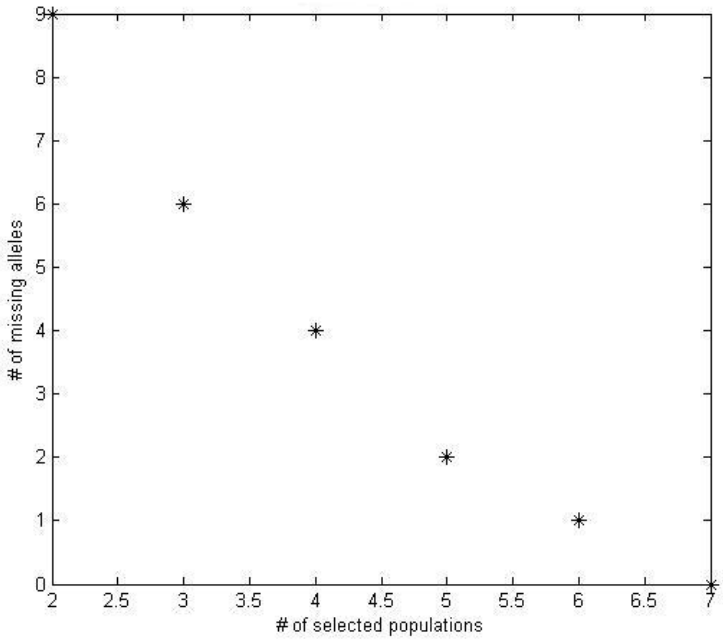


Figure 2. Pareto front obtained for problem Variation 1. Because there were 2 objectives (number of selected populations and of missing alleles), the results were plotted on a 2-D graph.

Table 5. Six solutions found for problem Variation 1. Solutions are Si, where i = 1-6. In columns Si, a value 1 indicates that the population j (j = 1-25) was included in the solution; 0, otherwise.

Population #	Population name	S1*	S2*	S3	S4	S5	S6	Irreplaceability*	Irreplaceability
1	CMT	1	1	1	1	1	1	1.00	1.00
2	ABMT	0	0	0	0	0	0	0.00	0.00
3	PGO	0	0	0	0	0	0	0.00	0.00
4	SMS	0	0	0	0	0	1	0.00	0.17
5	AMS	1	1	1	1	1	1	1.00	1.00
6	ATO	0	0	0	0	0	0	0.00	0.00
7	SMGO	0	0	0	0	1	0	0.05	0.17
8	LGO	0	0	0	0	0	0	0.00	0.00
9	ISP	0	0	0	0	0	0	0.00	0.00
10	MAMG	0	0	0	0	0	0	0.00	0.00
11	ENGO	0	0	0	0	0	0	0.00	0.00
12	STGO	0	0	0	0	0	0	0.00	0.00
13	AMG	0	0	0	0	0	0	0.00	0.00
14	PMG	0	0	0	0	0	0	0.00	0.00
15	PMS	0	1	0	1	0	0	0.50	0.33
16	PCMS	0	0	0	0	0	0	0.05	0.00
17	CMS	1	1	1	1	1	1	0.95	1.00
18	IGO	1	0	1	0	0	0	0.50	0.33
19	RAMT	1	1	1	1	1	1	1.00	1.00
20	RAGO	0	0	0	0	0	0	0.00	0.00
21	JGO	0	0	1	1	1	1	0.57	0.67
22	NTO	0	0	0	0	0	0	0.00	0.00
23	ARTO	0	0	0	0	0	0	0.00	0.00
24	AQMS	1	1	0	0	0	0	0.43	0.33
25	CAMT	1	1	1	1	1	1	1.00	1.00
-	Total	7	7	7	7	7	7	-	-

*Results found by Diniz-Filho et al. (2012).

Irreplaceability values, shown in the last 2 columns of Table 5, were obtained by considering the frequency by which a given population appeared in the solutions considering all determined solutions, including 6 in our experiment and 2 found by Diniz-Filho et al. (2012).

Local populations converging to 1 (1-CMT, 5-AMS, 17-CMS, 19-RAMT, 25-CAMT) were often irreplaceable, so that if they were lost, the conservation goal would not be achieved.

Variations 2 and 3: 4-D and 5-D optimizations

Results for 4 and 5 simultaneously optimized objectives were similar (Table 6), including the irreplaceability values shown in Figure 3.

Table 6. Partial results for minimum set of selected populations found by simultaneously optimizing 4 and 5 objectives.

No. of alleles (No. of missing alleles)	No. of selected populations	(Solution) selected populations	Variation 2		Variation 3
			4-D Heterozygosity	4-D HWE	5-D Heterozygosity & HWE
	7	(S1) 1-5-17-18-19-24-25	X	X	X
55(0)		(S2) 1-5-15-17-19-24-25	X	X	X
		(S3) 1-5-17-18-19-21-25	X	X	X
		(S4) 1-5-15-17-19-21-25	X	X	X
	Subtotal		4	4	4
55(0)	8	1-2-5-15-17-19-21-25	X	X	X
		1-2-5-15-17-19-24-25	X	X	X
		1-2-5-17-18-19-21-25	X	X	X
		1-2-5-17-18-19-24-25	X	X	X
		1-3-5-17-18-19-24-25	X	X	X
		1-4-5-15-17-19-21-25	X	X	X
		1-4-5-15-17-19-24-25	X	X	X
		1-4-5-17-18-19-21-25	X	X	X
		1-4-5-17-18-19-24-25	X	X	X
		1-5-9-17-18-19-24-25	X	X	X
		1-5-10-15-17-19-21-25	X	X	X
		1-5-10-15-17-19-24-25	X	X	X
		1-5-10-17-18-19-24-25	X	X	X
		1-5-13-15-17-19-21-25	X	X	X
		1-5-7-15-16-19-24-25		X	X
		1-5-7-16-18-19-24-25	X		X
		1-5-10-17-18-19-21-25	X	X	
		1-5-12-15-17-19-24-25	X		X
		1-5-12-17-18-19-21-25	X	X	
		1-5-13-15-17-19-24-25	X		X
		1-5-14-17-18-19-24-25		X	X
		1-5-15-17-18-19-21-25	X		X
		1-5-15-17-19-22-24-25	X		X
		1-5-16-17-18-19-24-25		X	X
		1-3-5-15-17-19-21-25	X		
		1-3-5-17-18-19-21-25		X	
		1-5-7-15-17-19-24-25			X
		1-5-8-17-18-19-24-25		X	
		1-5-9-15-17-19-24-25	X		
		1-5-9-17-18-19-21-25	X		
		1-5-11-15-17-19-24-25			X
		1-5-11-17-18-19-24-25			X
		1-5-12-17-18-19-24-25			X
		1-5-14-15-17-19-21-25		X	
		1-5-15-16-17-19-24-25	X		
		1-5-15-17-19-20-24-25	X		
		1-5-15-17-19-21-22-25	X		
		1-5-15-17-19-21-23-25			X
		1-5-17-18-19-20-21-25		X	
		1-5-17-18-19-21-22-25		X	
		1-5-17-18-19-21-24-25			X
	Subtotal		27	24	28
	Total		31	28	32

The optimization objectives are: 1) minimize number of selected population (set size); 2) minimize number of missing alleles; 3) maximize allele frequency; 4) maximize heterozygosity; 5) maximize HWE. Solutions Si, where i = 1-4, are the same corresponding to solutions on Table 5. X indicates that the solution was found by the corresponding approach.

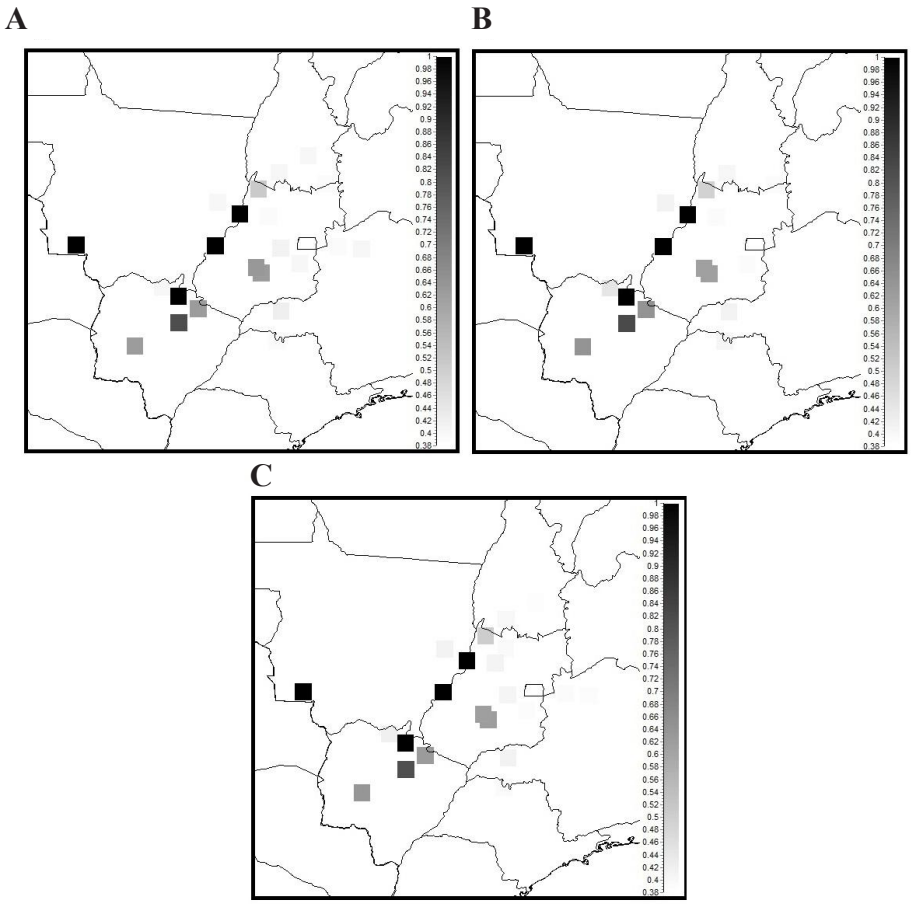


Figure 3. Irreplaceability values for (**A**) Variation 2 (4-D optimization) with heterozygosity as the 4th objective; (**B**) Variation 2 with HWE as the 4th objective and (**C**) Variation 3 (5-D optimization). Irreplaceability values were based on the frequency that a given population appeared in distinct solutions. A value converging to 1 means that the given local population was generally irreplaceable, in the sense that if it was not present in the solution, the conservation goal may not be achieved. Experimental irreplaceability values found for Variations 2 and 3 were very similar.

It is worth note that, in Variation 1, 7 was the minimum set size necessary to represent all 55 alleles, but because there were additional objectives, we used multi-objective optimization to obtain a much larger portfolio of solutions with populations ranging from 7-22 (only the solutions with a set of 7 and 8 populations are shown in Table 6), offering decision-makers a larger spectrum of options that fulfill the stated objectives.

We observed no hierarchy among the results shown in Table 6, indicating that all found solutions are optimal in the considered context, i.e., none can be declared the best when all optimized objectives are considered.

For 7 selected populations, both Variations 2 and 3 (4-D and 5-D optimizations) identified 4 from the 6 solutions identified in Variation 1, corresponding to S1, S2, S3, and S4 shown in Table 5. These 4 solutions are shown graphically in the right lower corner of the graph in Figure 4.

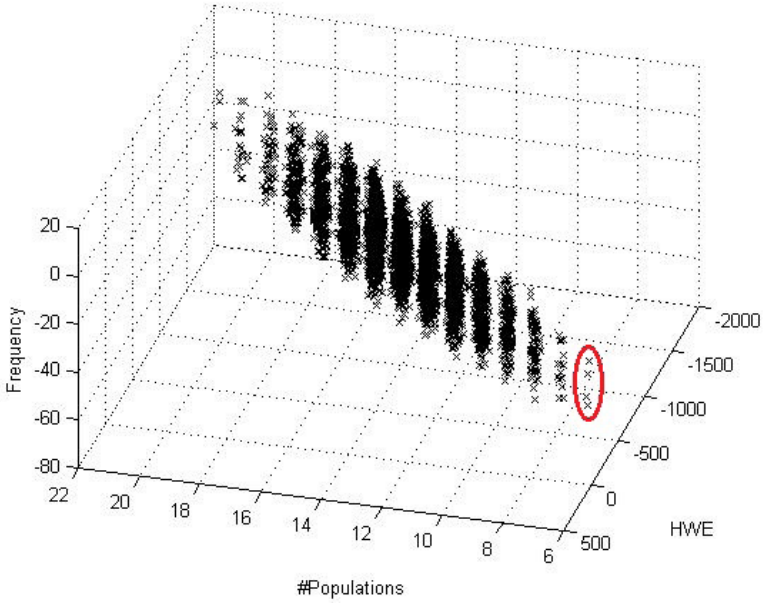


Figure 4. Pareto front obtained for problem Variation 3 (5-D optimization). There were 5 objectives (number of populations, allele frequency, HWE, heterozygosity, and number of missing alleles); the first 3 are plotted in this 3-D graph. Solutions with 7 populations are shown on the lower right corner (as indicated). It is worth note that the value for missing alleles of all solutions in this Pareto front was 0.

Table 7 shows the percentages for solutions with 7 populations found by optimization performed in Variations 2 and 3.

Table 7. Distribution (in percentage) of the solutions with 7 populations found by 4-D and 5-D optimizations. Solution Si, where i = 1-4, are the same corresponding to solutions on Table 5.

Solution	Variation 2		Variation 3
	4-D Heterozygosity	4-D HWE	5-D Heterozygosity & HWE
S1	22.78%	25.26%	1.58%
S2	20.06%	24.53%	98.34%
S3	31.44%	24.60%	0.02%
S4	25.72%	25.60%	0.05%

Null model

From 5,000,000 randomly generated individuals (10,000 populations of 500 individuals each), we obtained 351,547 with 7 populations (lower bound), approximately 7% of all randomly generated individuals. Furthermore, when compared to the results obtained for problem Variation 1, only 3 of the 351,547 results represented all 55 alleles (corresponding to 0.00006% of the randomly generated individuals). This indicates that NSGA-II experimental results were not generated by chance.

For each randomly generated individual, we determined heterozygosity and HWE values. Except for the 3 solutions with no missing alleles, the other 351,544 randomly generated solutions with 7 populations (number of populations presented in Table 6), heterozygosity and HWE values were worse (lesser) when compared to those found using NSGA-II.

DISCUSSION

The most popular methods, algorithms, and tools used for the SCP problem involve a monobjective approach; they aggregate the distinct objectives into a single function. Hence, despite the clear multi-objective aspect of SCP, optimization models often treat this as monobjective by assigning different weights to different objectives of the problem in order to aggregate the objectives into a single objective function, known as a fitness function (Zitzler et al., 2002). However, when 2 criteria represent different value systems, it may be impossible to combine such criteria in a meaningful way (e.g., comparing the preservation of endangered species and the promotion of economic development). A single objective function can result in the association of utterly disparate elements, requiring assumptions that decision-makers may consider inadequate and even leading to inaccurate results (Ciarleglio et al., 2010). This prompted us to use a multi-objective approach for solving SCP.

Moreover, it is not a trivial task to properly weight different conflicting objectives in order to combine them accordingly, and it is generally necessary to have expert knowledge that is not always available. In fact, depending on how the objectives are associated to compose the unique fitness function, one may obtain completely different results.

By applying multi-objective optimization, it is possible to overcome these issues because the objectives can simultaneously be optimized independently from one another.

Diniz-Filho et al. (2012) recently proposed an explicit complementary approach that could be used to optimize conservation of genetic variability, expressed as allelic variation (presence-absence of alleles) derived from microsatellite loci, and solved an SCP problem for the conservation of *D. alata* using a simulated annealing algorithm.

Nonetheless, these previous attempts were monobjective as well as used only the presence-absence of alleles in local populations, which is not as informative as using the allele frequencies directly. This approach more properly reflects the ecological and evolutionary processes driving genetic diversity in local populations and may be more related to population persistence. Using allele frequencies is equivalent to applying more complex characteristics of a species, such as abundance and environmental suitability in SCP higher hierarchical levels, potentially improving the long-term persistence of the conservation networks and providing the opportunity to test the multi-objective method in this new conservation problem at lower hierarchical level than the species level.

As stated above, in standard applications of SCP using software such as Marxan and based on simulated annealing (Table 1), the constraints are typically expressed as weights, obtained from a complex function combining several conflicting attributes. Hence, the approach proposed here, in addition to being the first to use an evolutionary algorithm to address the SCP problem, is also the first to use a legitimate multi-objective approach. This allows more flexibility by including additional decision objectives, as well as adds more complexity to results while increasing decision-maker options.

In Variation 1, developed to reproduce the experiment of Diniz-Filho et al. (2012), we identified a larger number of distinct solutions (6 rather than 2 as previously reported).

For Variations 2 and 3, we found similar results and significantly expanded the portfolio of solutions when considering options globally. Thus, in Variation 1, the method was not able to identify new options with more than 7 populations. By including a larger number of objectives, we introduced a degree of flexibility that allowed the identification of other solutions with more populations but with optimized allele frequency, heterozygosity, or HWE. Although it is desirable to have fewer populations, this is advantageous because a minimal set of populations representing all alleles is not necessarily the one with the best results when considering intra-specific diversity and persistence.

A key point is that Variations 2 and 3 were able to refine the Variation 1 solutions, indicating the optimization of persistence features (S1-S4) (Table 6). Additionally, the distribution (in percentage) of solutions with 7 populations found using 4-D optimizations (Table 7 2nd and 3rd columns) showed no significant differences between the results obtained by optimization performed in Variation 2, as they were similarly distributed from S1-S4. However, for 5-D optimization (Table 7 last column), although there was no hierarchy among these solutions (S1-S4) in the sense that they were optimal when all optimized objectives were considered, the results for Variation 3 clearly highlighted S2, as S2 corresponds to 98.34% of the solutions with 7 populations. This result suggests that optimization of 5 objectives can be used to further refine the results.

As the decision-makers' portfolio increases considerably, the method can indicate the most adequate options in this portfolio, which is advantageous.

In addition, it appears that the simultaneous optimization of heterozygosity and HWE in Variation 3 is advantageous, as applying it allowed the method to identify more distinct solutions for 8 populations. There were 28 solutions compared to 27 and 24 from Variation 2, which used heterozygosity and HWE, respectively, as the 4th objective (Table 6).

Considering that for 7 populations, the same results were found using 4-D and 5-D optimizations, it can be said that heterozygosity and HWE guided the set of solutions in the same direction. For 8 populations, of the 41 distinct solutions, 14 (34%) were determined using the 4-D and the 5-D optimizations, and 24 (58%) by at least 2 of the 3 optimizations performed.

The results from the null model assured that solutions obtained using NSGA-II did not emerge randomly. Only 3 results from the null model included 7 populations and all 55 alleles, while more than 65,000 solutions retrieved using NSGA-II presented these characteristics.

The possibility of dealing with these more complex situations clearly shows the advantage of our method as compared with the standard approach based on simulated annealing, as implemented by Diniz-Filho et al. (2012).

Considering that the most commonly used tools for SCP apply algorithms on a monobjective approach, our results show that dealing with various conflicting objectives in more than one dimension (i.e. using multi-objective optimization) allows for amore sophisticated and general solution to the SCP problem. This can be observed by the variety of different solutions generated using our method (instead of a single point generated using monobjective methods), thus improving the decision-maker support.

In conclusion, we used a multi-objective approach to solve variations of the SCP problem with more than two objectives, which added complexity and increased the decision-makers options. We showed that a multi-objective algorithm is more powerful and opens more possibilities than the methods previously used, such as simulated annealing.

We implemented a more refined search for optimal solutions to the problem of finding the smallest set of local populations of *D. alata* that should be conserved in order to represent genetic diversity based on allele frequency information associated with heterozygosity and

Hardy-Weinberg equilibrium. This was the first time these parameters (objectives) were used in the context of SCP.

We found that the smallest set of populations needed to represent all alleles under study was 7, corroborating the 2 solutions determined by Diniz-Filho et al. (2012), but we obtained more options of distinct solutions (the previous 2 solutions as well as 4 additional solutions, for a total of 6). By optimizing 4 and 5 objectives simultaneously (4-D and 5-D optimizations), we found 4 solutions for 7 populations, refining the 6 previously determined solutions. Additionally, we obtained a larger portfolio in terms of intra-specific diversity and persistence with populations ranging from 8-22. In particular, for 8 populations we found 41 solutions.

Although additional experiments should be conducted to improve the NSGA-II fitness function in a more precise and controlled manner, our results demonstrate the advantages of the new approach with respect to previous solutions. Additionally, this was the first attempt to apply multi-objective algorithms to an SCP problem with more than 2 dimensions based on molecular data at the population level as basic units. Our results can be used to propose a more specialized multi-objective algorithm for SCP problems, allowing researchers to deal with such problems in a more efficient and appropriate manner.

ACKNOWLEDGMENTS

S. Schlottfeldt wishes to thank the University of York and Prof. Jon Timmis for the PhD stay and the support from CNPq, through the Science without Borders Program. The research program in molecular ecology and conservation of Cerrado plants has been continuously supported by several grants and fellowships to the research network GENPAC (Geographical Genetics and Regional Planning for Natural Resources in Brazilian Cerrado) from CNPq/MCT/CAPES (projects #564717/2010-0 and 563624/2010-8) and by the "Núcleo de Excelência em Genética e Conservação de Espécies do Cerrado" - GECER (PRONEX/FAPEG/CNPq CP 07-2009). Field work was supported by Systema Natura e Consultoria Ambiental LTDA and work by M.E.M.T. Walter, A.C.P.L.F. Carvalho, M.P.C. Telles, R.D. Loyola, and J.A.F. Dinz-Filho have been continuously supported by productivity fellowships from CNPq.

REFERENCES

Ackery PR and Vane-Wright RI (1984). Milkweed butterflies, their cladistics and biology: being an account of the natural history of the Danainae, a subfamily of the Lepidoptera, Nymphalidae. Natural History Museum, London.

Ardron JA, Possingham HP and Klein CJ (2010). Marxan Good Practices Handbook, Pacific Marine Analysis and Research Association (PacMARA), Victoria.

Arthur JF, Hachey M, Sahr K, Huso M, et al. (1997). Finding all optimal solutions to the reserve site selection problem: formulation and computational analysis. *Environ. Ecol. Stat.* 4: 153-165.

Bäck T (1996). Evolutionary algorithms in theory and practice: evolution strategies, evolutionary programming, genetic algorithms. Oxford University Press, Oxford.

Ball IR (2000). Mathematical Applications for Conservation Ecology: The Dynamics of Tree Hollows and the Design of Nature Reserves. Doctoral Thesis. Depts. of Applied Mathematics, Environmental Science and Management, University of Adelaide, Adelaide.

Barton DN, Rusch G, Gjershaug JO, Faith DP, et al. (2004). TARGET as a tool for prioritising biodiversity conservation payments on private land - a sensitivity analysis. Technical Report SNR 4856-2004. Norwegian Institute for Water Research (NIVA), Oslo.

Brooks TM, da Fonseca GAB and Rodrigues ASL (2004). Species, data, and conservation planning. *Conserv. Biol.* 18: 1682-1688.

Ciarleglio M, Barnes JW and Sarkar S (2010). ConsNet: A tabu search approach to the spatially coherent conservation area network design problem. *J. Heuristics* 16: 537-557.

Coello-Coello CA, Lamont GB and Veldhuizen DAV (2007).Evolutionary Algorithms for Solving Multi-Objective Problems. 2nd edn. Springer-Verlag, New York.

Cormen TH, Stein C, Rivest RL and Leiserson CH (2001). Introduction to Algorithms. 2nd edn. MIT Press, Cambridge, MA.

Deb K, Agarwal S, Pratap A and Meyarivan T (2002). A fast elitist non-dominated sorting genetic algorithm for multi-objective optimization: NSGA-II. *IEEE T. Evol. Comput.* 6: 182-197.

Diniz-Filho JAF and Telles MPC (2002). Spatial autocorrelation analysis and the identification of operational units for conservation in continuous populations. *Conserv. Biol*. 16: 924-935.

Diniz-Filho JAF, Loyola RD, Melo DB, Oliveira G, et al.(2012). Planning for optimal conservation geographical genetic variability within species. *Conserv. Genet*. 13: 1085-1093.

Fonseca CM and Fleming PJ (1995). An overview of evolutionary algorithms in multi-objective optimization. *Evol. Comput*. 3: 1-16.

Fraser DJ and Bernatchez LB (2001). Adaptive evolutionary conservation: toward a unified concept for defining conservation units. *Mol. Ecol*. 10: 2741-2752.

Garson J, Aggarwal A and Sarkar S (2002). ResNet Manual Version 1.2. User Manual. Biodiversity and Biocultural Conservation Laboratory, University of Texas at Austin, Austin.

Gotelli NJ and Graves GR (1996). Null Models in Ecology. Smithsonian Institution Press, Washington, DC.

Kirkpatrick JB (1983). An interactive method for establishing priorities for the selection of nature reserves - an example from Tasmania. *Biol. Conserv*. 25: 127-134.

López-Jaimes A and Coello-Coello CA (2009). Multi-Objective Evolutionary Algorithms: A Review of the State-of-the-Art and some of their Applications in Chemical Engineering (Pandu RG, ed.). Multi-Objective Optimization Techniques and Applications in Chemical Engineering, World Scientific, Singapore, 61-90.

Margules CR and Nicholls A (1987). Assessing the conservation value of remnant habitat "islands": Mallee patches on the Western Eyre peninsula (Saunders DA, Arnold GW, Burbridge AA and Hopkins AJM, eds.). Nature Conservation: The role of remnants of native vegetation. Surrey Beatty and Sons, Baulkham Hills, 89-102.

Margules CR and Pressey RL (2000). Systematic conservation planning. *Nature* 405: 243-253.

Margules CR, Nicholls A and Pressey R (1988). Selecting networks of reserves to maximize biological diversity. *Biol. Conserv*. 43:63-76.

Moffett A, Garson J and Sarkar S (2005). MultCSync: a software package for incorporating multiple criteria in conservation planning. *Environ. Modell. Softw.* 20: 1315-1322.

Moilanen A (2005). Reserve selection using nonlinear species distribution models. *Am. Nat.* 165: 695-706.

Nicholls A and Margules CR (1993). An updated reserve selection algorithm. *Biol. Conserv*. 64: 165-169.

Paes E and Blinder PB (1995). Modelos nulos e processos de aleatorização: algumas aplicações em Ecologia de Comunidades (Peres-Neto PR, Valentin JL, Fernandez FAZ, eds.). Oecologia Brasiliensis Volume II: Tópicos em Tratamento de Dados Biológicos. Instituto de Biologia, UFRJ, Rio de Janeiro, 119-139.

Possingham HP, Ball I and Andelman S (2000). Mathematical methods for identifying representative reserve networks (Ferson S and Burgman M, eds.). Quantitative methods for conservation biology. Springer-Verlag, New York, 291-305.

Pressey RL (2004). Conservation planning and biodiversity: Assembling the best data for the job. *Conserv. Biol.* 18: 1677-1681.

Pressey RL, Possingham HP and Margules CR (1996). Optimality in reserve selection algorithms: when does it matter and how much? *Biol. Conserv*. 76: 259-267.

Rebelo AG and Siegfried W (1990). Protection of Fynbos vegetation: ideal and real-world options. *Biol. Cons*. 54: 15-31.

Sarkar S (2012). Complementarity and the selection of nature reserves: algorithms and the origins of conservation planning, 1980-1995. *Arch. Hist. Exact Sci*. 66: 397-426.

Schlottfeldt S, Walter MEMT, Diniz-Filho JAF and Telles MPC (2012). Multi-objective Optimization in Systematic Conservation Planning to Represent Genetic Variability within Species. In: 8th International Conference on Ecological Informatics, ISEI2012, Brasília.

Soares TN, Melo DB, Resende LV, Vianello RP, et al. (2012). Development of microsatellite markers for the neotropical tree species *Dipteryx alata* (*Fabaceae*). *Am. J. Bot.* 99: e72-e73.

Underhill LG (1994). Optimal and suboptimal reserve selection algorithms. *Biol. Conserv*. 70: 85-87.

Valenzuela-Rendón M (1998). Reinforcement learning in the fuzzy classifier system. *Expert Syst. Appl*. 14: 237-247.

Zitzler E, Laumanns M and Thiele L (2002). SPEA2: Improving the Strength Pareto Evolutionary Algorithm for Multi-objective Optimization (Giannakoglou K, Tsahalis D, Periaux J, Papaliliou K, et al., eds.). Evolutionary Methods for Design, Optimisation and Control with Application to Industrial Problems. Proceedings of the EUROGEN2001 Conference, Athens, Greece, September 19-21, 2001, 95-100.

8.4 *Paper 3*

Schlottfeldt, S.; Walter, M.E.M.T.; de Carvalho, A.C.P.L.F; Soares, T.N.; Loyola, R.D.; Telles, M.P.C.; Diniz-Filho, J.A.F. Multi-Objective Optimization for Plant Germplasm Collection Conservation of Genetic Resources Based on Molecular Variability. Tree Genet Genomes, 11(2):16, 2015. ISSN 1614-2942 (versão impressa)/ISSN 1614-2950 (versão eletrônica).

A publicação final está disponível em *Spriger* via
http://dx.doi.org/10.1007/s11295-015-0836-3

Material suplementar disponível no *Dryad Digital Repository* via
http://doi.org/10.5061/dryad.hq8pd

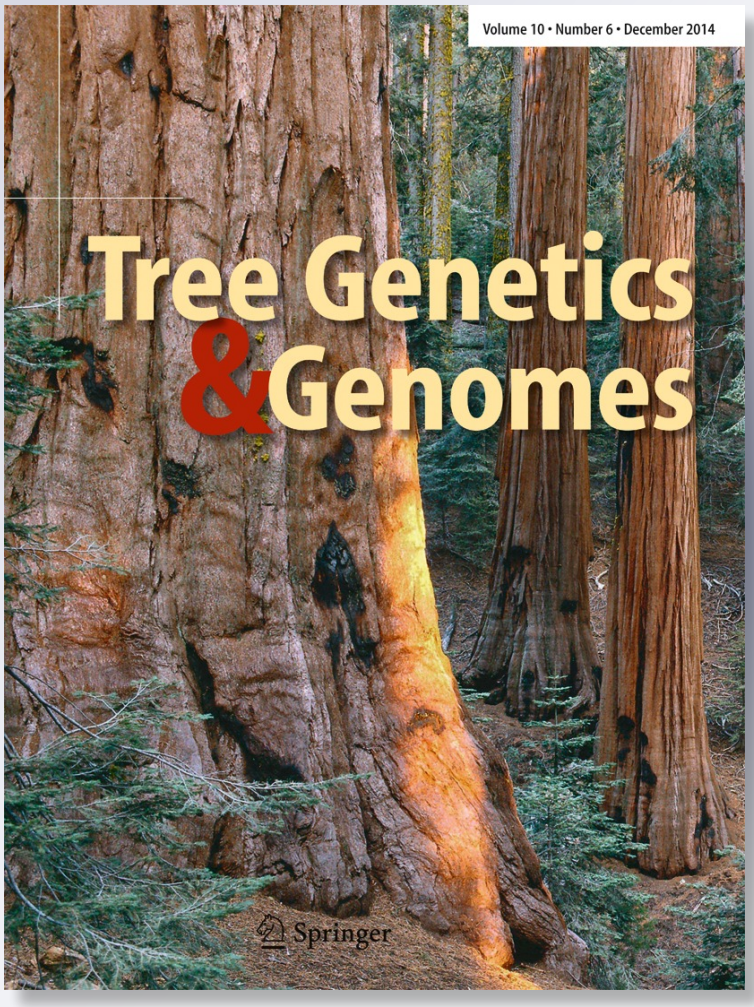
Multi-objective optimization for plant germplasm collection conservation of genetic resources based on molecular variability

Shana Schlottfeldt, Maria Emília M. T. Walter, André Carlos P. L. F. de Carvalho, Thannya N. Soares, Mariana P. C. Telles, et al.

Tree Genetics & Genomes

ISSN 1614-2942
Volume 11
Number 2

Tree Genetics & Genomes (2015)
11:1-10
DOI 10.1007/s11295-015-0836-3



Springer

Tree Genetics & Genomes
DOI 10.1007/s11295-015-0836-3

ORIGINAL PAPER

Multi-objective optimization for plant germplasm collection conservation of genetic resources based on molecular variability

Shana Schlottfeldt · Maria Emília M. T. Walter · André Carlos P. L. F. de Carvalho · Thannya N. Soares · Mariana P. C. Telles · Rafael D. Loyola · José Alexandre F. Diniz-Filho

Received: 21 September 2014 / Revised: 22 December 2014 / Accepted: 6 January 2015

Abstract Germplasm collections play a significant role among strategies for conservation of diversity. It is common to select a core collection to represent the genetic diversity of a germplasm collection, in order to minimize the cost of conservation, while ensuring the maximization of genetic variation. We aimed to solve two main problems: (1) to select a set of individuals, from an in situ data set, that is genetically complementary to an existing germplasm collection, and (2) to define a core collection for a germplasm collection. We proposed a new multi-objective optimization (MOO) approach based on principles of systematic conservation planning (SCP) incorporating heterozygosity information; therefore, optimization takes genotypic diversity and variability patterns into account as well. As a case study, we used *Dipteryx alata* microsatellite loci information from two sources, an ex situ germplasm collection located at the Agronomy School of the Federal University of Goiás (UFG-AS), and an in situ data set composed of 642 sampled individual trees. We were able to identify within a population of several individuals, the exact accessions/samples that should be chosen in order to preserve the species diversity. We found that material from nine in situ individual trees are enough to complement the UFG-AS germplasm collection as it is, and that it is possible to define a core collection of 20 individual trees representing all studied genetic diversity. Moreover, we defined a method (a protocol) to deal with large amounts of accessions in the context of MOO. The proposed approach can be used to help constructing collections with maximal allelic richness and can also be extended to the in situ conservation. As far as we know, this is the first time that principles of SCP and the MOO approach are applied to the problem of complementing a germplasm collection and of finding a core collection for a germplasm collection.

Keywords Biodiversity · Conservation planning · Core collection · Genetic variability · Germplasm · Multi-objective optimization

Communicated by: J. Beaulieu

Electronic supplementary material The online version of this article (doi:10.1007/s11295-015-0836-3) contains supplementary material, which is available to authorized users.

S. Schlottfeldt (✉) · M. E. M. T. Walter
Department of Computer Science, University of Brasilia, Brasilia, DF, 70910-900 Brazil
e-mail: shanass@unb.br

M. E. M. T. Walter
e-mail: mariaemilia@unb.br

A. C. P. L. F. Carvalho
Department of Computer Science, University of São Paulo, SCC–ICMC–USP, CxP 668, São Carlos, SP, 13560-970 Brazil
e-mail: andre@icmc.usp.br

T. N. Soares · M. P. C. Telles
Department of General Biology, Federal University of Goiás, CxP 131, Goiânia, GO, 74001-970 Brazil
e-mail: tnsoares@gmail.com

M.P.C. Telles
e-mail: tellesmpc@gmail.com

R. D. Loyola · J. A. F. Diniz-Filho
Department of Ecology, Federal University of Goiás, CxP 131, Goiânia, GO, 74001-970 Brazil
e-mail: rdiasloyola@gmail.com

J.A.F. Diniz-Filho
e-mail: diniz@ufg.br

Published online: 05 February 2015

Introduction

Germplasm corresponds to the living tissues from which new plants can be grown, and thus are an important component for maintenance of plant genetic resources (Roederer et al. 2000). Plant germplasm can be stored as seed collections, pollen storage, in a nursery (field), in vitro (Engels et al. 2003) constituting what is called germplasm bank or collection. These germplasm collections are kept to represent the genetic diversity of plants, their wild relatives, and/or plants present and unique in a local region (Dawson and Were 1997). They play a significant role among strategies for conservation of diversity (Rao et al. 2006). Additionally, to commonly occurring species, rare, threatened, or endangered species are made available for study or for habitat restoration projects.

Core collections are useful tools for organizing and analyzing representative sets of genotypes in a germplasm collection and can be defined based on several criteria, including explicit evaluation using molecular markers. A core collection is a subset from a larger collection of a particular species that represents, with a minimum level of repetitiveness, the genetic diversity of that species and its wild relatives (Frankel 1984; Brown 1989). A core collection should not be considered a substitute of the whole collection, but it captures the complete diversity of the entire collection it was derived from. Therefore, a core collection should correspond to a set representing all of the species alleles, while ensuring minimum redundancy of those alleles and high reproducibility of entries. Core collections are being adopted as a useful instrument to improve conservation, accessibility and the use of plant genetic resources (Zhang et al. 2011). In this conservation context, one could understand that remaining natural populations of the species could be viewed as in situ genetic resources, whereas a core collection could be an ex situ sample that could be stored for further conservation applications. In this conservation context, the use of molecular markers to achieve the representativeness definition of a core collection is important because population persistence and resilience to environmental changes are usually positively correlated with genetic diversity. Methodologically, efforts to create germplasm core collections commonly use statistical and clustering methods (Holbrook et al. 1993; Li et al. 2004; Laghetti et al. 2008; Wang et al. 2011; Zhang et al. 2011; Belaj et al. 2012; Mei et al. 2012; Rao et al. 2012).

There is always a cost involved in maintaining germplasm collections such as maintenance of storage space, controlled temperature, and relative humidity (Rao et al. 2002). Various studies for estimating the costs of conservation have been carried out adopting different methodologies (Gupta et al. 2002). Germplasm collection operations and facilities vary substantially in size, capacity, types, and amount of equipments. These factors depend on the quantities of germplasm to be stored in the germplasm collection, which will in turn depend upon several circumstances that vary from objectives of the germplasm collection, range of species and breeds to be conserved, to financial resources available for the conservation program.

Cerrado is a large biome in Central Brazil, occupying about $1,500,000$ km^2 and includes significant environmental heterogeneity. It is a typical tropical savanna environment, but actually has different types of vegetation, ranging from open grasslands to dense woodlands and dry forests. Cerrado has been considered one global biodiversity hot spot (Myers et al. 2000), because of the strong plant endemism and high rates of habitat loss and fragmentation due to a recent expansion of soybean cultures and cattle ranching (Diniz-Filho et al. 2008). The conservation of indigenous plant species germplasm is important since considerable material has been identified as unique to the Brazilian Cerrado biodiversity. In this context, germplasm collections exist to conserve, increase utilization, and catalogue germplasm of plants that might otherwise be lost.

At the Agronomy School of the Federal University of Goiás (UFG-AS), Brazil, there is a large accessible germplasm collection for *Dipteryx alata* (also known as *baru*, a Fabaceae tree species widely distributed and endemic to Brazilian Cerrado), composed of 178 fully grown adult trees. Besides, data from *D. alata* was collected in situ for a total of 642 individual trees sampled in 25 local populations throughout species' geographic range (Fig. 1).

In this context, using data from *D. alata* as a case study for our method, we aim to solve two main general problems:

1. To select a set of individuals from in situ collected data that is genetically complementary to an existing germplasm collection;
2. To define a core collection for a germplasm collection.

In both problems, we look for minimizing the overall cost of conservation while maximizing the allele representation. On the one hand, these problems can be mapped to the systematic conservation planning (SCP), a widely accepted biodiversity-focused approach to selecting priority areas for protection. In a simplified way, SCP is the problem of finding a minimum set of sites (in this case, individual trees) with the maximum representation of some feature (Margules and Pressey 2000). On the other hand, the SCP problem can be modeled by the minimum set covering problem, that was shown to be NP-hard (Cormen et al. 2001), i.e., there is no known efficient exact solution for the problem, therefore, when the input grows arithmetically, the time to find a solution increases exponentially. Clearly, there are two conflicting objectives (minimize selected individuals

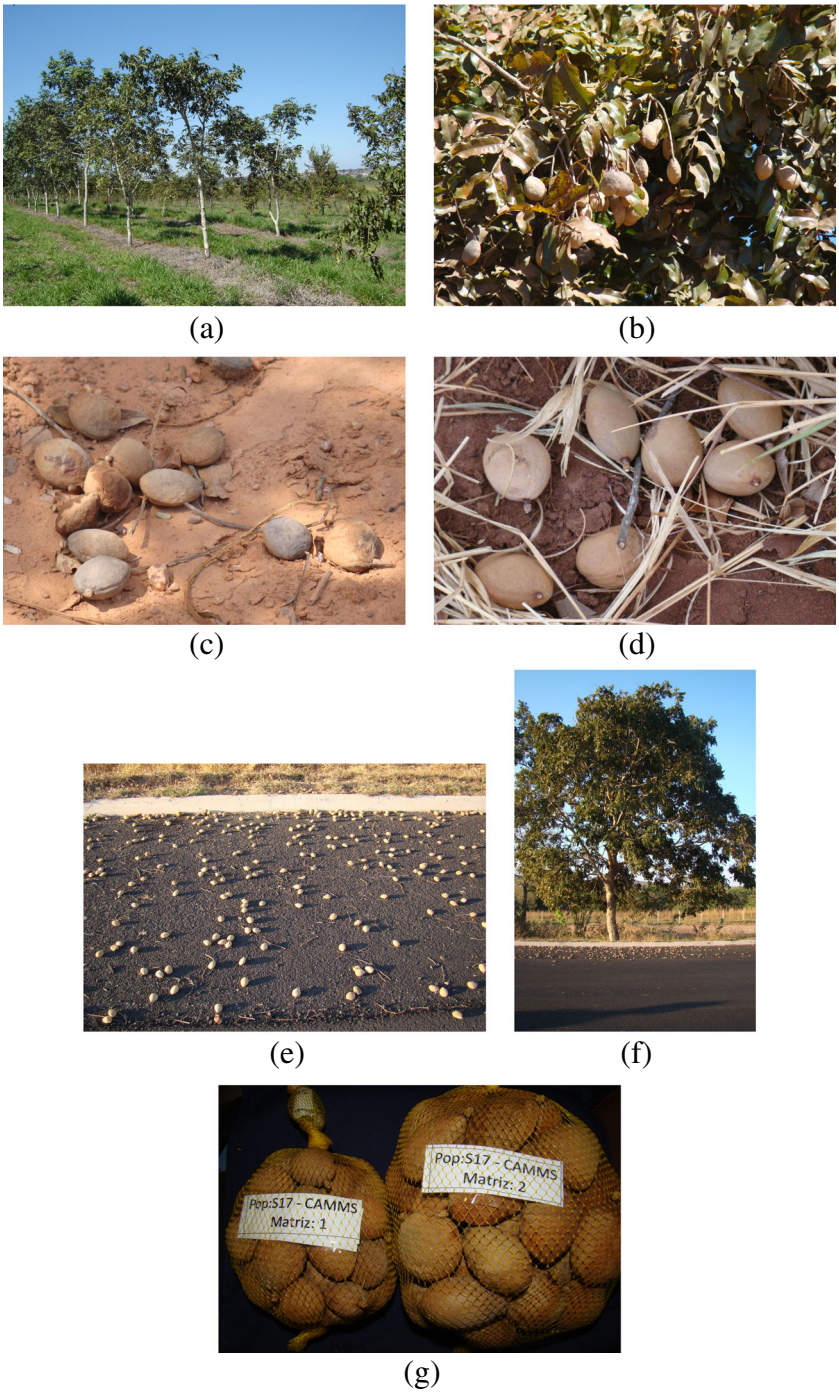


Fig. 1 *Dipteryx alata*, also known as *baru*, is a widely distributed tree species in the Cerrado biome, Central Brazil, although restricted to seasonal savannas that grow in eutrophic and drained soil. Fruits have a very woody endocarp with an edible nut that is eaten and dispersed by mammals, e.g., bats and monkeys. The *D. alata* is used as lumber, for charcoal production, shade in pasture, and it is a source of raw material for handicraft, cosmetics, and food industries, playing an important role in the local economy (Correa et al. 2008; Collevatti et al. 2013). **a** The UFG-AS *D. alata* germplasm collection (nursery), composed of 178 fully grown adult trees. **b** *Baru* fruit in situ. **c**–**e** Ripe baru fruits are collected from the ground. **f** Individual tree in situ. **g** In situ collected samples (samples were collected for a total of 642 individual trees in 25 local populations distributed throughout species' geographic range)

while maximizing allele representation), making the problem a perfect candidate for multi-objective optimization (MOO).

Optimization problems with more than one objective are called vector optimization or multi-objective problems (MOP) (Brockhoff 2009; Deb 2008; Coello Coello 2001; Coello Coello et al. 2007; Zitzler 2002). In these cases, there is no single optimal solution, but rather, a set of solutions that should be considered equivalents in the absence of information about the relevance of each objective related to the others (Fonseca and Fleming 1995). These solutions are known as Pareto optimal solutions (Fig. 2), and their plot form what is called Pareto front (Fig. 3) (Coello Coello et al. 2007).

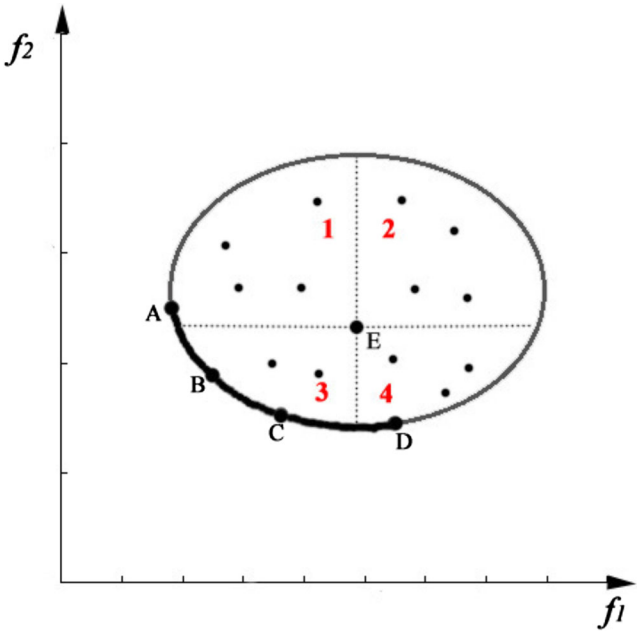


Fig. 2 Pareto optimality. In multiobjective problems (MOP) there is not only one objective function to optimize, but many, thus, the aim is to find good trade-offs rather than a single solution. The notion of optimum is based on the Pareto dominance: considering two objective functions (f_1 and f_2), a solution (a point) can be better, worse, equal, or also indifferent to another solution with respect to the objective values computed from functions f_1 and f_2. "Better" means that a solution is not worse in any objective and it is better in at least one objective; this solution is also said to dominate the other ones. Using this concept, an optimal solution is not dominated by any other solution. Such a solution is called Pareto optimal, and the entire set of optimal trade-offs is called the Pareto optimal set, which is represented here by points *A*, *B*, *C*, and *D*. Considering a minimization problem, taking solution *E* as reference, solutions located in: area *2* are dominated by *E* (are worse than *E*); area *3* dominate *E* (are better than *E*); areas *1* and *4* are indifferent (it is not possible to compare them, since if *E* is better in f_1, it is worse in f_2 and vice versa)

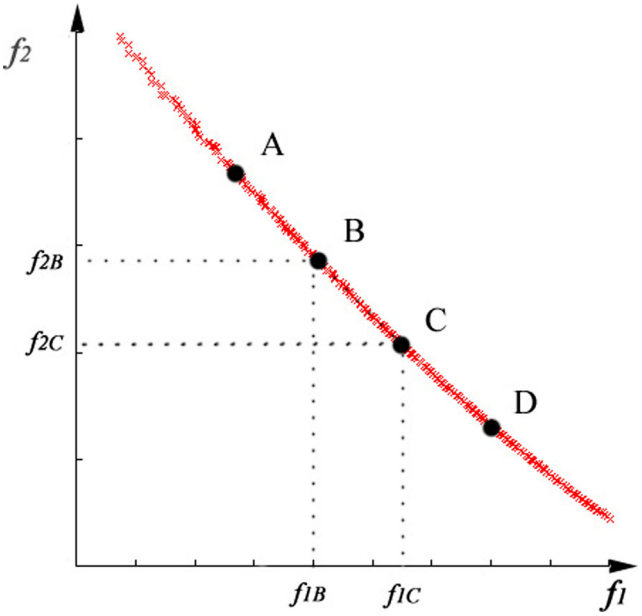


Fig. 3 Pareto front. An example of a problem with two objective functions to be simultaneously minimized: f_1 and f_2. Observe that for points *B* and *C*, it is not possible to obtain an improvement in one objective without a degradation of the other objective, i.e., $f_{1B} < f_{1C}$, $f_{2B} > f_{2C}$. The solutions (points) *A*, *B*, *C*, and *D* are optimal when f_1 and f_2 are considered. The Pareto front (or trade-off surface) is indicated by the *curve*

SCP has been generally used at species level (or hierarchically higher), but has also been applied to conservation genetics, aiming to maximize molecular variation within populations (Diniz-Filho and Telles 2002; 2006; Diniz-Filho et al. 2012b). To properly use this method to organize germplasm collections, it would be important to improve the approach to analyze individuals and their genotypes.

In a previous work, our group solved a problem that looked for the smallest set of local populations of *D. alata* that might be preserved in order to represent the genetic diversity of this species to its in situ conservation (Schlottfeldt-Santos et al. 2012; Schlottfeldt et al. 2014). That study was pioneer in the use of information about allele frequency, heterozygosity, and Hardy-Weinberg equilibrium as objectives for simultaneous optimization.

Diniz-Filho et al. (2012b) used a mono-objective approach (in particular, simulated annealing) to find the minimum amount of populations, not specifically individuals, needed to represent all diversity in in situ conservation of *D. alata*. Here, we propose a multi-objective optimization approach for the conservation of genetic resources based on individual molecular variability aiming to inform ex situ conservation strategies, and to guide sampling for germplasm collections. We represented the known alleles, but incorporated individual heterozygosity information, thus optimization also takes genotypic diversity and variability patterns into account. By including these characteristics, individuals can better represent the genetic diversity, allowing to identify sets of individuals with a higher probability of persistence throughout time.

As far as we know, this is the first time that a SCP approach dealing with multi-objective optimization is applied to the problems of finding a core collection for a germplasm collection and complementing this core collection using heterozygosity information as well.

Material and methods

Data

As a case study for our method, we used data from *D. alata* (data available on Supplementary_File_S2, and Supplementary_File_S3):

1. An ex situ data set corresponding to the germplasm collection for *D. alata*, composed of 178 fully grown adult trees, located at the Agronomy School of the Federal University of Goiás. These 178 individuals came from several locations in the state of Goiás, Brazil.
2. An in situ data set composed of 642 individual trees sampled in 25 local populations throughout *D. alata*'s geographic range (Fig. 4). Sample sizes within local populations ranged from 12 to 32 (see Diniz-Filho et al. (2012a) and Soares et al. (2012) for sampling methodological details).

The *D. alata* samples were genotyped for nine microsatellite loci (Bm164, DaE06, DaE12, DaE20, DaE34, DaE41, DaE63, DaE67, DaE46) (Soares et al. 2012; Diniz-Filho et al. 2012b) finding a total of 55 distinct alleles. All 55 alleles are represented among the in situ sampled trees. The ex situ germplasm collection has 40 from the 55 studied alleles.

Based on these data, we produced three matrices:

1. Matrix A: an allele-by-tree matrix for the ex situ data set. In matrix A_{txa}, $t = 178$ (UFG-AS adult trees) and $a = 55$ (alleles), a_{ij} represents the occurrence of allele j in tree i.
2. Matrix B: an allele-by-tree matrix for the in situ data. In matrix C_{txa}, $t = 642$ (sampled individual trees) and $a = 55$ (alleles), b_{ij} represents the occurrence of allele j in tree i.
3. Matrix C: an allele-by-population presence-absence matrix for the in situ data. In matrix C_{pxa}, $p = 25$ (populations) and $a = 55$ (alleles), c_{ij} represents the occurrence of allele j in population i.

In these matrices, alleles in homozygosity received lower score compared to alleles in heterozygosity. By doing this, we benefited solutions with higher content of heterozygosity.

The problem

The overall problem is to maximize the number of alleles while minimizing the number of individuals required to represent all alleles, maximizing at the same time heterozygosity. A candidate solution for the problem is a vector

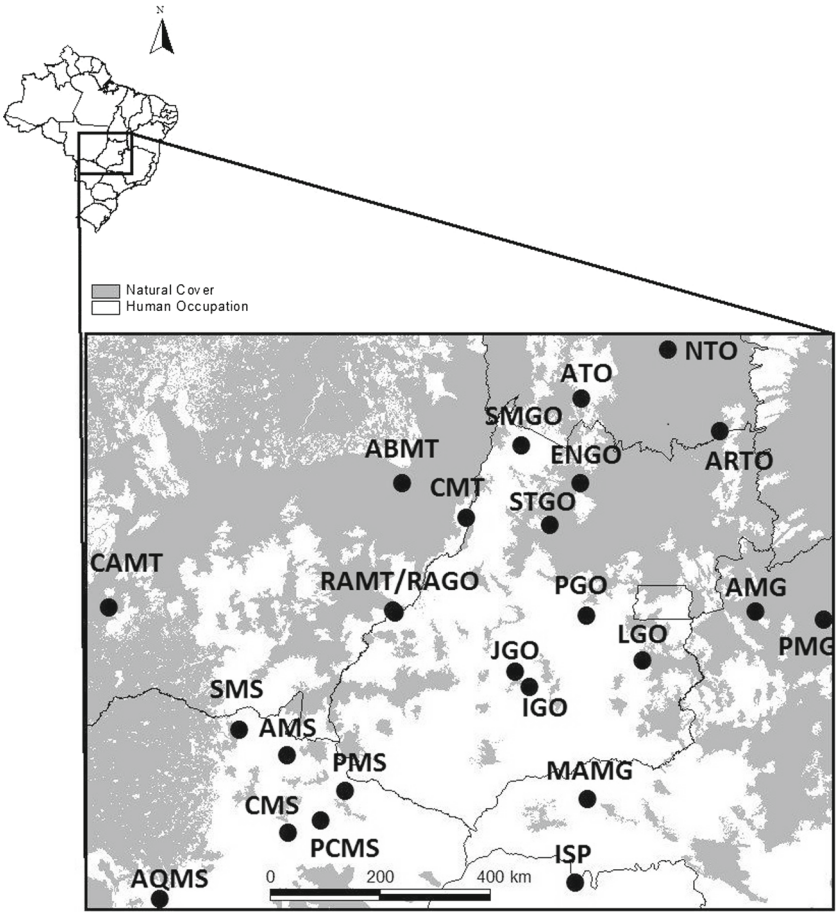


Fig. 4 Geographic location of the 25 sampled local populations of *Dipteryx alata* in Central Brazil. Regions shown in *dark tone* are still covered by natural remnants of Cerrado vegetation. It is worth to note that the studied region is comprised of small fragments or isolated individuals in a matrix of pastures and crops, notably maize and soybean. The only populations connected by continuous original vegetation are *RAGO* and *RAMT*. The names and geographical coordinates of the populations can be found in Supplementary file S1. Modified from Diniz-Filho et al. (2012b)

$\vec{x} = x_1, ..., x_k$, where k is the number of accessions (individual trees), $x_i \in \{0, 1\}$, such that $x_i = 1$, if the individual tree i is selected to compose the solution; or 0, otherwise.

The aim is to obtain:

$$\min\left(\sum_{i=1}^{k} x_i\right) \quad (1)$$

Subject to:

$$\forall j \in \{1, 2, ..., n\}, \sum_{i=1}^{k} a_{ij} x_i \geq 1 \quad (2)$$

Where $n = 55$ (alleles).

Regarding the multi-objective optimization (MOO), there are three objectives to be optimized:

1. Minimize the number of selected individuals (3);

$$\min(f_1(\vec{x})) = \min(\text{individual\_trees}(\vec{x})) \quad (3)$$

2. Maximize the number of alleles (4). For simplicity, this second objective function was transformed into an equivalent minimization objective function by using the number of lacking alleles (those that are not present in the solution);

$$\begin{aligned} \max(f_2(\vec{x})) &= \max(\text{alleles}(\vec{x})) \Leftrightarrow \min(\text{lacking\_alleles}(\vec{x})) \\ &= \min(55\text{-alleles}(\vec{x})) \end{aligned} \quad (4)$$

3. Maximize the heterozygosity (5).

$$\max(f_3(\vec{x})) = \max(\text{heterozygosity}(\vec{x})) \quad (5)$$

Next, we describe the experiments used to select individuals from the in situ sample in order to complement the germplasm collection and to select entries from the existing germplasm collection to form a core collection.

Experiment 1: Complementing the UFG-AS germplasm collection. The optimization problem defined here is to find the smallest set of *D. alata* individuals that better complement the genetic variability already preserved in the UFG-AS germplasm collection, thus, representing the genetic diversity of this species, aiming at its conservation and persistence. This problem is similar to the one stated in Schlottfeldt-Santos et al. (2012), but instead of treating populations as a unit, here we solved the problem at individual level.

As mentioned before, from the 55 studied alleles, 40 were already presented in the germplasm collection, therefore, individual trees belonging to the in situ data set should be selected in order to represent those 15 still lacking alleles.

Experiment 2: Defining the core collection (reducing the existing germplasm collection). One of the challenges of maintaining a germplasm collection is to keep the genetic variability while reducing the maintenance costs, and this is the objective of this experiment.

We have two scenarios:

1. *Scenario 1*: we addressed the problem of defining a core collection by identifying, within the UFG-AS germplasm collection, the minimal set of individuals needed to represent all the genetic variability exhibited by the germplasm collection.

   For this first scenario, we optimized the three objectives stated in Eqs. 3 to 5 on the ex situ data set.
2. *Scenario 2*: considering that maybe within the universe of all sampled trees (ex situ and in situ individual trees) it could be found an even smaller core collection, we performed the optimization joining information from the ex situ data set (178 trees) and the in situ data set (642 trees), in a total of 820 individual trees.

   For this scenario, we tested two approaches in order to define a protocol to deal with this optimization in a more efficient way, aiming the attainment of better results (a smaller, still representative, core collection):

   (a) *Method 1*: to perform the MOO directly on the 820 individual tree data set.

   (b) *Method 2*:

   Step 1: to perform the MOO in population level, by considering the 25 in situ data populations and assuming that the germplasm collection was the 26th population. This step identifies the smallest set of populations needed to represent all 55 alleles.

   Step 2: considering only individual trees belonging to populations selected in step 1, to perform a new MOO.

The null model. We executed a null model in order to verify if the same results would be found without the use of MOO.

Results

Table 1 summarizes the obtained results. It is worth noting that here, we are dealing with solutions having a minimum set of individuals, but since there was more than one optimized objective, we obtained a larger portfolio of solutions that are equally good in the sense that none of them is better when all the objectives are simultaneously considered (they are all Pareto optimal).

We performed experiment 1 aiming at complementing the UFG-AS germplasm collection. We found that the

Table 1 Sets of selected populations identified by simultaneously optimizing three objectives, where the first objective is the minimization of selected population/individuals set, the second objective is the minimization of lacking alleles, and the third objective is the maximization of heterozygosity

					Results	
		Data source	Number of accessions	Number of alleles	Core collection	Pop. minimum
Experiment 1		In situ data set	642	15	9	8
	Scenario 1	Ex situ data set	178	40	18	-
	Scenario 2	(In situ + ex situ)	820	55	42	20
Experiment 2	Method 1	data set				
	Scenario 2	Selected pop. from	336	55	20	6
	Method 2	(In situ + ex situ)	(in 8 pop.)			
		data set				

The column "Core collection" shows the minimal quantity of individuals found as solution to represent all the diversity (all the alleles) indicated in the experiment. The column "Pop. minimum" shows the number of populations to which the individuals found as minimum solution belong

smallest set of individuals needed to complement the UFG-AS germplasm collection has nine individuals selected from eight populations (AMS, ABMT, CAMT, CMT, MAMG, PMS, RAMT, SMS) (Fig. 5).

In experiment 2, we had the objective of finding a core collection for the UFG-AS germplasm collection, trying to reduce the number of individuals (and therefore, the maintenance costs), but preserving the overall diversity. For scenario 1, we found that 18 individuals were enough to represent all 40 alleles from the UFG-AS germplasm collection

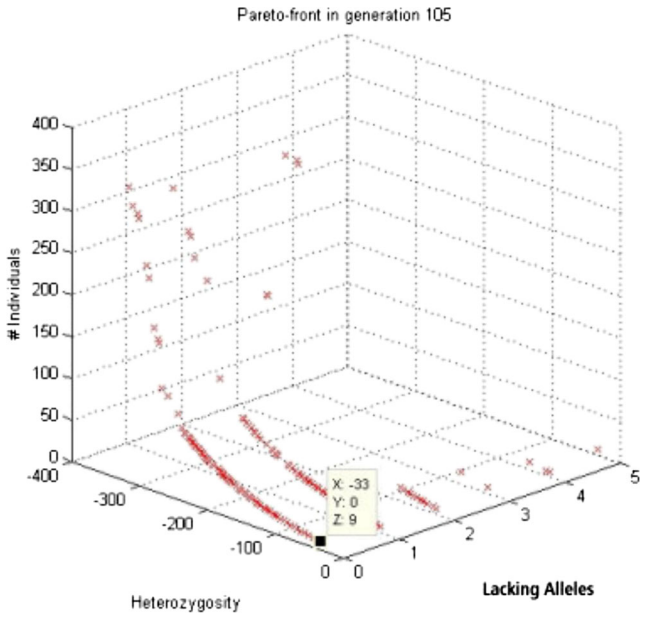


Fig. 5 Pareto front for experiment 1 (complementing the UFG-AS germplasm collection with selected individuals from the in situ data set)

(Fig. 6). For scenario 2 - method 1, we found that 42 individuals were enough to represent all 55 alleles taking into account the 820 individuals trees from joined in situ and ex situ data set (Fig. 7a). Finally, for scenario 2 - method 2, we found, in step 1, that the smallest set of *populations* needed to represent all 55 alleles was 5, but among all the obtained solutions, from the initial universe of 26 populations, 8 appeared in at least one solution with no lacking alleles (AMS, CAMT, CMT, ENGO, PMS, RAMT, STGO, UFG-AS). In step 2, MOO was performed considering only individual trees belonging to those eight populations. In the end of the process, we found that the smallest set of individuals needed to represent all 55 alleles is composed of 20 individuals (within six populations: AMS, CAMT, CMT, PMS, RAMT, UFG-AS) (Fig. 7b).

Finally, we generated a *null model* as follows. We randomly produced the same amount of solution generated for each previously described experiment. For each randomly generated solution, we computed values for heterozygosity, number of lacking alleles, and number of populations/individuals, following the computations performed for the MOO. For all the randomly generated solutions, the comparison values were worse than solutions found with MOO, as well as the number of lacking alleles (meaning that less alleles were presented in the null model solutions).

Discussion

The proposed MOO approach based on principles of SCP incorporating heterozygosity information can be used to help construct collections with maximal allelic richness.

The most important contribution of this work is to identify, in the context of SCP, within a population of several

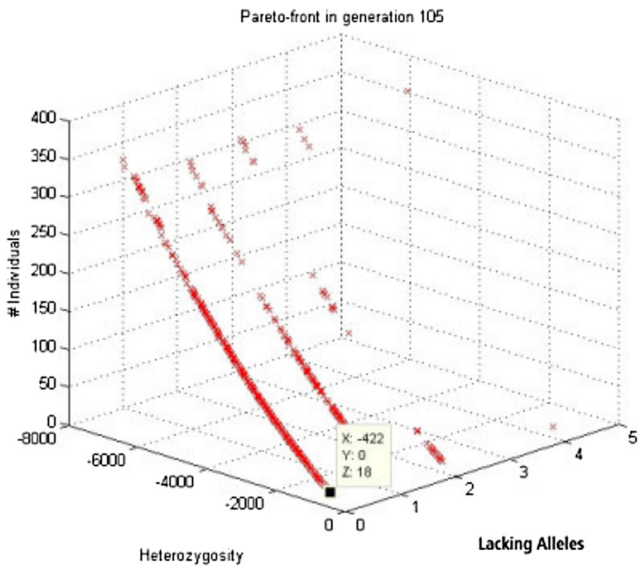
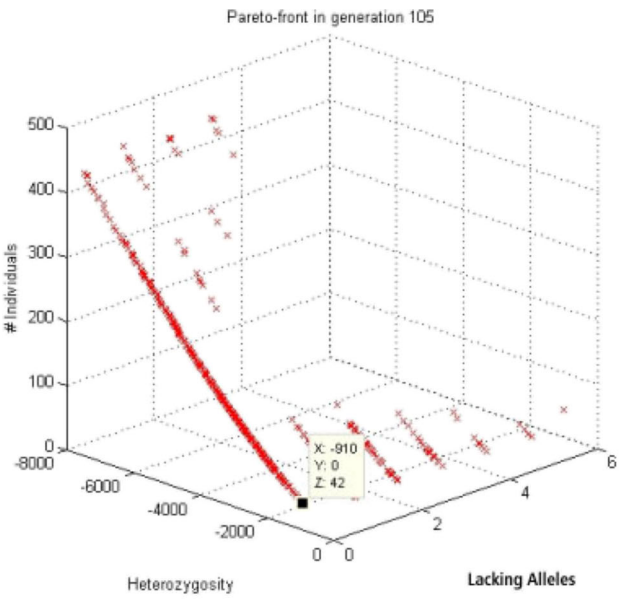
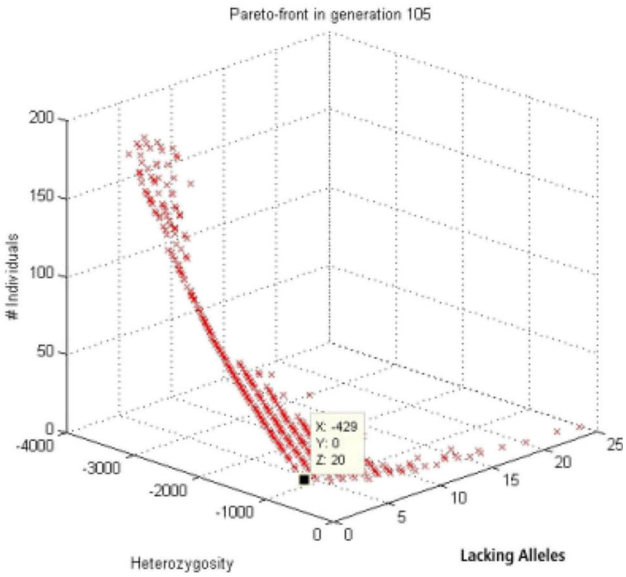


Fig. 6 Pareto front for experiment 2 - scenario 1 (defining the minimal set of individuals needed to represent all the genetic variability exhibited within the UFG-AS germplasm collection (ex situ data set))



(a) Method 1: performing MOO directly on all 26 populations (*in-situ* + *ex-situ* data set = 820 individual trees).



(b) Method 2: using previously selected populations.

Fig. 7 Pareto front for experiment 2 - scenario 2 (ex situ and in situ data set altogether)

individuals, the exact samples that should be chosen in order to preserve the species diversity.

This approach can be extended to the in situ conservation. Previous approaches generally indicate a population to be preserved; the proposed method indicates exactly which individuals within the population should be sampled/kept.

Even if the aim is to obtain a minimum set, this method identifies a portfolio of solutions, indicating sets with more individuals that better fulfill the stated objectives, providing decision makers with more options for achieving their conservation targets. These sets are optimal in the sense that none of them can be considered the best when all the objectives are simultaneously considered.

In the context of the case study carried out, most *D. alata* diversity is found only in nature, and many such populations are increasingly threatened by habitat reduction. In nature, there are many potential useful populations, yet, for practical purposes, only a fraction of this material can be afforded protection or maintenance in germplasm collections or in protected areas. Additionally, individual trees are often geographically wide ranging, making it costly to collect representative samples of material.

Experiment 1 results show that samples of only nine ex situ individual trees are necessary to complement the existing UFG-AS germplasm collection, completing the already represented 40 alleles with the 15 lacking ones.

Moreover, using the proposed approach, it is possible to include additional optimization objectives, such as the distance from in situ individual trees to the UFG-AS, reducing, therefore, displacement costs associated with collection of samples for complementing the germplasm collection.

Experiment 2 - scenario 1 found that from the 178 trees in the UFG-AS germplasm collection, it is possible to preserve the allele diversity (40 alleles) by keeping a core collection of 18 individual trees (only 10 % from the current germplasm collection). Considering that the lifespan of a *D. alata* tree is 60 years, the proposed method is important to

define strategies to provide a set of genetically diverse material while selecting the most representative individual trees. By maximizing genetic diversity in germplasm collections, resources available for conservation of biodiversity can be allocated to a larger number of species.

Associating results of experiment 1 (minimum set of 9 individual trees) and experiment 2 - scenario 1 (minimum set of 18 individual trees), we obtained a core collection of 27 individual trees ($9 + 18 = 27$) representing the entire allelic richness of the nine coded microsatellite *loci*.

Considering that within the universe of all the available data (ex situ and in situ data sets, e.g., 820 individual trees) a smaller core collection could be found, we executed experiment 2 - scenario 2 in order to verify this hypothesis. Experiment 2 - scenario 2 - method 1 found a minimum set of 42 individual trees representing all the allelic richness. Likely, this result was not better because we have done the same number of executions performed in the previous experiments, but with a bigger input (820 individual trees). As an alternative to overcome this issue, we proposed a refining method that initially selected the main populations integrating the optimum solutions, and within these populations we looked for the specific individuals that would compose our minimum set. By doing so, experiment 2 - scenario 2 - method 2 was able to find a core collection with only 20 individual trees representing all 55 alleles; less than the numbers obtained for experiment 2 - scenario 2 - method 1 (42 individual trees) and experiment 1 associated with experiment 2 - scenario 1 (27 individual trees), suggesting that this is an adequate method to be applied.

We verified that, for all the experiments, there was a significant improvement in the retention of alleles found in selected accessions.

Results from the null model assured that solutions obtained with MOO have not emerged randomly.

An effective genetic strategy should maximize the retention of genetic variation associated with long-term species survival, yet it is not easy to assess the level and properties of such variation, not to mention which traits might become important in future.

The new proposed approach using MOO and SCP for the conservation of genetic resources based on individual molecular variability helps design ex situ conservation strategies, and guide in situ sampling for germplasm core collections aiming for effective conservation and future species utilization.

Conclusion

Genetic diversity is basic for meaningful and effective conservation for the species-specific traits. This paper showed how principles of SCP and the MOO approach associated to microsatellite loci information can be applied to successfully help construct germplasm collections having maximum allelic richness and minimum number of accessions.

Having established that the approach is viable, our future work will focus on applying this approach to other kinds of genomic information (e.g., single nucleotide polymorphism – SNP) in order to verify its feasibility for this kind of data.

Acknowledgments SS acknowledges the support from CNPq, through the Science without Borders Program. The research program in molecular ecology and conservation of Cerrado plants has been continuously supported by several grants and fellowships to the research network GENPAC (Geographical Genetics and Regional Planning for Natural Resources in Brazilian Cerrado) from CNPq/MCT/CAPES (projects#564717/2010-0 and #563624/2010-8) and by the "Núcleo de Excelência em Genética e Conservação de Espécies do Cerrado"-GECER (PRONEX/FAPEG/CNPq CP 07-2009). Field work has been supported by Systema Naturae Consultoria Ambiental LTDA and work by MEMTW, ACPLFC, TNS, MPCT, RDL, and JAFDF have been continuously supported by productivity fellowships from CNPq.

Data Archiving Statement

Raw data available from the Dryad Digital Repository: http://doi.org/10.5061/dryad.hq8pd. (Dryad informs that we must wait until our manuscript has been accepted before submitting data to the repository). Supplementary_File_S1 (geografic_location): Geographic location of *D. alata* populations. Supplementary_File_S2 (ex-situ_data): Excel file containing *D. alata ex-situ* genotype data (UFG-AS germplasm collection). Individuals are identified in each row. Loci are designated by column with the nine microsatellite listed. Supplementary_File_S3 (in-situ_data): Excel file containing *D. alata in-situ* genotype data. Individuals are identified by population in each row. Loci are designated by column with the nine microsatellite listed. (?) refers to missing data.

References

Belaj A, del C Dominguez-Garcia M, Atienza SG, Urdiroz NM, de la Rosa R, Satovic Z, Martin A, Kilian A, Trujillo I, Valpuesta V, Del Rio C (2012) Developing a core collection of olive (Olea europaea L.) based on molecular markers (DArTs, SSRs, SNPs) and agronomic traits. Tree Genet Genomes 8(2):365–378. doi:10.1007/s11295-011-0447-6

Brockhoff D (2009) Theoretical aspects of evolutionary multiobjective optimization—a review. Rapport de Recherche RR-7030, INRIA Saclay—Île-de-France

Brown A (1989) Core collections: a practical approach to genetic resources management. Genome 31(2):818–824

Coello Coello CA (2001) A short tutorial on evolutionary multiobjective optimization. In: Proceedings of the first international conference on evolutionary multi-criterion optimization. Springer-Verlag, London. EMO '01, pp 21–40. http://dl.acm.org/citation.cfm?id=647889.736510

Coello Coello CA, Lamont GB, Van Veldhuizen DA (2007) Evolutionary algorithms for solving multi-objective problems, 2nd edn. Springer-Verlag, New York. ISBN 978-0-387-33254-3

Collevatti R, Telles M, Nabout J, Chaves L, Soares T (2013) Demographic history and the low genetic diversity in *Dipteryx alata* (Fabaceae) from Brazilian Neotropical Savannas. Heredity doi:10.5061/dryad.1cd80

Cormen TH, Leiserson CE, Rivest RL, Stein C (2001) Introduction to algorithms, 2nd edn. The MIT Press

Correa GC, Naves RV, Rocha MR, Chaves LJ, Borges JD (2008) Physical determinations in fruit and seeds of Baru (*Dipteryx alata* Vog.), Cajuzinho (*Anacardium othonianum* Rizz.) and Pequi (*Caryocar brasiliense* Camb.), aiming genetic breeding. Biosci J 24(4):42–47

Dawson I, Were J (1997) Collecting germplasm from trees—some guidelines. Agrofor Today 9(2):6–9

Deb K (2008) Multiobjective optimization, Springer-Verlag, chap introduction to evolutionary multiobjective optimization, pp 59–96

Diniz-Filho J, Telles M (2006) Optimization procedures for establishing reserve networks for biodiversity conservation taking into account population genetic structure. Genet Mol Biol 29:207–214

Diniz-Filho J, Bini LM, Vieira C, Blamires D, Terribile L, Bastos R, Oliveira G, Souza B (2008) Spatial patterns of terrestrial vertebrate species richness in the Brazilian Cerrado. Zool Stud 47(2):146–157

Diniz-Filho JAF, Telles MPC (2002) Spatial autocorrelation analysis and the identification of operational units for conservation in continuous populations. Conserv Biol 16(4):924–935

Diniz-Filho JAF, Collevatti RG, Chaves LJ, Soares TN, Nabout JC, Rangel TF, Melo DB, Lima JS, Telles MPC (2012a) Geographic shifts in climatically suitable areas and loss of genetic variability in *Dipteryx alata* ("Baru" Tree; Fabaceae). Genet Mol Biol 11(2):1618–1626

Diniz-Filho JAF, Melo DB, de Oliveira G, Collevatti RG, Soares TN, Nabout JC, de S Lima J, Dobrovolski R, Chaves LJ, Naves RV, Loyola RD, de C Telles MP (2012b) Planning for optimal conservation geographical genetic variability within species. Conserv Genet 13:1085–1093

Engels J, Visser L, International Plant Genetic Resources Institute (2003) A guide to effective management of germplasm collections. IPGRI handbooks for genebanks, international plant genetic resources institute

Fonseca CM, Fleming PJ (1995) An overview of evolutionary algorithms in multiobjective optimization. IEEE T Evolut Comput 3:1–16

Frankel OH (1984) Genetic manipulation: impact on man and society, Cambridge University Press, chap Genetic perspectives of germplasm conservation, pp 161–170

Gupta A, Sanjeev S, Vikas C, B GS, Riya S, Neeru J (2002) Cost of conservation of agrobiodiversity. IIMA Working Papers WP2002-05-03, Indian Institute of Management Ahmedabad, Research and Publication Department, http://EconPapers.repec.org/RePEc:iim:iimawp:wp00015

Holbrook CC, Anderson WF, Pittman RN (1993) Selection of core collection from the U.S. Germplasm collection of peanut. Crop Sci 33(4):859–861

Laghetti G, Pignone D, Sonnante G (2008) Statistical approaches to analyze Gene Bank data using a lentil germplasm collection as a case study. Agric Conspec Sci 73(3):175–181

Li CT, Shi CH, Wu JG, Xu HM, Zhang HZ, Ren YL (2004) Methods of developing core collections based on the predicted genotypic value of rice (Oryza sativa L.) Theor Appl Genet 108(6):1172–1176. doi:10.1007/s00122-003-1536-1

Margules CR, Pressey RL (2000) Systematic conservation planning. Nature 405(6783):243–253. doi:10.1038/35012251

Mei Y, Zhou J, Xu H, Zhu S (2012) Development of sea island cotton (*Gossypium barbadense* L.) core collection using genotypic values. Aust J Crop Sci 6(4):673–680

Myers N, Mittermeier RA, Mittermeier CG, da Fonseca GAB, Kent J (2000) Biodiversity hot spots for conservation priorities. Nature 403:853–858

Rao ES, Kadirvel P, Symonds RC, Geethanjali S, Ebert AW (2012) Using SSR markers to map genetic diversity and population structure of *Solanum pimpinellifolium* for development of a core collection. Plant Genet Resour 10:38–48. doi:10.1017/S1479262111000955

Rao NK, Bonwoo K, Sastry DVSSR (2002) Pod and seed storage: cost-benefit study for groundnut germplasm conservation. International Arachis Newsletter 22:10–12

Rao NK, Hanson J, Dulloo ME, Ghosh K, Nowell A, Larinde M (2006) Manual of seed handling in genebanks. IPGRI Handbooks for Genebanks, Renouf Publishing Company Limited

Roederer C, Nugent R, Wilson P (2000) Economic impacts of genetically modified crops on the agri-food sector: a synthesis, working document, vol 1. Directorate-General for Agriculture, European Commission, http://ec.europa.eu/agriculture/publi/gmo/gmo.pdf

Schlottfeldt S, Timmis J, Walter MEMT, Carvalho ACPLF, Diniz-Filho JAF, Simon LM, Loyola RD, Telles MPC (2014) Multiobjective optimization applied to systematic conservation planning and spatial conservation priorities under climate change. In: Proceedings of the 2014 conference companion on genetic and evolutionary computation companion. GECCO Comp'14, pp 177–178, doi:10.1145/2598394.2598404

Schlottfeldt-Santos S, Walter MEMT, Diniz-Filho JAF, de C Telles MP (2012) Multiobjective optimization in systematic conservation planning to represent genetic variability within species. In: Proceedings of the 8th International Conference on Ecological Informatics

Soares TN, Melo DB, Resende LV, Vianello RP, Chaves LJ, Collevatti RG, Telles MPC (2012) Development of microsatellite markers for the neotropical tree species *Dipteryx alata* (Fabaceae). Am J Bot 99:e72–e73. doi:10.3732/ajb.1100377

Wang Y, Zhang J, Sun H, Ning N, Yang L (2011) Construction and evaluation of a primary core collection of apricot germplasm in China. Sci Hort 128:311–319. doi:10.1016/j.scienta.2011.01.025

Zhang P, Li J, Li X, Liu X, Zhao X, Lu Y (2011) Population structure and genetic diversity in a rice core collection (*Oryza sativa* L.) Investigated with SSR Markers. PLoS One 6(12):e27,565+ doi:10.1371/journal.pone.0027565

Zitzler E (2002) Evolutionary algorithms for multiobjective optimization. In: Evolutionary methods for design, optimisation, and control (EUROGEN 2001), CIMNE, pp 19–26

8.5 *Paper 4*

Schlottfeldt, S; Timmis, J.; Walter, M.E.M.T; de Carvalho, A.C.P.L.F.; Diniz-Filho, J.A.F.; Simon, L.M.; Loyola, R.D.; Telles, M.P.C. Multi-objective Optimization Applied to Systematic Conservation Planning and Spatial Conservation Priorities Under Climate Change. In Proceedings of the 2014 Conference Companion on Genetic and Evolutionary Computation Companion, GECCO Comp '14, pages 177–178, Vancouver, BC, Canada, 2014. ACM. ISBN 978-1-4503-2881-4.

Multi-Objective Optimization Applied to Systematic Conservation Planning and Spatial Conservation Priorities under Climate Change

Shana Schlottfeldt
Dep. of Computer Science
University of Brasilia
Brasilia, DF, Brazil
shanass@unb.br

Jon Timmis
Department of Electronics
University of York
York, UK
jon.timmis@york.ac.uk

Maria Emilia M. T. Walter
Dep. of Computer Science
University of Brasilia
Brasilia, DF, Brazil
mia@cic.unb.br

Andre C.P.L.F. Carvalho
Dep. of Computer Science
University of São Paulo
São Carlos, SP, Brazil
andre@icmc.usp.br

Jose Alexandre F. Diniz-Filho
Department de Ecology, ICB
Federal University of Goiás
Goiânia, GO, Brazil
diniz@icb.ufg.br

Lorena M. Simon
Institute of Biological Sciences
Federal University of Goiás
Goiânia, GO, Brazil
lores_bio@hotmail.com

ABSTRACT

Biodiversity problems require strategies to accomplish specific conservation goals. An underlying principle of these strategies is known as Systematic Conservation Planning (SCP). SCP is an inherently multi-objective (MO) problem but, in the literature, it has been usually dealt with a monobjective approach. In addition, SCP analysis tend to assume that conserved biodiversity does not change throughout time. In this paper we propose a MO approach to the SCP problem which increases flexibility through the inclusion of more objectives, which whilst increasing the complexity, significantly augments the amount of information used to provide users with an improved decision support system. We employed ensemble forecasting approach, enriching our analysis by taking into account future climate simulations to estimate species occurrence projected to 2080. Our approach is able to identify sites of high priority for conservation, regions with high risk of investment and sites that may become attractive options in the future. As far as we know, this is the first attempt to apply MO algorithms to a SCP problem associated to climate forecasting, in a dynamic spatial prioritization analysis for biodiversity conservation.

Categories and Subject Descriptors

J.2 [**Computer Applications**]: Physical Sciences and Engineering—*Earth and atmospheric sciences*

GECCO'14, July 12–16, 2014, Vancouver, BC, Canada
ACM 978-1-4503-2881-4/14/07
http://dx.doi.org/10.1145/2598394.2598404.

Keywords

multi-objective optimization; systematic conservation planning; biodiversity conservation; climate change.

1. INTRODUCTION

The growing interest and concern regarding biodiversity demand strategies to target conservation goals. These strategies include the Systematic Conservation Planning (SCP), which determines the most cost effective way of investing in conservation actions.

Computationally speaking, SCP is formalized by the well-known NP-hard Set-Covering Problem [1]. In a simplified way, SCP is the problem of finding a minimum set of sites maximizing at the same time the characters under study. There are clearly two conflicting objectives to be optimized, which makes SCP a natural candidate for Multi-Objective Optimization (MOO). Furthermore, several other parameters, e.g., social and political objectives, can be incorporated to SCP, adding further dimensions to the problem, therefore increasing its complexity.

Although SCP is inherently multi-objective, it is frequently dealt with using a monobjective approach, assigning weights to the problem dimensions, in order to obtain an unique objective function [7]. Two main reasons justifies the use of MOO when dealing with the SCP problem; first, it is possible to find a set of solutions to the problem instead of a single one; and second, there is an increase in flexibility of both data type and problem constraints, at the same time that the problem is kept tractable [2].

Typically, SCP analyses are static, i.e., they assume that biodiversity does not change over time [6]. However, scientific evidences urge to incorporate climate change analysis into conservation plans [5].

In this paper, we propose a more sophisticated, yet general, solution to the SCP problem using MOO. This approach increases flexibility by including more decision objectives, which whilst increasing the complexity, significantly augments the amount of information used to provide users with an improved decision support system.

As far as we know, this is the first attempt to apply MOO to SCP associated to climate forecasting, in a dynamic spatial prioritization analysis for biodiversity conservation.

2. MATERIALS AND METHODS

2.1 Data

Plant species. We used data of occurrence of 96 plants with economic importance in Cerrado, a large biome in Central Brazil, in which many endemic and rare species are under high threat levels or extinction [4].

Climate Forecast. We used an ensemble forecast approach, obtaining the likely distribution of species in the considered region by 2080 [8].

Additional Objectives. *Annual Actual Evapotranspiration (AET):* a measure of the joint availability of energy and water in the environment. *Human Occupancy (H_O):* a measure obtained compiling data on social and economic variables indicating conservation conflicts. *Vegetation Remnants (VR):* the proportion of each grid cell covered by natural vegetation, based on remote sense information.

Conservation scenarios. For present and future data, we have a presence-absence matrix $A_{m \times n}$, $m = 181$ *sites* and $n = 96$ *plant species*. We have *five different objectives*: 1) minimize the number of sites; 2) maximize the number of represented plant species; 3) maximize AET; 4) minimize H_O; 5) maximize VR.

Our fitness functions were developed by having as many equations as objectives to be optimized. This allowed to simultaneously optimize distinct objectives instead of aggregating variables into one single function.

We defined three conservation scenarios: *Scenario 1*: to represent all species in current time, applying optimization in 2 dimensions (optimizing objectives 1 and 2); *Scenario 2*: to represent all species in current time, using optimization in 5 dimensions (i.e., optimizing simultaneously objectives 1 to 5); and *Scenario 3*: to represent all species in 2080 (since it happens to be a forecast, objectives 3 to 5 are not available, and optimization was performed considering objectives 1 and 2).

2.2 Experimental Setup

Algorithm and computer infrastructure. We performed 19,120 individual runs for each scenario previously described. For each run, a population of 500 initial solutions was randomly generated. These solutions were then evolved using Non-Dominated Sorting Genetic Algorithm-II (NSGA-II) [3], implemented in Matlab®. The experiments were performed on two servers, a HP ProLiant DL585 G7, 4xAMD 2.8Ghz 16-cores, 512GB RAM, and a HP ProLiant DL385p Gen8, 2xAMD 2.8Ghz 16-cores, 256GB RAM

Evaluation metric. Due to the algorithm stochasticity, we used the *selection frequency* metric (SF) [6], which represents the number of times each site is selected in the solutions to the overall problem. Once the SF to all cells was calculated, the grid cells were ranked based on their score. We ordered all the grid cells in a bi-dimensional plot showing the relative importance of each cell both to the current time and to 2080 (importance axis). This graph epitomizes the scheme for dynamic spatial prioritization analyses for biodiversity conservation [6].

3. RESULTS AND DISCUSSION

We found that optmization using additional objectives (Scenario 2) allowed us to supply decision makers (DM) with a more diversified portfolio of sets of sites to be conserved.

Furthermore, our method was able to identify (if they exist) sites of high priority for conservation, regions with high risk of investment and sites that may become attractive options in the future. In this context, this study focused on showing that predicted climate change could cause shifts on the distribution patterns of economically important plants, and that these data could be used in order to help DM to select their schemes of conservation. Supported by scientific data, DM can examine options made available to current time related to the future, and decide how to define their spatial conservation priorities, reviewing them if necessary.

4. CONCLUSIONS

Our results show the advantages of the new approach (MOO) with respect to previous solutions (monobjective). Moreover, the associated climate forecast showed that having a picture of how future scenarios will look like can be extremely useful for DM.

Next step is to develop a more specialized MOO algorithm for the SCP problem. Using artificial immune system, we intend to improve the work presented in this paper.

5. ACKNOWLEDGMENT

ACPLFC, JAFDF, MEMTW, MPCT, RDL and SSS have been continuously supported by CNPq. JT is part funded by The Royal Society.

6. ADDITIONAL AUTHORS

Rafael D. Loyola (Federal University of Goiás–UFG, email: diasloyola@gmail.com) and Mariana P.C. Telles (UFG, email: tellesmpc@gmail.com).

7. REFERENCES

[1] M. Cabeza and A. Moilanen. Design of reserve networks and the persistence of biodiversity. *Trends Ecol Evol*, 16(5):242–248, May 2001.

[2] C. A. Coello Coello at al. *Evolutionary Algorithms for Solving Multi-Objective Problems*. Springer-Verlag, New York, 2nd edition, Sept. 2007. ISBN 978-0-387-33254-3.

[3] K. Deb at al. A Fast Elitist Non-Dominated Sorting Genetic Algorithm for Multi-Objective Optimization: NSGA-II. *Trans. Evol. Comp*, 6(2):182–197, Apr. 2002.

[4] J. A. F. Diniz-Filho et al. Spatial Patterns of Terrestrial Vertebrate Species Richness in the Brazilian Cerrado. *Zool Stud*, 47(2):146–157, 2008.

[5] C. R. Groves et al. Incorporating climate change into systematic conservation planning. *Biodivers Conserv*, 21:1651–1671, 2012.

[6] R. D. Loyola et al. A straightforward conceptual approach for evaluating spatial conservation priorities under climate change. *Biodivers Conserv*, 22:483–495, 2013.

[7] S. Schlottfeldt-Santos at al. Multiobjective Optimization in Systematic Conservation Planning to Represent Genetic Variability within Species. In *Proceedings of the 8th International Conference on Ecological Informatics.*, 2012.

[8] L. M. Simon et al;. Effects of Global Climate Changes on Geographical Distribution Patterns of Economically Important Plant Species in Cerrado. *Revista Árvore*, 37(2):267–274, 2013.

8.6 *Paper 5*

Schlottfeldt, S; Timmis, J.; Walter, M.E.M.T.; de Carvalho, A.C.P.L.F.; Diniz-Filho, J.A.F.; Simon, L.M.; Loyola, R.D.; Telles, M.P.C. A Multi-Objective Optimization Approach Associated to Climate Change Scenario to Improve Systematic Conservation Planning and Spatial Conservation Priorities Setting. In Proceedings of the 8th International Conference on Evolutionary Multi-Criterion Optimization, EMO'2015, pages 458–472, Guimarães, Portugal, 2015. Part II. Lect Notes Comput Sc. LNCS 9019. 2015. ISBN 978-3-319-15891-4 (versão impressa), 978-3-319-15892-1 (versão eletrônica).

A Multi-objective Optimization Approach Associated to Climate Change Analysis to Improve Systematic Conservation Planning

Shana Schlottfeldt[1,2(✉)], Jon Timmis[2], Maria Emilia Walter[1], André Carvalho[3], Lorena Simon[4], Rafael Loyola[4], and José Alexandre Diniz-Filho[4]

[1] Department of Computer Science, University of Brasilia, Brasilia, Brazil
shanass@unb.br
[2] Department of Electronics, University of York, York, UK
[3] Department of Computer Science, SCC-ICMC-USP, São Paulo, Brazil
[4] Institute of Biological Sciences, Federal University of Goiás, Goiânia, Brazil

Abstract. Biodiversity conservation has been since long an academic community concern, leading scientists to propose strategies to effectively meet conservation goals. In particular, Systematic Conservation Planning (SCP) aims to determine the most cost effective way of investing in conservation actions. SCP can be formalized by the Set-Covering Problem, which is NP-hard. SCP is inherently multi-objective, although it has been usually treated with a monobjective and static approach. Here, we propose a multi-objective solution for SCP, increasing its flexibility and complexity, and, at the same time, augmenting the quality of provided information, which reinforces decision-making. We used ensemble forecasting, considering future climate simulations to estimate species occurrence projected to 2080. Our method identifies sites: 1) of high priority for conservation; 2) with significant risk of investment; and, 3) that may become attractive in the future. To the best of our knowledge, this application to a real-world problem in ecology is the first attempt to apply multi-objective optimization to SCP associated to climate forecasting, in a dynamic spatial prioritization analysis for biodiversity conservation.

Keywords: Multi-objective optimization · Systematic conservation planning · Spatial conservation prioritization · Biodiversity conservation · Climate change · Uncertainty in simulations · Parameter tuning

1 Introduction

Effective conservation of biodiversity is essential for continued human well-being, and has been since long a concern of the academic community, but only in recent years it has been faced as a political, economic and social affair [20]. In this context, the growing interest and concern regarding biodiversity is leading scientists

A. Gaspar-Cunha et al. (Eds.): EMO 2015, Part II, LNCS 9019, pp. 458–472, 2015.
DOI: 10.1007/978-3-319-15892-1_31

to develop effective strategies to meet conservation goals. The underlying principle of these strategies lies on the Systematic Conservation Planning (SCP), which determines the most cost effective way of investing in conservation actions.

SCP can be formalized by the Set-Covering Problem [5], which is NP-hard [9]. SCP can be enunciated as the problem of finding a minimum set of sites (among several available ones), simultaneously maximizing the other features under study. Thus, there are at least two conflicting objectives to be optimized, making SCP a natural candidate for Multi-Objective Optimization (MOO).

Several parameters, e.g., vegetation remnants, annual actual evapotranspiration (AET), and human occupation, among other environmental, social, and political objectives, can be incorporated to SCP, adding more dimensions to the problem, therefore increasing its complexity.

Albeit inherently multi-objective, SCP has been usually dealt with a monobjective approach through the assignment of weights to dimensions of the problem aiming to obtain a unique objective function [2,3,7,8,19,22]. Moreover, the most known techniques for SCP are static, implicitly adopting the hypothesis that conserved biodiversity does not change throughout time [17]. However, this is not really accurate, and climate change analyses should be incorporated into conservation plans to more properly reflect the biodiversity dynamics [15].

Quite a few reasons justify the use of the multi-objective approach to deal with SCP. First, a set of solutions can be found, instead of just one, and this can be of great interest to decision makers. In addition, flexibility of data type is increased and constraints can be integrated, at the same time that the problem is kept tractable [9].

In this paper, we propose a MOO approach for SCP, which significantly augments the amount and quality of information provided to users, reinforcing decision-making. We employ the well known NSGA-II, given the wide success of the algorithm, this seemed a logical place to start before developing more sophisticated multi-objective approaches.

To the best of our knowledge, this application to a real-world problem in ecology is the first attempt to apply MOO to SCP associated to climate forecasting, in a dynamic spatial prioritization analysis for biodiversity conservation. In particular, our analysis considered future climate simulations to estimate species occurrence projected to 2080. Our method suggests sites of high priority for conservation, regions with significant risks of investment and those ones that may become attractive in the future.

The remainder of the paper is structured as follows. In Section 2 we discuss the approaches previously used to deal with SCP. Section 3 describes the materials and methods adopted in this study. In Section 4, we discuss the results obtained so far. Conclusions and possible future work are presented in Section 5.

2 Previous Approaches to the SCP Problem

The SCP problem aims to minimize the number of sites, total area or cost and at the same time guarantee the representation of natural features (objects of

conservation) [25]. In order to achieve this, the problem can been formulated as follows [5]:

Let $A_{m \times n}$ be a matrix where $m = sites$ and $n = natural\ features$, whose element $a_{ij} \in \{0, 1\}$, and $a_{ij} = \begin{cases} 1, & \text{if the natural feature } j \text{ occurs in the site } i; \\ 0, & \text{otherwise.} \end{cases}$

Let each site i have a cost c_i, and each feature j a desired representation level r_j. Let $x_i \in \{0, 1\}$, where $x_i = \begin{cases} 1, & \text{if the site } i \text{ is included in the solution;} \\ 0, & \text{otherwise.} \end{cases}$

The SCP problem consists in minimizing Eq. 1:

$$\sum_{i=1}^{m} c_i x_i \tag{1}$$

Subject to Eq. 2 (for all j, each feature should be represented at least r_j times):

$$\forall j \in \{1, 2, ..., n\}, \ \sum_{i=1}^{m} a_{ij} x_i \geq r_j \tag{2}$$

The development of algorithms and tools for SCP began in the 1980s [24]. Since then, several approaches have been suggested, ranging from a simple scoring system to more complex optimization techniques. Commonly, in these approaches, algorithms select complementary sites in a sequential order, until they reach the goal of representing all the species (in effect, a greedy algorithm). Alternatively, the adoption of an exact approach (which ensures the production of optimal solutions, e.g., integer linear programming) was initially discussed by Cocks and Baird, in 1989 (mentioned in [29]). However, as SCP is a NP-hard problem, even the available software packages computing exact algorithms are not able to solve some large data sets [26]. Due to these characteristics, metaheuristics are used as an alternative approach to SCP. The most widely used metaheuristics for SCP are Simulated Annealing (SPEXAN [3], SITES [22], and Marxan [2]), and the Tabu Search (ConsNet [7]). Nonetheless, as previously mentioned, these approaches have treated SCP in a monobjective way by combining the different problem objectives in one single objective function.

On many occasions, it is difficult to work exclusively with agregated values in a monobjective function. Often the subjectivity associated to such an approach can drive to distinct results for the same data set [3,19]. Furthermore, when two criteria represent distinctive value systems it can be impossible to combine and/or compare such criteria in a meaningful manner. To insist in a single objective function can lead to disparate values, conducing to inaccurate results and/or requiring assumptions that some decision makers would find inappropriate [8].

3 Materials and Methods

3.1 Data

Plant Species. We used data of occurrence of 96 plants with economic importance in Cerrado, a large biome in Central Brazil, occupying around $1,500,000$ km^2. Satellite-based estimates of habitat transformations in Cerrado show rates that are still very high and far from diminishing, which will likely put many endemic and rare species under high threat levels or extinction [14]. Besides the importance of the biome conservation, plant species used in this research have historical and cultural relevance, being widely used as part of the culture and development of regional communities [10].

Information of the 96 plant species under study were obtained from Centro de Referência em Informação Ambiental (CRIA; www.cria.org.br), from Flora Integrada da Região Centro-Oeste (Florescer; www.florescer.unb.br), from the scientific literature index in ISI (apps.isiknowledge.com) and from Scielo (www.scielo.org). A total of 8,896 points were compiled and used for modelling the 96 species. The Cerrado region was overlapped by a 181-cell grid, in which cells were 1^o of latitude by 1^o of longitude. The occurrences were modelled as a function of several environmental variables using different methods (for details see [31]), and results were combined to generate the distribution data, which was later converted into a matrix of presence-absence of species.

Climate Forecast. To evaluate the effects of future climate changes on the species geographical distribution, we used an ensemble forecast approach, a conjunction of different climate models, modelling methods and carbon emission scenarios [13], obtaining what would be the distribution of the species in the considered region by 2080 (for details see [31]).

Additional Objectives. Three additional objectives were used in this study: annual actual evapotranspiration, human occupancy and vegetation remnants.

Annual Actual Evapotranspiration (AET). A measure of the joint availability of energy and water in the environment. Information came from many databases, and our dataset was obtained according to Rodriguez et al. [28].

Human Occupancy (H_O). Human population density (H) has been often used as a criterion to be minimized [18] or as an evidence of conflicts between economic/social interests and biological conservation. Although, Rangel et al. [27] showed that, in Brazilian Cerrado, species richness was positively correlated with patterns of modern agriculture and cattle ranching, but not with human population density. Consequently, other socio-economic variables should be considered to minimize costs when establishing regional programs for conservation planning in Brazilian Cerrado. Therefore, this study considered the human occupancy (H_O), a measure obtained compiling data on social and economic variables indicating conservation conflicts [14,27]. Data was obtained from the Brazilian Institute of Geography and Statistics (IBGE; www.ibge.gov.br).

Vegetation Remnants (VR). These refer to the proportion of each 1^o grid cell covered by natural vegetation, based on remote sense information (Moderate-Resolution Imaging Spectroradiometer (MODIS)). Data used in this article are detailed described in Carvalho et al [6].

Conservation Scenarios. For present and future data, we have a presence-absence matrix $A_{m \times n}$, where $m = 181$ *sites* and $n = 96$ *plant species*. In addition, for each site over time, we have information about AET, H_O, and VR. Hence, we have five different objectives to be optimized: 1) minimize the number of sites (among the 181 grid cells); 2) maximize the number of 96 represented plant species; 3) maximize AET; 4) minimize H_O; 5) maximize VR.

Our fitness functions were developed by having as many representations of Eq. 1 as objectives to be optimized, and varying c_i according to the objective under consideration. This allowed to simultaneously optimize distinct objectives instead of aggregating them into one single function.

We worked with minimization, so, based on the duality principle, w.l.o.g., we converted all objectives to their equivalent minimization representation (e.g., for the objective mentioned in item 2, optimization consisted in minimizing the number of missing species – which is the same as maximizing the number of represented species).

The experts defined that all the species should be represented at least once, i.e. in Eq. 2, $r_j = 1$, $j \in \{1, ..., 96\}$.

We defined three conservation scenarios:

- *Scenario 1*: to represent all species in current time, applying optimization in 2 dimensions (we optimized objectives 1 and 2, respectively, the number of sites and the number of plant species);
- *Scenario 2*: to represent all species in current time, using optimization in 5 dimensions (i.e., optimizing simultaneously objectives 1 to 5); and,
- *Scenario 3*: to represent all species in 2080 (since it happens to be a forecast, objectives 3 to 5 are not available, and optimization was performed in two dimensions, considering only objectives 1 and 2).

3.2 Experimental Setup

Algorithm. We used the Non-Dominated Sorting Genetic Algorithm-II (NSGA-II) [11]. For each run, a population of initial solutions was randomly generated. These solutions were then evolved using NSGA-II, which was implemented in Matlab®.

Aleatory Uncertainties. In order to determine the number of runs required to mitigate aleatory uncertainty in the stochastic algorithm employed, we used Spartan (Simulation Parameter Analysis R Toolkit Application) [1], a package of statistical techniques designed to support the identification of which simulation results can be attributed to the dynamics of the modelled system, rather than artefacts of uncertainty or parametrisation, or simulation stochasticity. More specifically, we applied the Spartan's Technique 1 (Aleatory Uncertainty

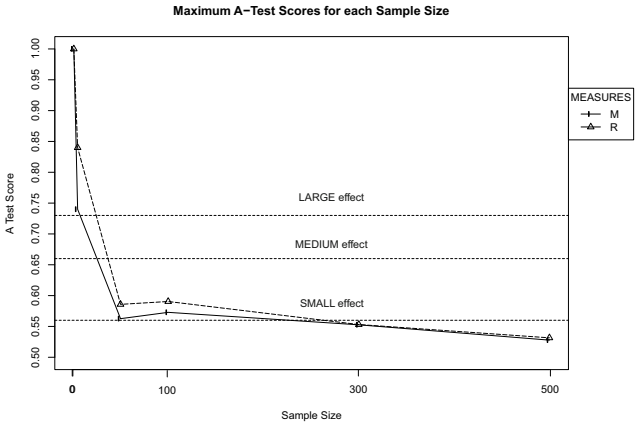
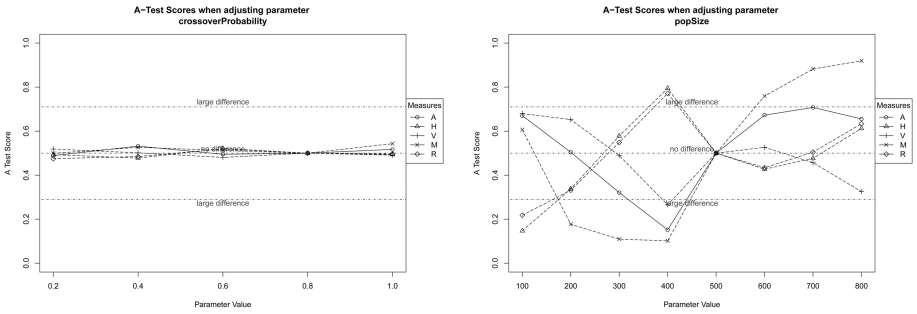


Fig. 1. Spartan's Technique 1 applied to Scenario 1. At 300 runs, stochasticity over the measures M (missing species) and R (number of selected sites) attains a small effect.

Analysis). In order to do so, we analysed 20 subsets sample sizes of 1, 5, 50, 100, 300 and 500 runs each, requiring, therefore, 19,120 individual runs for each previously described optimization scenario (a total of 57,360 individual runs). It was found that, for all the scenarios, 300 runs were sufficient to reduce the effect magnitude of aleatory uncertainty on results to less than "small" (the desired level) (Fig.1).

Parameter Settings. Almost all of the heuristic procedures involve some parameter tuning. The task of setting parameter values is notably challenging because we do not know, in advance, the impact of parameter values on the performance of the algorithm, specially when the algorithm to be tuned is stochastic in nature [32] . We used Spartan's Technique 2 (Robustness Analysis) [1] to investigate the impact of different parameter settings on the quality of the solutions, and to estimate the most suitable values for the following parameters: population size, crossover probability, mutation probability, and mutation rate. The sampling method begins at the parameters lower value and increases the value by a set increment until the upper limit is reached. Each parameter is addressed in turn, and simulation results for value assigned to that parameter analysed (Fig.2). We analysed 26 subsets sample (resulting from the combinations of the different parameter values) sizes of 300 runs each (in accordance with results obtained from Spartan's Technique 1), requiring, therefore, 7,800 individual runs for each optimization scenario (a total of 23,400 individual runs). Based on the obtained results, parameter values were set to: population size = 500; crossover probability = 0.90; mutation probability = 1/L (where L is the number of regions); mutation rate = 0.5. Besides we used: crossover operator = single point crossover (SPX); selection by binary tournament; number of objective functions evaluation = 250.000.

Computer Infrastructure. The experiments were performed on two servers running Ubuntu Linux 12.04 LTS, a HP ProLiant DL585 G7, 4xAMD Opteron



(a) (b)

Fig. 2. Spartan's Technique 2 applied to Scenario 2. The x-axis shows the range of values explored and the y-axis displays the scores obtained by contrasting response values for perturbed parameter values to calibrated values. Solutions are considered over the measures A (AET), H (human occupancy), V (vegetation remnants), M (missing species), and R (number of seleted sites). (a) Scores for different values of crossover probability, which when perturbed has no significant effect on solutions. (b) Scores for simulations varying population size, this parameter has a strong effect on the obtained solutions, and its most suitable value is 500. Results suggest that a change in the population size has a statistically significant effect on solutions, and it is more critical than the crossover probability, which has no statistically significant impact.

6386 SE 2.8Ghz 16-cores (64 physical CPU cores), 512GB RAM, and a HP ProLiant DL385p Gen8, 2xAMD Opteron 6386 SE 2.8Ghz 16-cores (32 physical CPU cores), 256GB RAM.

Evaluation Metric. Due to the stochasticity of the algorithm, we used the *selection frequency* metric (SF) [17] to compare the outcomes of our analysis. This measure represents the number of times each site is selected in the solutions to the overall problem. Once the SF to all cells was calculated, grid cells were ranked based on the result. Grid cells with the highest SF were assigned the first rank and those having SF value zero received the last rank. Next, cell relative importance in both axes was rescaled to 0–100 (zero being not important, and 100 being highly important). Then, grid cells with value zero were excluded and all the remaining grid cells ordered in a bi-dimensional plot showing the relative importance of each cell related to current time and to 2080. This graph epitomizes the scheme for dynamic spatial prioritization analyses for biodiversity conservation.

Cells with rank higher than 90 for both axes were considered high-priority. Cells ranking higher than 90 in the present, but not in 2080, are important now, but will become climatically unsuitable in the future. Cells ranking higher than 90 in 2080, but not now, will become suitable in the future. High-priority cells are those ranking higher than 90, now and in the future.

It is worth noting that we settle the lower limit rank to 90 following the literature [17], but this value is arbitrarily defined and, depending on the context, can be relaxed assuming other lower reference values (e.g., considering cells ranking higher than 50 as important, instead of higher than 90).

4 Results and Discussion

4.1 Scenario 1

The objective of the optimization in this context was to select the smallest set of sites, among the 181 available ones, capable of representing all the 96 species (the species diversity) in current time. This also allowed to establish a lower bound for Scenario 2.

We found that the minimum number of sites required to represent all of the species was 2. We found 35 distinct solutions with these characteristics, reflecting diversity in solution, which is important since it provides more options to decision makers. It is important to note that we have no hierarchy amongst results, which means that all the solutions are equal in the considered context.

A relative frequency map of the multiple solutions indicates the relative importance of a cell in order to fullfil the objectives of optimization (Fig.3). This frequency can be taken as an estimator of *irreplaceability*[1] of the cell [21], e.g., the rarest plant appears in only 10 regions in current time, these sites tend to be irreplaceables, so that if at least one of them is not selected, the conservation goal may not be achieved. One of such a site is #105, the most frequent site in solutions (associated to the presence of the rarest specie, it has the greatest diversity of species).

4.2 Scenario 2

The objective of the optimization was to select the smallest set of sites capable of representing all the 96 species in current time, but at the same time optimizing the additional objectives AET, H_O, and VR.

Although there is some empirical inferences, and correlational data in the literature, experts did not know, in principle, what to expect from the optimization in 5 dimensions, since a behaviour was not determined with respect to optimizing AET, H_O and VR simultaneously. The initial expectation was that this additional information would bring some advantage selecting sites to compose solutions, improving, therefore, the overall quality of results.

NSGA-II was not able to find (at least in the number of evaluations performed) the lower bound of 2 sites established in Scenario 1, being 3 the smallest set of sites found. This can be due to the use of a multi-objective algorithm in this scenario, when maybe the most appropriated would be to apply a many-objective

[1] A measure that indicates the proportion a cell contributes to the overall solution, e.g., cells with this measure converging to 1 often tends to be irreplaceable, in the sense that if they are lost, the conservation goal is not accomplished.

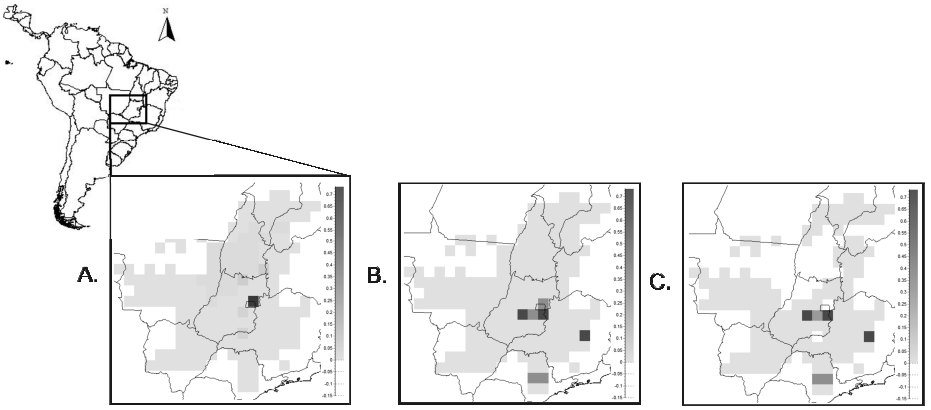


Fig. 3. Irreplaceability for: A. Scenario 1. B. Scenario 3. C. Synthesis of Scenarios 1 and 3. Irreplaceability scales from 0 to 1 (since it express the frequency a given site appears in the solutions). Cells shown in the darkest colour tend to be irreplaceable, which means that if they are lost, the conservation target (to represent all existing species) may not be achieved. Sites presenting rare species tend to have higher irreplaceability scores. Regarding the synthesis of two scenarios, irreplaceability can assume a negative value (which is plotted as value zero; white cells), this indicates that the site has lost importance (its capability to fullfil the requirements to achieve the objective).

approach [16]. Although, it can indicate that, in the context of using additional objectives, better results are obtained not through the minimum absolute possible representation, but through a trade-off between minimum representation and the other considered objectives.

The portfolio of solutions increased significantly, which was expected, since it is known in the literature that as the number of objectives increases, the number of solutions enlarge exponentially [9,16]. Thus, almost all new combination of sites will give a different result with all species being represented, so it is included in the portfolio.

Results (using Pearson correlation) confirmed the empirical conflict between H_O and VR ($r = -.84, p < .0001$), as well as between H_O and AET ($r = -.93, p < .0001$) (Fig. 4.a and Fig. 4.b), which corroborates with the evidence of *conservation conflicts* [4]. This means that H_O reflects properly the antropical effect over biodiversity by the conversion of natural habitats in antropical ones [14].

This is a strong evidence that in addition to the standard biological data used to guide planning decisions, some kind of human settlement patterns (here H_O) have to be explicitly considered from the very beginning of planning processes [18]. This is essential to reduce the conflict between population density and biodiversity and to minimize the cost of conservation (since land prices inexorably rise as human population density increases).

In addition, a positive relationship between H_O and species richness may be expected because both increase with AET [27]. This was confirmed by results obtained with a steady number of sites, where for higher values of H_O, higher

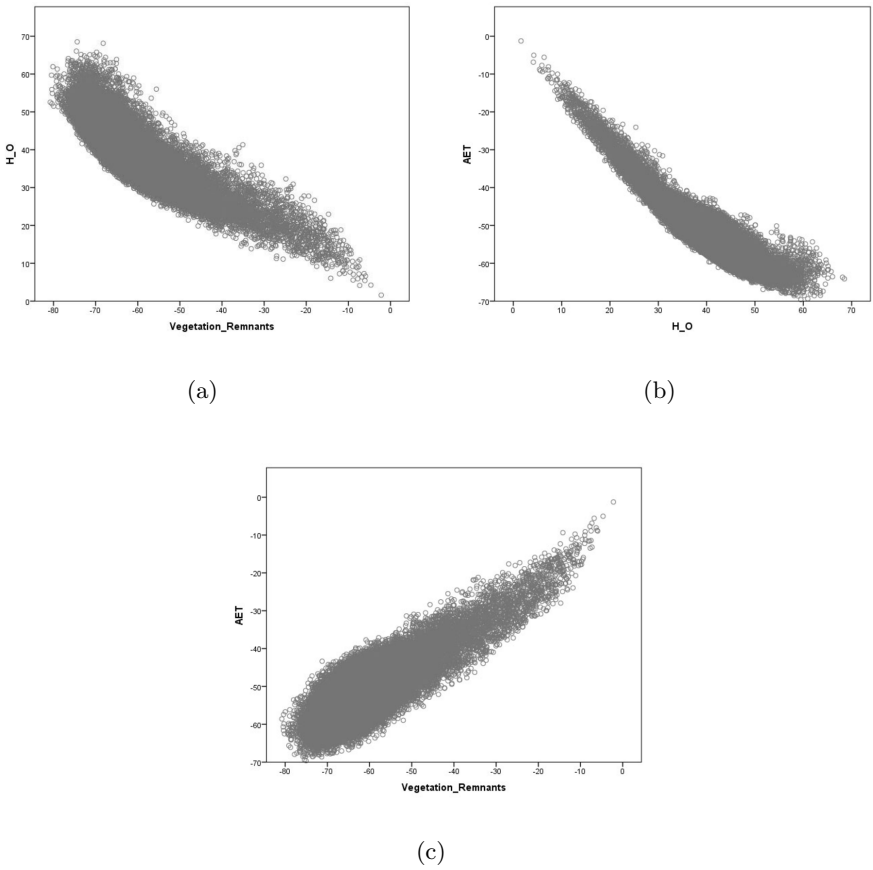


Fig. 4. Scenario 2, optimization of five objectives simultaneously. Scatterplot of additional objectives in pairs. A. VR vs H_O ($r = -.84, p < .0001$). B. H_O vs AET ($r = -.93, p < .0001$). C. VR vs AET, showing a positive correlation between them ($r = .84, p < .0001$), this is a poor pair for optimization. A and B show a negative correlation, revealing a conflicting behaviour which means that they are proper candidates for optimization.

values of AET were observed (human settlement follows better conditions patterns), even though the relationship between H_O and AET is inversely proportional (H_O has a deleterious effect on AET).

The scatterplot of AET and VR (Fig. 4.c) shows that these objectives have a positive correlation ($r = .84, p < .0001$), i.e., there is no conflict between them, revealing that this pair would be a poor candidate for multi-objective optimization. But since by definition, in optimal solutions, improving the value in one dimension of the objective function vector leads to a degradation in at least one other dimension of it, the other objectives (H_O and AET) hold optimization conditions.

As result of Scenario 2 experiments, we found that optimization in 5 dimensions allowed to supply decision makers with a more diversified portfolio, increasing the problem flexibility through the inclusion of more decision objectives, which whilst increasing the complexity, significantly augments the amount of information that can be used to provide users with an improved decision support system.

Although we shall investigate these aspects further, the current study makes a significant contribution by applying a multi-objective optimization method to a real-world problem. This reveals important relationships among objectives that are common to conservation scenarios of a practical SCP problem.

4.3 Scenario 3

The objective of this optimization was to locate the smallest set of sites that would be required to represent all species in 2080.

First, it is important to mention that it is not possible to represent all the 96 species, since one of them was extinct (species #80). Moreover, it is worthy of note that projections by Simon et al. [31] to 2080 show that the species under study will reduce about 78% of their geographic distribution in Cerrado due to climate change that will have a strong influence on the distribution pattern of these species, regardless the conservation plan adopted.

In this new scenario, the mimimum set of sites that represent the highest diversity of plants (95 species) is 5. We found 4 distinct solutions with these characteristics. Irreplaceability for Scenario 3 can be seen in Fig. 3.B, while Fig. 3.C corresponds to the synthesis of information from Scenarios 1 and 3, where positive values imply gain of irreplaceability and negative, loss. The irreplaceability map has the advantage of showing the flexibility degree of systematic conservation sites [23].

It is worth noting that despite ensemble forecast approach allows more accurate predictions on changes in the species bioclimatic envelope, it is not possible to remove all uncertainty associated with projections of future climates [31].

4.4 Dynamic Spatial Prioritization

Having a picture of how future scenarios will look like can be extremely useful for decision makers. To assess the relative importance of sites in achieving conservation targets both for current time (Scenarios 1 and 2) and for future (Scenario 3), we compared the variation of site *selection frequency* scores under dynamic conditions, using a bi-dimensional graph (Fig. 5).

Grid cells populating the *upper right corner* of the graph (framed by a square) are important both for current time and for future scenarios of climate change, therefore it would be a good choice to invest in them. However, grid cells located in the *lower right corner* (framed by a rectangle) represent a risk of conservation investment given their low relative importance in 2080, so, based in this information, the decision maker might opt not to invest in these regions, redirecting

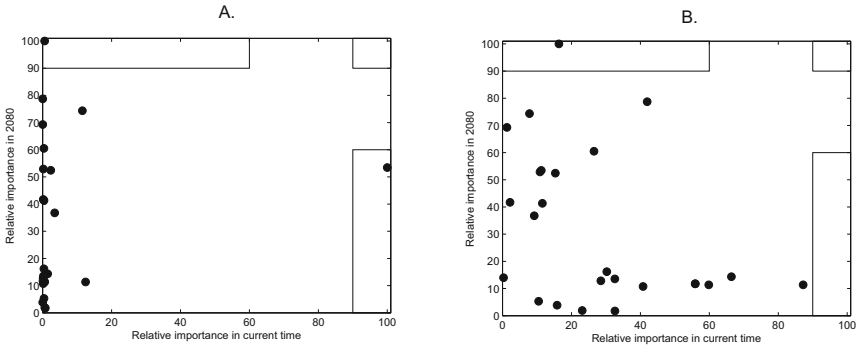


Fig. 5. Graphs for establishing a dynamic spatial conservation prioritization analysis. A. Scenario 1 x Scenario 3. B. Scenario 2 x Scenario 3. The relative importance of grid cells is given by a rank ranging from 0 to 100, based on their selection frequency. Grid cells placed in the *upper right corner* are important in the present time and in a future scenario. Grid cells in the *lower right corner* represent a risk of conservation investment as they seems to be very important in current time, but of low relative importance in future. Grid cells in the *upper left corner* gain attention since they are not critical for current time, but might became important in the future.

funds to another more promising area. Nevertheless, grid cells in the *upper left corner* (framed by a rectangle) deserve attention as they might become very important in the future even though they are not critical at present time. In this case, careful land-use planning is imperative because these regions can represent good cost-benefit in the long-term.

This information, associated to data displayed in Fig. 3.C, provides important knowledge support to decision makers. Supported by scientific data, they can scrutinize the options available in current time and decide how to define their spatial conservation priorities, reviewing them if necessary.

For optimization in two dimensions (Scenario 1) (Fig. 5.A), results show data concentrated along the vertical axis (that represents importance in 2080). We were able to identify a region in the *upper left corner*, representing a location that probably will become very important in the future, although not being critical at present time. We also found a region in the *lower right corner* that can represent a risk to conservation investment given its low relative importance in 2080.

Our results show that with optimization using additional objectives (in 5 dimensions; Scenario 2) (Fig. 5.B), we were able to find data more smoothly spread along both relative importance axes, and specially leading to the *upper right corner* and closer to the *upper left corner* that could be the most attractive locations to invest.

Although it would be interesting to find data in the *upper right corner* of Fig. 5, the results reflected the available data, and it strongly indicates that these solutions simply do not exist. However, our method is able to identify (if they exist) sites of high priority for conservation, regions with high risk of investment and sites that may become attractive options in the future. And these data can be used in order to help decision makers to select their schemes of conservation.

5 Conclusions and Future Work

As far as we know, this is the first attempt to apply multi-objective algorithms to a SCP problem associated to climate forecasting, in a dynamic spatial prioritization analysis for biodiversity. Our work improves the methods used by most of the tools for SCP, which in general apply a static and monobjective approach.

We applied the proposed new approach to a real and important SCP problem that is the conservation of the Brazilian Cerrado obtaining consistent and useful results.

The use of more dimensions allows to incorporate relevant information in the context of SCP, increasing the complexity of the process but in a more intuitive and simpler way (without the assistance of an expert).

We suggest that priorities for conservation could be integrated into a strategy that considers different additional objectives helping to select areas, which results in a conservation plan that is likely to be more effective taking into account the impact of climate change.

The dynamic analysis is an improvement compared to the static approach since it reflects a significant opportunity to adjust priorities into biodiversity conservation plan, by comparing the relative importance of conservation targets in current time and in the future.

Although bioclimatic models are effective and widely used to evaluate the consequences of climate changes for biodiversity, there are still many uncertainties associated to projections to the future.

Our results show that, despite the encouraging achievements, efforts to address the loss of biodiversity need to be strengthened by complementary policies, since changes in climate are inevitable and tend to strongly affect conservation projects as result of the direct influence on the persistence of species.

This was an exploratory study that showed the advantages of the new approach with respect to previous solutions. Having established that the approach is viable with a standard MOO algorithm, our future work will focus on the development of a multi-objective algorithm more specialized to the SCP problem. Given the success of a variety of work in the Artificial Immune System area, e.g. [30], who showed better solutions (closer to the origin axes an more regularly spread throughout the known Pareto Front), we will build on that work to improve the work presented in this paper.

We also plan to perform further comparative studies addressing SCP problem scenarios that deal with optimization of more than three objectives (e.g. Scenario 2), applying approaches as many-objective optimization [16] and bilevel optimization [12].

Acknowledgments. SS wishes to thank the University of York and Prof. Jon Timmis for the PhD stay, and the support from CNPq throughout a Science without Borders scholarship. Jon Timmis is part funded by The Royal Society. GENPAC has been supported by CNPq/MCT/CAPES (projects #564717/2010-0 and #563624/2010-8) and by the GECER (PRONEX/FAPEG/CNPq CP 07-2009). Work by MEW, AC, RL and JADF have been continuously supported by productivity fellowships from CNPq.

References

1. Alden, K., Read, M., Timmis, J., Andrews, P.S., Veiga-Fernandes, H., Coles, M.: Spartan: A Comprehensive Tool for Understanding Uncertainty in Simulations of Biological Systems. PLoS Comput. Biol. 9(2), e1002916+ (2013)
2. Ardron, J., Possingham, H.P., Klein, C. (eds.): Marxan Good Practices Handbook. Pacific Marine Analysis and Research Association (PacMARA), Victoria, BC, Canada (July 2010)
3. Ball, I.R.: Mathematical Applications for Conservation Ecology: The Dynamics of Tree Hollows and the Design of Nature Reserves. PhD thesis, University of Adelaide, Dept. Applied Mathematics, Env. Science and Management (2000)
4. Balmford, A., Moore, J., Brooks, T., Burgess, N., Hansen, A., Williams, P., Rahbek, C.: Conservation Conflicts Across Africa. Science **291**, 2616–2619 (2001)
5. Cabeza, M., Moilanen, A.: Design of Reserve Networks and the Persistence of Biodiversity. Trends Ecol. Evol. **16**(5), 242–248 (May 2001)
6. Carvalho, F., Ferreira, L., Lobo, F., Diniz-Filho, J., Bini, L.: Spatial Autocorrelation Patterns of The Modis Vegetation Indices for the Cerrado Biome. Revista Árvore. **32**(4), 279–290 (2008)
7. Ciarleglio, M.: Modular Abstract Self-Learning Tabu Search (MASTS) Metaheuristic Search Theory and Practice. PhD thesis, Univ. Texas at Austin, Texas (2008)
8. Ciarleglio, M., Barnes, J., Sarkar, S.: ConsNet - A Tabu Search Approach to the Spatially Coherent Conservation Area Network Design Problem. J. Heuristics **16**, 537–557 (2010)
9. Coello Coello, C.A., Lamont, G.B., Van Veldhuizen, D.A.: Evolutionary Algorithms for Solving Multi-Objective Problems, 2nd edn. Springer, New York (2007) ISBN 978-0-387-33254-3
10. de Almeida, S.P.: Cerrado: Aproveitamento Alimentar (Cerrado: Food Utilization). Embrapa - CPAC, Planaltina (1998) (in Portuguese)
11. Deb, K., Rudolph, G., Lutton, E., Merelo, J.J., Schoenauer, M., Schwefel, H.-P., Yao, X. (eds.): PPSN 2000. LNCS, vol. 1917, pp. 849–858. Springer, Heidelberg (2000)
12. Deb, K., Sinha, A.: Evolutionary Bilevel Optimization (EBO). In: Proceedings of the 2014 Conference Companion on Genetic and Evolutionary Computation Companion, GECCO Comp 2014, pp. 857–876. ACM, New York (2014)
13. Diniz-Filho, J., Bini, L., Rangel, T., Loyola, R., Hof, C., Nogués-Bravo, D., Araújo, M.: Partitioning and Mapping Uncertainties in Ensembles of Forecasts of Species Turnover Under Climate Change. Ecography **32**(6), 897–906 (2009)
14. Diniz-Filho, J., Bini, L., Vieira, C., Blamires, D., Terribile, L., Bastos, R., Oliveira, G., Souza, B.: Spatial Patterns of Terrestrial Vertebrate Species Richness in the Brazilian Cerrado. Zool. Stud. **47**(2), 146–157 (2008)
15. Groves, C., Game, E., Anderson, M., Cross, M., Enquist, C., Ferdaña, Z., Girvetz, E., Gondor, A., Hall, K., Higgins, J., Marshall, R., Popper, K., Schill, S., Shafer, S.: Incorporating Climate Change into Systematic Conservation Planning. Biodivers. Conserv. **21**, 1651–1671 (2012)
16. Jain, H., Deb, K.: An Evolutionary Many-Objective Optimization Algorithm Using Reference-Point-Based Nondominated Sorting Approach, Part I: Solving Problems With Box Constraints. IEEE Journal **18**(4), 577–601 (2013)
17. Loyola, R., Lemes, P., Nabout, J., Trindade-Filho, J., Sagnori, M., Dobrovolski, R., Diniz-Filho, J.: A Straightforward Conceptual Approach For Evaluating Spatial Conservation Priorities Under Climate Change. Biodivers. Conserv. **22**, 483–495 (2013)

18. Luck, G., Ricketts, T., Daily, G., Imhoff, M.: Alleviating Spatial Conflict Between People and Biodiversity. Proc. Natl. Acad. Sci. USA **101**(1), 182–186 (Jan. 2004)
19. Margules, C.R., Pressey, R.L., Nicholls, A.O.: Nature Conservation: Cost Effective Biological Surveys and Data Analysis, chapter Selecting Nature Reserves. Commonwealth Scientific & Industrial Research (CSIRO), Dickson, Australia (1991)
20. McCarthy, D.P., Donald, P.F., Scharlemann, J.P.W., et al.: Financial Costs of Meeting Global Biodiversity Conservation Targets: Current Spending and Unmet Needs. Science **338**(6109), 946–949 (2012)
21. Meir, E., Andelman, S., Possingham, H.P.: Does Conservation Planning Matter in a Dynamic and Uncertain World? Ecol. Lett. **7**(8), 615–622 (2004)
22. Possingham, H.P., Ball, I., Andelman, S.: Mathematical Methods for Identifying Representative Reserve Networks, ch. 17, pp. 291–305. Springer, New York (2000)
23. Pressey, R.L.: Ad Hoc Reservations: Forward or Backward Steps in Developing Representative Reserve Systems? Conserv. Biol. **8**, 662–668 (1994)
24. Pressey, R.L.: The First Reserve Selection Algorithm: a Retrospective on Jamie Kirkpatrick's 1983 Paper. Prog. Phys. Geog. **26**(3), 434–441 (2002)
25. Pressey, R.L., Possingham, H.P., Day, J.R.: Effectiveness of Alternative Heuristic Algorithms for Identifying Indicative Minimum Requirements for Conservation Reserves. Biol. Conserv. **80**(2), 207–219 (1997)
26. Pressey, R.L., Possingham, H.P., Margules, C.R.: Optimality in Reserve Selection Algorithms: When Does it Matter and How Much? Biol. Conserv. **76**(3), 259–267 (1996)
27. Rangel, T., Bini, L., Diniz-Filho, J., Pinto, M., Carvalho, P., Bastos, R.: Human Development and Biodiversity Conservation in Brazilian Cerrado. Appl. Geogr. **27**(1), 14–27 (2007)
28. Rodríguez, M., Belmontes, J.A., Hawkins, B.: Energy, Water and Large-scale Patterns of Reptile and Amphibian Species Richness in Europe. Acta. Oecol. **28**, 65–70 (2005)
29. Sarkar, S.: Complementarity and the Selection of Nature Reserves: Algorithms and the Origins of Conservation Planning, 1980–1995. Arch. Hist. Exact. Sci. **66**, 397–426 (2012)
30. Schlottfeldt, S., Saéz, Y., Isasim P.: Sistemas Inmunológicos Artificiales aplicados al Problema de Optimización Multiobjetivo Radio Network Design (Artificial Immune Systems applied to the Radio Network Design Problem). Technical Report UC3M-TR-CS-2009-01, Universidad Carlos III de Madrid (2009) (in Spanish)
31. Simon, L., Oliveira, G., Barreto, B., Nabout, J., Rangel, T., Diniz-Filho, J.: Effects of Global Climate Changes on Geographical Distribution Patterns of Economically Important Plant Species in Cerrado. Rev. Árvore. **37**(2), 267–274 (2013)
32. Sinha, A., Malo, P., Xu, P., Deb, K.: A Bilevel Optimization Approach to Automated Parameter Tuning. In: Proceedings of the 2014 Conference Companion on Genetic and Evolutionary Computation Companion, GECCO Comp 2014, pp. 847–854, ACM, New York (2014)

8.7 *Paper 6*

Shana Schlottfeldt, Maria Emilia M.T. Walter, Jon Timmis, Andre C.P.L.F. Carvalho, Mariana P.C. Telles, and Jose Alexandre F. Diniz-Filho. 2015. Using Multi-Objective Artificial Immune Systems to Find Core Collections Based on Molecular Markers. In Proceedings of the 2015 on Genetic and Evolutionary Computation Conference, GECCO'15, Sara Silva (Ed.). ACM, New York, NY, USA, 1271-1278. DOI=10.1145/2739480.2754653 http://doi.acm.org/10.1145/2739480.2754653.

Using Multi-Objective Artificial Immune Systems to Find Core Collections Based on Molecular Markers

Shana Schlottfeldt
Dept of Computer Science
University of Brasilia
Brasilia, DF, Brazil
shanass@unb.br

Maria Emilia M. T. Walter
Dept of Computer Science
University of Brasilia
Brasilia, DF, Brazil
mia@cic.unb.br

Jon Timmis
Dept of Electronics
University of York
York, United Kingdom
jon.timmis@york.ac.uk

André C.P.L.F. Carvalho
Dept of Computer Science
University of Sao Paulo
São Carlos, SP, Brazil
andre@icmc.usp.br

Mariana P.C. Telles
Institute of Biological Sciences
Federal University of Goias
Goiania, GO, Brazil
tellesmpc@gmail.com

José Alexandre F. Diniz-Filho
Institute of Biological Sciences
Federal University of Goias
Goiania, GO, Brazil
diniz@ufg.br

ABSTRACT

Germplasm collections are an important strategy for conservation of diversity, a challenge in ecoinformatics. It is common to select a core to represent the genetic diversity of a germplasm collection, aiming to minimize the costs of conservation, while ensuring the maximization of genetic variation. For the problem of finding a core for a germplasm collection, we proposed the use of a constrained multi-objective artificial immune algorithm (MAIS), based on principles of systematic conservation planning (SCP), and incorporating heterozygosity information. Therefore, optimization takes genotypic diversity and variability patterns into account. As a case study, we used *Dipteryx alata* molecular marker information. We were able to identify within several accessions, the exact entries that should be chosen to preserve species diversity. MAIS presented better performance measure results when compared to NSGA-II. The proposed approach can be used to help construct cores with maximal genetic richness, and also be extended to *in situ* conservation. As far as we know, this is the first time that an AIS algorithm is applied to the problem of finding a core for a germplasm collection using heterozygosity information as well.

CCS Concepts

•Applied computing → Environmental sciences; •Mathematics of computing → *Evolutionary algorithms;* **•Computing methodologies** → *Discrete space search;*

GECCO '15, July 11 - 15, 2015, Madrid, Spain

DOI: http://dx.doi.org/10.1145/2739480.2754653

Keywords

multi-objective optimization; artificial immune systems; systematic conservation planning; germplasm; core collection; genetic variability; biodiversity

1. INTRODUCTION

Nowadays, biodiversity, sustainability, and climate change represent strategic research issues, compelling researchers to deal with integration of data within and among distinct disciplines, and rapid conversion of those data into information required by decision-makers [16].

In this context, ecoinformatics plays an important role as an emerging multidisciplinary field that develops methods and tools for the understanding, generation, processing, and dissemination of ecological data [22]. While the components of this field are not new [16], there is a novel emphasis on the integrated treatment of the area, associated with the increasing social awareness of ecological and environmental issues and their social, economic, and political impacts [20, 27]. As a consequence, ecoinformatics is experiencing a fast growing, similarly to the one seen in bioinformatics, some decades ago.

A focal research challenge in ecoinformatics is related to plant germplasm maintenance. *Plant germplasm* is a living tissue from which new plants can be grown. There are several ways to store germplasm, e.g., as seed collections, pollen storage, in a nursery (field), *in vitro* [13], comprising what is called *germplasm collection.* Germplasm collections are an important strategy for conservation and maintenance of genetic resources since they preserve the genetic diversity of plants [9], which, therefore, is made available for further study or for habitat restoration projects.

The maintenance of a germplasm collection is expensive and its costs are related to storage space, controlled temperature, relative humidity, and amount of equipment [15]. These factors depend mainly on the quantities of germplasm to be stored. This leads us to the concept of *core collection*, a subset from a larger collection of a particular species that represents, with a minimum level of repetitiveness, the ge-

netic diversity of that species [3, 14]. A core collection is not a surrogate of the whole collection, but it captures the complete diversity of the entire collection it is derived from; thus, it is a useful tool for organizing and analyzing representative sets of genotypes in a germplasm collection.

We use the term *accession* to refer to any sample in the whole collection, and *entry* to refer to any accession selected for inclusion in the core [3].

Methodologically, efforts to create germplasm core collections commonly use complex but primitive tools based on statistical and clustering methods [2, 28, 29].

In the development of a core collection, we look for minimizing the overall cost of conservation while maximizing the genetic diversity. This problem can be mapped to the systematic conservation planning (SCP), a widely accepted biodiversity-focused approach to determine the most cost effective way of investing in conservation actions. In short, SCP is the problem of finding a minimum set of elements (in this case, entries for the core collection) with the maximal representation of some feature (here the genetic diversity) [19]. Evidently, there are at least two conflicting objectives, which makes the problem a perfect candidate for multi-objective optimization (MOO). SCP can be modeled by the well known NP-hard minimum set covering problem [6].

SCP has been generally applied at species level (or hierarchically higher), but has also been used in conservation genetics, aiming to maximize molecular variation within populations [12]. In this study, this is attained by using molecular markers. The use of molecular markers to achieve the representativeness of a core collection is important chiefly because population persistence and resilience to environmental changes are consistently correlated with genetic diversity.

Previously, our group successfully applied MOO to a problem of *in situ* conservation (i.e., preserving plant species in their natural habitat) [24]. At that time, the well known NSGA-II [11] was employed.

Here we propose a more sophisticated MOO approach using a constraint-handling multi-objective algorithm inspired by the immune system (i.e., an artificial immune system – AIS) and based on individual molecular variability aiming to guide sampling to find germplasm core collections. We represented the known alleles, but incorporated individual *heterozygosity* information[1]. Thus, optimization also takes genotypic diversity and variability patterns into account. By including these characteristics, accessions can better represent the genetic diversity, allowing to identify sets with a higher probability of persistence throughout time.

As far as we know, this is the first time that an AIS multi-objective algorithm is applied to a SCP problem, in particular to the problem of finding a germplasm core collection using heterozygosity information as well.

The remainder of the paper is structured as follows. Section 2 describes the material and methods adopted in this study, in special, briefly presents the used constrained multi-objective AIS algorithm. In Section 3, we discuss the results obtained so far. Conclusions and future work are presented in Section 4.

[1]Heterozygosity is the state of being heterozygous, i.e., having two different alleles of the same gene; it is positively correlated with genetic variation and evolutionary potential.

2. MATERIAL AND METHODS

2.1 Data

As a case study for our method, we used a *Dipteryx alata* data set composed of 642 individual trees sampled in 25 local populations throughout *D. alata*'s geographic range (Fig. 1) (see [26] for sampling methodological details). This data set will be hereafter referred as our germplasm collection, and each sample as an accession.

The *D. alata* samples were genotyped for nine microsatellite *loci*, finding a total of 55 distinct alleles. These microsatellites (also known as simple sequence repeats – SSRs) are our molecular markers.

Figure 1: Samples of *Dipteryx alata*, also known as baru, a widely distributed tree species in Cerrado biome, Central Brazil. It is used as lumber, for charcoal production, shade in pasture, and it is source of raw material for handcraft, cosmetics, and food industries, playing an important role in local economy [7]. Most of the *D. alata* diversity is found only in nature, and many such populations are increasingly threatened by habitat reduction. Additionally, individual trees are often geographically wide ranging, making it costly to collect representative samples.

Based on these data, we produced an allele-by-accession presence-absence matrix $A_{k\times m}$, where $k = 642$ (accessions corresponding to the sampled individual trees) and $m = 55$ (alleles), a_{ij} represents the occurrence of allele j in accession i. In this matrix, alleles in homozygosity received value 1, and alleles in heterozygosity, value 2; by doing this, we benefited solutions with higher content of heterozygosity.

2.2 Problem Formalization

The overall problem is to maximize the number of alleles while minimizing the number of entries required to represent these alleles, maximizing, at the same time, the heterozygosity.

A candidate solution for the problem is a vector $\vec{x} = \{x_1, ..., x_k\}$, where k is the number of accessions (sampled individual trees), $x_i \in \{0, 1\}$, such that $x_i = 1$, if the accession i is selected to compose the solution; or 0, otherwise. Each site i have a cost c_i, and each feature j a desired representation level r_j. The aim is to obtain:

$$min\left(\sum_{i=1}^{k} c_i x_i\right) \quad (1)$$

Subject to:

$$\forall j \in \{1, 2, ..., m\}, \sum_{i=1}^{k} a_{ij} x_i \geq r_j \quad (2)$$

Where m = total number of alleles (i.e., $m = 55$), and $r_j = 1$.

As fitness functions, we have as many representations of Eq. 1 as objectives to be optimized, varying c_i according to the objective under consideration. This allowed us to simultaneously optimize distinct objectives instead of aggregating them into one single function. So, regarding the MOO approach, we optimized three objectives:

1. Minimize the number of selected accessions (i.e., the number of entries for the core);

$$min(f_1(\overrightarrow{x})) = min(selected\_accessions(\overrightarrow{x})) \quad (3)$$

2. Maximize the representation of alleles;

$$\begin{aligned} max(f'_2(\overrightarrow{x})) = max(alleles(\overrightarrow{x})) \Leftrightarrow \\ min(f_2(\overrightarrow{x})) = min(lacking\_alleles(\overrightarrow{x})) \\ = min(m - alleles(\overrightarrow{x})) \end{aligned} \quad (4)$$

3. Maximize the heterozygosity.

$$\begin{aligned} max(f'_3(\overrightarrow{x})) = max(heterozygosity(\overrightarrow{x})) \Leftrightarrow \\ min(f_3(\overrightarrow{x})) = min(-heterozygosity(\overrightarrow{x})) \end{aligned} \quad (5)$$

We worked with minimization, henceforth, based on the duality principle, w.l.o.g., we converted all objectives to their equivalent minimization representation (e.g., for Eq. 4, optimization consisted in minimizing the number of lacking alleles, which is the same as maximizing the number of alleles).

According to the literature, the core chosen to represent as much as possible of the collection diversity is composed of about 10% of the total collection [3]. Hence, we defined $\approx 10\%$ of lacking alleles, and $\approx 15\%$ from the number of accessions as constraints to delimit our objective space. Corroborated this decision, the experts (biologists) information that solutions with more than 5 lacking alleles and 107 entries would not be acceptable in practice. As a result, constraints were defined as penalties based on the objective functions as follow:

$$c_1(\overrightarrow{x}) = \begin{cases} f_1(\overrightarrow{x}) - 107, & \text{if } f_1(\overrightarrow{x}) > 107; \\ 0, & \text{otherwise.} \end{cases} \quad (6)$$

$$c_2(\overrightarrow{x}) = \begin{cases} f_2(\overrightarrow{x}) - 5, & \text{if } f_2(\overrightarrow{x}) > 5; \\ 0, & \text{otherwise.} \end{cases} \quad (7)$$

In this context, a solution that violates any constraint is said to be *infeasible*, otherwise, it is *feasible*.

Henceforth, we used the concept of *constrained-dominance* [11], where a solution i is said to constrained-dominate a solution j ($i \preceq_{cd} j$) if any of the following conditions is true:

1. Solution i is feasible and solution j is not;
2. Solution i and j are both infeasible, but solution i has smaller overall constraint violation;
3. Solutions i and j are feasible and solution i dominates solution j.

2.3 MAIS: the Constrained Multi-Objective AIS Algorithm

The constrained multi-objective AIS algorithm (MAIS), proposed here, is based on the clonal selection principle [4, 10] and consider two entities: *antigens (Ag)* and *antibodies (Ab)*. The input is the antigen-problem, and the output is composed of antibodies-solutions that recognize-solve Ag.

Algorithm 1 MAIS

```
1: P ← generateNewAb()
2: Pm ← ∅
3: while not(stopCondition()) do
4:     evaluate(P)
5:     Psel ← select(P)
6:     Pm ← updateMemory(Pm, Psel)
7:     Pc ← clone(Psel)
8:     Phyp ← hypermutate(Pc)
9:     Pmut ← mutate(P \ Psel)
10:    P ← P ∪ Phyp ∪ Pmut
11:    if (#OfGenerations mod X) = 0 then
12:        Pnew ← generateNewAb()
13:        P ← P ∪ Pnew
14:    end if
15:    return2OriginalSize(P)
16: end while
17: Pm ← updateMemory(Pm, P)
18: return Pm
```

We used a secondary population (memory – Pm), which keeps the best Ab's found (the *known Pareto front* – PF_{known}), and an *adaptive grid* (see [18]), to maintain the spread of solutions in the memory.

MAIS steps are as follow (Algorithm 1): randomly generate an initial population (P) (line 1) – at this moment, the memory (Pm) is empty (line 2). Evaluate P based on constrained-dominance (line 4). At the end of this step, P is sorted according to the following hierarchy of solutions:

1. Feasible non-dominated;
2. Feasible dominated;
3. Infeasible non-dominated;
4. Infeasible dominated.

Dominated solutions are sorted in ascending order, according to the number of solutions that dominates them. Infeasible solutions are sorted in ascending order as well, but according to the amount of constraint violations. Select the best Ab's (i.e., all feasible non-dominated Ab; if the number of feasible non-dominated individuals is less than 10% of the population size, then select Ab's following the constrained-dominance hierarchy until reaching a number of individuals equal to the 10% of the population size) to be cloned (line 5). Copy the best Ab's obtained in the previous step into Pm (line 6). Entrance into memory is regulated using the adaptive grid. For each Ab selected in line 5, the constrained-dominance is verified against those that are already in memory:

1. If the selected Ab is dominated by any Ab already present in the memory, the new Ab is discarded;
2. All the Ab's belonging to the memory that are dominated by the new Ab are removed. Then, the possibility of the new Ab composing the memory is verified:
   (a) If the memory is not full, the Ab is allowed to enter;
   (b) Otherwise, if the new Ab belongs to the most populated region, it is not allowed to enter;

(c) Otherwise, it enters the memory, but an individual from the most populated cell is removed leaving space for the new Ab (the memory size is maintained).

For Ab's selected in line 5, clone them proportionally to their distance to the k-nearest neighbor, obtaining Pc (line 7). Hypermutate Pc inversely proportional to the hierarchy defined by constrained-dominance – feasible non-dominated Ab's suffer less mutation than hierarchically worse solutions, e.g., infeasible dominated ones (line 8). In line 9, a uniform mutation is applied to Ab's that were not selected in line 5 (we empirically verified that this step improved the overall final solution). At each set of X generations (here, $X = 25$), a number of new Ab's is generated and added to the main population (as a way of generating diversity, explore the objective space, and rescue the search from local maxima) (lines 10-14). Return the population P to its original size (select as many individuals as the original population size using the criterion of hierarchies taken by the constrained-dominance) (line 15). Repeat the process from line 3 until a stop criterion is achieved (here, number of evaluation = $500,000$). At the end of execution, return Pm, the set of the best Ab's found.

2.4 Experiments

The objective of this study is two fold: (1) to present a new approach to deal with the problem of finding a core to a germplasm collection; and (2) to assess the performance of MAIS by comparing it to a state-of-the-art MOO algorithm (e.g., NSGA-II).

We addressed the problem of defining a core collection by identifying, within our germplasm collection, the minimal set of accessions needed to represent all the genetic variability exhibited by the germplasm collection; therefore, we considered the heterozygosity. Our aim was to find out if the proposed MOO approach is viable, and also if the optimization of heterozygosity as an additional objective would result in some advantage when selecting accessions to compose the core, hence, improving the overall quality of results. We optimized the three objectives stated in Eqs. 3 to 5, subject to constraints expressed in Eqs. 6 and 7.

2.5 Experimental Setup

Algorithms. We used MAIS, and, as baseline, NSGA-II in its constrained version [11]. For each run, a population of initial solutions was randomly generated. These solutions were then evolved using MAIS and NSGA-II. Both algorithms were implemented in Matlab®.

Aleatory Uncertainties. To determine the number of runs required to mitigate aleatory uncertainty inherent to the stochastic algorithms, we used Spartan (Simulation Parameter Analysis R Toolkit Application) [1]. Following Spartan protocol, we analyzed 20 subsets sample sizes of 1, 5, 10, 50, 100, and 300 runs each, requiring, therefore, 9,320 individual runs (for each algorithm, MAIS and NSGA-II). For both algorithms, it was found that 300 runs were sufficient to reduce the effect magnitude of aleatory uncertainty on results to less than "small" (the desired level) (Fig. 2).

Parameter Settings. Before running the experiments, we used a sample set to estimate, empirically, the most suitable parameter values, which for MAIS were set to: pop.[2] size = 500; secondary pop. size = 500; clone rate = proportionally to k-nearest neighbor; hypermutation rate = inversely proportional to the hierarchy of constrained-dominance; uniform mutation = 1/L (where L is the number of accessions); number of new Ab's created at each 25 generations=20% of pop. size. NSGA-II parameter values were set to: pop. size = 500; crossover probability = 0.90; mutation probability = 0.5; mutation rate = 1/L. Besides, we used: crossover operator = single point crossover (SPX); selection by binary tournament.

[2]population.

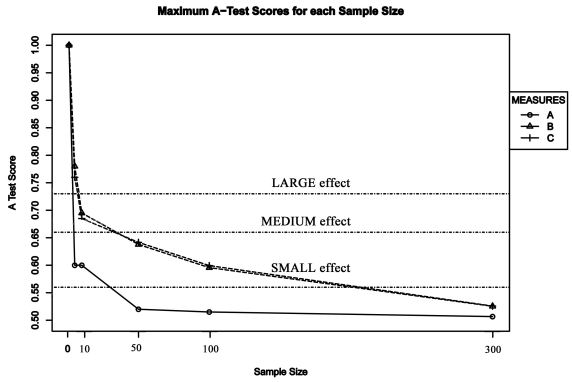


Figure 2: Spartan's Technique 1 applied to MAIS. This technique identifies which simulation results can be attributed to the dynamics of the modeled system, rather than artifacts of uncertainty or parametrization, or simulation stochasticity. At 300 runs, stochasticity over objectives A (lacking alleles), B (number of selected accessions), and C (heterozygosity) attains a small effect. Due to space limitation we present only MAIS results, but NSGA-II results were similar.

To grant an adequate comparison, we adopted the criteria suggested by Coello et al [5]: all algorithms executed the same number of fitness evaluations (500,000) ensuring a very nearly equivalent computational effort; pop. sizes and parameter values (whenever it was possible) were the same as well.

Computer Infrastructure The experiments were performed on a computer cluster consisting of 49 computers equipped with Intel(R) Core(TM) i5-2500 CPU 3.30GHz, 8GB RAM; and 9 computers with Intel(R) Core(TM) Duo CPU E7500 2.93GHz, 2GB RAM.

2.6 Performance Measures (Metrics)

According to Zitzler et al. [30], an optimization should:

1. Minimize the distance of the non-dominated set to the Pareto-optimal front;
2. Obtain a good spread of solutions;
3. Maximize the extent of the obtained non-dominated front (i.e., for each objective, a wide range of values should be covered by the non-dominated solutions).

In this study, these targets were assessed by the following performance measures:

Function C [30]**.** Let $\vec{x}'$ and $\vec{x}''$ be two sets of decision vectors. The function C maps the ordered pair $(\vec{x}', \vec{x}'')$ to the interval [0,1]:

$$C(\vec{x}', \vec{x}'') = \frac{|\{x' \in \vec{x}'; \exists x'' \in \vec{x}'' : x' \preceq_{cd} x''\}|}{|\vec{x}''|} \quad (8)$$

Using the function C, it can be seen if the outcomes of an algorithm dominate the outcomes of the other (i.e., a pair of non-dominated sets is compared by calculating the fraction of each set that is covered by the other). $C(\overrightarrow{x}', \overrightarrow{x}'') = 1$ implies that all solutions in $\overrightarrow{x}''$ are constrained-dominated by solutions in $\overrightarrow{x}'$, while $C(\overrightarrow{x}', \overrightarrow{x}'') = 0$, indicate that none of the solutions in $\overrightarrow{x}''$ is covered by the set $\overrightarrow{x}'$. Both $C(\overrightarrow{x}', \overrightarrow{x}'')$ and $C(\overrightarrow{x}'', \overrightarrow{x}')$ should be considered, since $C(\overrightarrow{x}', \overrightarrow{x}'')$ is not necessarily equal to $1 - C(\overrightarrow{x}'', \overrightarrow{x}')$.

Empirical attainment function (EAF) [5, 8]**.** EAF is a quality indicator used for stochastic algorithmic evaluation. It is computed from the combined collection of approximation sets. Let $b_1(z)...b_n(z)$ be n runs of the optimizer, then the EAF is defined as $EAF : \mathbb{R}^d \longmapsto [0, 1]$ with

$$EAF = \frac{1}{n} \sum_{i=1}^{n} b_i(z) \tag{9}$$

It offers a useful description of the solution distribution location. Differences in the frequency with which certain goals are met by the respective algorithms are represented graphically. The intensity of the shading corresponds to the frequency of the solution.

Hypervolume (H) [18]**.** For the set X of non-dominated vectors $\overrightarrow{x_i}$, and a reference vector $\overrightarrow{x}^{ref}$, which is dominated by all members of X and whose components are the maximum value on each objective (i.e., $\overrightarrow{x}^{ref} = \{max(f_1(\overrightarrow{x})), max(f_2(\overrightarrow{x})), max(f_3(\overrightarrow{x}))\} = \{642, 55, 0\}$), the *hypervolume* is the summation of all rectangular areas bounded by X and $\overrightarrow{x}^{ref}$ according to:

$$H(X, \overrightarrow{x}^{ref}) \triangleq \bigcup_{i \in 1..|X|} H(\overrightarrow{x_i}, \overrightarrow{x}^{ref}) \tag{10}$$

Higher values of H correspond to better solutions.

Spacing S [25]**.** This metric is used to numerically assess the spread of vectors in PF_{known}. It measures the distance variance between each solution and its nearest neighbor according to:

$$S = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (\overline{d} - d_i)^2} \tag{11}$$

Where $d_i = min_j ||f(\overrightarrow{i}) - f(\overrightarrow{j})||$, i.e., the norm/distance to the nearest neighbor of i; $i, j = 1, ..., n$; $\overline{d}$ is the mean of all d_i; f is an objective function (Eqs. 3 to 5); n is the number of vectors in PF_{known}. If $S = 0$, the algorithm has found the ideal distribution of non-dominated vectors (all vectors are uniformly spaced).

Extent E [30]**.** Let X be a set of objective vectors. E uses the maximum extent of decision vectors $\overrightarrow{a}, \overrightarrow{b} \in X$ in each dimension m to estimate the range to which the front spreads out:

$$E = \sqrt{\sum_{i=1}^{m} max \|a_i - b_i\|} \tag{12}$$

Where $\| \cdot \|$ is the norm/distance between two points. Concerning the value of E, the bigger the better.

Function C, EAF, and H metrics assess optimization target 1; S metric, target 2; and E metric, target 3.

3. RESULTS AND DISCUSSION

3.1 MAIS to find a Core Collection

The objective of the optimization was to select the smallest set of sites capable of representing the most amount of alleles (preferably all alleles), but at the same time optimizing heterozygosity, an additional objective.

We found that from the 642 accessions composing our germplasm collection, it is possible to preserve the allele diversity (55 alleles) by keeping a core of 16 selected accessions (only 2,5% of the germplasm collection). Even if the aim is to obtain a minimum set, our method identifies a portfolio of solutions, indicating sets with individuals that fulfill the objectives, providing decision-makers with additional alternatives for achieving their conservation targets. It is worth noting that there is no hierarchy among results, i.e., all of the solutions are equally good in the considered context and in the absence of additional preferences. As mentioned before, there is a flexibility on the core size, that is generally composed of about 10% of the total collection. The proposed method is important to define strategies to provide a set of genetically diverse material while selecting the most representative accessions. By maximizing genetic diversity in germplasm collections (throughout maximization of heterozygosity), resources available for conservation of biodiversity can be allocated more efficiently.

The output of 6,000 individual runs were unified[3], and we calculated the frequency for each accession in the union set. This frequency indicates the relative importance of an accession in order to fulfill the optimization objectives. This frequency can be taken as an estimator for accession *irreplaceability*[4] [21], e.g., the rarest alleles (there are three) appear in only one accession each, if one of them is not selected, the conservation goal of representing all the alleles is not achieved. These accessions are #145, #474, and #477, not surprisingly, they are among the most frequently selected in solutions. The used approach also privileged accessions with greater diversity of alleles, e.g., accessions #99 and #441, both with 16 different alleles[5] and 17 other accessions, with 15 different alleles each (Fig. 3).

We verified that there was a significant improvement in the retention of alleles in the selected accessions, suggesting that this approach is adequate to define a core collection.

The proposed approach is straightforward. Once the matrix based on the molecular markers is generated, there is no need of expertise to proceed the selection of accessions. Furthermore, it is simpler than the statistical and clustering methods, traditionally used.

The use of constraints allowed us to concentrate the exploration of the objective space in a more rational and efficient way, privileging feasible solutions over non-dominated ones, since the former are more valuable to decision-makers.

[3]since we already had 20 folders of 300 individual runs ($20 \times 300 = 6,000$) as a result of Spartan analysis, we used these data.

[4]a measure that indicates the proportion that an accession contributes to the overall solution, e.g., accessions with irreplaceability converging to 1 tend to be irreplaceable, i.e., if they are lost, the conservation goal may not be accomplished.

[5]*D. alata* germplasm collection was genotyped for 9 *loci*, since it is an diploid species, each accession has $9 \times 2 = 18$ possibly distinct alleles.

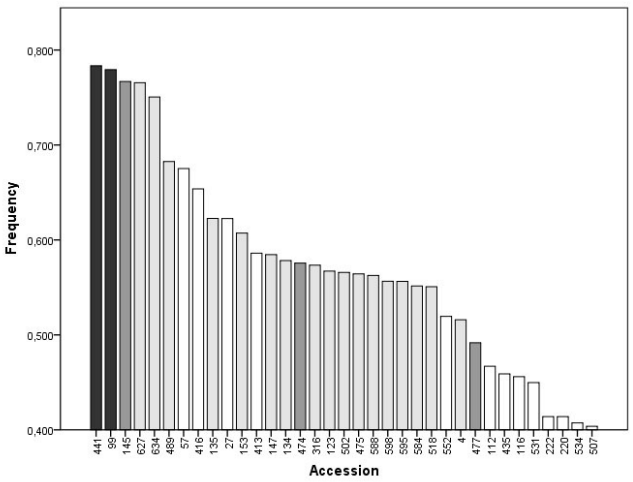


Figure 3: Solutions' accession frequency. Frequency scales from 0 to 1. Only accessions with frequency higher than 0.4 were plotted. It is possible to identify accessions: (1) with greater diversity (with 16 different alleles, in dark gray – #99 and #441 –, and with 15 different alleles, in light gray); (2) with the rarest alleles (in medium gray – #145, #474, and #477).

To the biologists, one of the most important contributions of this work is to identify, in the context of SCP, within a population of several individuals, the exact samples that should be chosen in order to preserve the species diversity.

This approach can be extended to the *in situ* conservation (selection of individual species to be preserved in their own habitat). Previous approaches [12] generally indicate a population to be preserved; the proposed method indicates exactly which individuals within the population should be sampled/kept. Indeed, this approach may be extended to any problem that can be mapped into the minimum set covering problem.

In Eq. 2, experts defined $r_j = 1$, $j \in \{1, ..., 55\}$, i.e., all of the alleles should be represented at least once. Nevertheless, it could be settled a different desired representation level $r_j \geq n > 1$, where n is the minimum number of times an allele should be present in solution. By doing so, allele representation in the core collection is heightened, benefiting persistence throughout time.

3.2 Comparison between MAIS and NSGA-II

To assess the performance of MAIS we compared it to NSGA-II. Firstly, per algorithm, the outcome of the 6,000 runs were unified, and then the dominated solutions were removed from the union set [30]. The remaining points were plotted (Fig. 4). It can be seen that MAIS is more effective in exploring the objective space, has better extent, was able to find smaller core collections, and has a better spread of solutions.

Additionally to the graphical presentation, the algorithms were assessed by using the performance measures defined in Subsection 2.6.

We calculated the function C using the previously described non-dominated front as input. We found *C(MAIS, NSGA-II)= 0.515217* and *C(NSGA-II,MAIS) = 0.5032*,

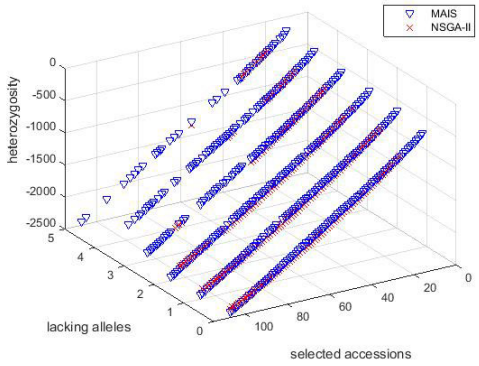


Figure 4: Non-dominate fronts achieved by MAIS and NSGA-II.

meaning that, although slightly, solutions found by MAIS covered NSGA-II, i.e., there were more MAIS solutions that constrained-dominated NSGA-II solutions than the opposite. This support the hypothesis that MAIS was able to find a better approximation to the true Pareto front (PF_{true}) – even though in this real-world problem PF_{true} is unknown.

The remaining metrics were computed for 6,000 independent runs of each algorithm[6].

By analyzing the EAF surfaces (Fig. 5), it can be seen that MAIS was able to obtain smaller core collections, and closer to the origin axis (i.e., likely closer to PF_{true}). Furthermore, MAIS has solutions more regular and smoothly distributed on the objective space delimited by the constraints, being able to better explore it.

Table 1 shows H, S, and E values for MAIS and NSGA-II.

In all individual runs of MAIS, H values where higher than the ones found for NSGA-II. The graphical analyses of Fig. 4 and EAF surfaces (Fig. 5), associated with the values for function C and H metric (Table 1 and Fig. 6a) corroborates the assumption that MAIS was able to find solutions closer to the PF_{true}.

Moreover, MAIS was able to find core collections more regularly spread throughout the PF_{known}, as can be seen in the S metric (Table 1 and Fig. 6b).

The E metric (Table 1 and Fig. 6c) shows that MAIS was more effective in extent the non-dominated front, exploring a wider range of the objective space.

Given these points, it can be said that in the problem of finding a core collection, MAIS presented better performance measure results when compared to NSGA-II, being able to find smaller and better distributed core collections, and to explore a larger extent of the objective space.

4. CONCLUSIONS AND FUTURE WORK

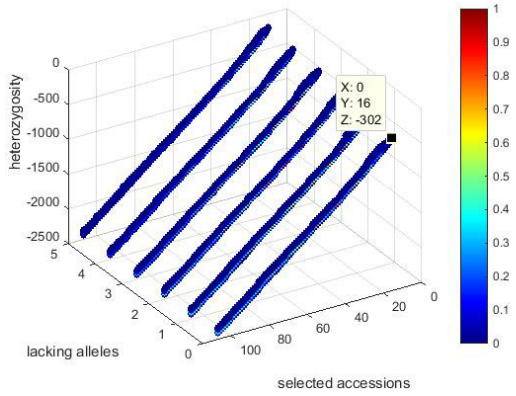
The problem of finding a core for a germplasm collection is NP-hard, nevertheless, decision-makers still need a solution for it. This is a relevant problem in ecology with real impact on resources availability (whether financial or genetic), and on biodiversity conservation. This work is inserted in the context of the emerging field of ecoinformatics.

As far as we know, this paper is pioneer in showing how principles of SCP and the use of a constrained multi-objective AIS algorithm (MAIS) associated to molecular marker information can be applied to successfully help construct core col-

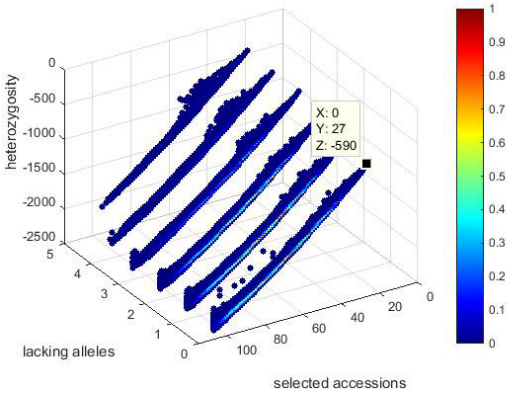
[6]the same observation made at footnote 3 applies here.

Table 1: Performance measures for MAIS and NSGA-II calculated for 6,000 independent runs. Best results are indicated in bold.

	Hypervolume (H)		Spacing (S)		Extent (E)	
	MAIS	NSGA-II	MAIS	NSGA-II	MAIS	NSGA-II
Mean	$\mathbf{7.75}\times10^7$	7.12×10^7	**2.1996**	2.2498	**46.966**	36.834
(Std.dev.)	$(\mathbf{0.35}\times10^4)$	(1.64×10^4)	(**0.0029**)	(0.0079)	(**0.399**)	1.624
Best	$\mathbf{7.88}\times10^7$	7.55×10^7	1.4430	**0.9904**	**48,166**	42,273
Worst	$\mathbf{7.58}\times10^7$	6.35×10^7	**3.0489**	6.2388	**44.944**	30.364

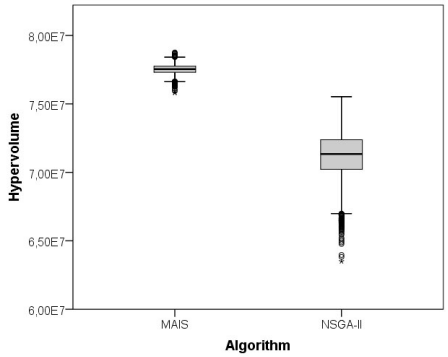


(a) MAIS

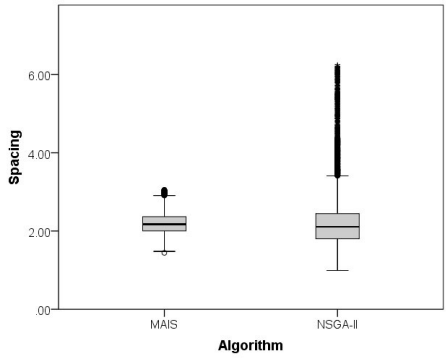


(b) NSGA-II

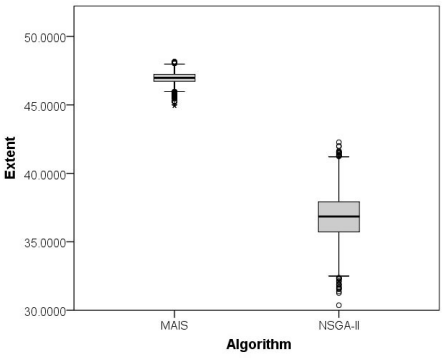
Figure 5: EAF surfaces showing the probabilities of attaining goals with (a) MAIS, and (b) NSGA-II. Results obtained after 6,000 independent runs. MAIS was able to find not only smaller cores when compared to NSGA-II, but also more regular and smoothly distributed solutions on the constrained objective space. X = lacking alleles; Y = selected accessions; Z = heterozygosity.



(a) Hypervolume (higher values are better)



(b) Spacing (lower values are better)



(c) Extent (higher values are better)

Figure 6: Boxplot of performance measures for MAIS and NSGA-II: (a) Hypervolume (H), (b) Spacing (S), and (c) Extent (E).

lections with maximum allelic richness and minimum number of accessions.

We performed comparisons of MAIS with NSGA-II and found that MAIS surpassed NSGA-II in all tested performance measures, being able to find better solutions, i.e., closer to PF_{true}, more evenly spread and exploring a larger extent of the objective space.

Moreover, using the proposed approach, it is possible to include additional optimization objectives to the problem, e.g., the distance from the germplasm collection facility to *in situ* individual trees; thus, reducing displacement costs

associated with collection of samples for complementing the germplasm collection.

Having established that the proposed approach is viable, a future work is to apply it to other kinds of molecular markers (e.g., single nucleotide polymorphism – SNP, and diversity arrays technology – DArT) in order to verify its feasibility for this kind of data.

We did not use domain information in MAIS, so there is still room for improvement. MAIS uses standard operators for hypermutation and cloning. Our future work will focus on the use of more complex operators (e.g., contiguous hypermutation, aging), which have shown interesting results in the literature [17, 23].

5. ACKNOWLEDGMENTS

SS wishes to thank the University of York and Prof. Jon Timmis for the PhD stay, and the support from CNPq throughout a Science without Borders scholarship. GENPAC has been supported by CNPq/MCT/CAPES. Work by MEMTW, ACPLFC, MPCT, RDL and JAFDF have been continuously supported by productivity fellowships from CNPq. Jon Timmis is part funded by The Royal Society and The Royal Academy of Engineering.

6. REFERENCES

[1] K. Alden et al. Spartan: A Comprehensive Tool for Understanding Uncertainty in Simulations of Biological Systems. *PLoS Comput Biol*, 9(2):e1002916+, Feb. 2013.

[2] A. Belaj et al. Developing a Core Collection of Olive (Olea europaea L.) Based on Molecular Markers (DArTs, SSRs, SNPs) and Agronomic Traits. *Tree Genet Genomes*, 8(2):365–378, Apr. 2012.

[3] A. Brown. Core Collections: a Practical Approach to Genetic Resources Management. *Genome*, 31(2):818–824, 1989.

[4] C. A. Coello Coello and N. C. Cortés. Solving Multiobjective Optimization Problems Using an Artificial Immune System. *Genet Program Evol M*, 6(2):163–190, 2005.

[5] C. A. Coello Coello, G. B. Lamont, and D. A. Van Veldhuizen. *Evolutionary Algorithms for Solving Multi-Objective Problems*. Springer-Verlag, New York, second edition, Sept. 2007. ISBN 978-0-387-33254-3.

[6] T. H. Cormen et al. *Introduction to Algorithms*. The MIT Press, 2 edition, 2001.

[7] G. C. Correa et al. Physical Determinations in Fruit and Seeds of Baru (*Dipteryx alata* Vog.), Cajuzinho (*Anacardium othonianum* Rizz.) and Pequi (*Caryocar brasiliense* Camb.), Aiming Genetic Breeding. *Biosci J*, 24(4):42–47, 2008.

[8] V. G. da Fonseca et al. Inferential Performance Assessment of Stochastic Optimisers and the Attainment Function. In E. Zitzler et al., editors, *EMO'2001*, volume 1993 of *Lect Notes Comput Sc*, pages 213–225. Springer-Verlag, 2001.

[9] I. Dawson and J. Were. Collecting Germplasm from Trees – Some Guidelines. *Agroforestry Today*, 9(2):6–9, 1997.

[10] L. N. de Castro and J. Timmis. *Artificial Immune Systems: A New Computational Intelligence Approach*. Springer-Verlag New York, Inc., NJ, USA, 2002.

[11] K. Deb et al. A Fast Elitist Non-Dominated Sorting Genetic Algorithm for Multi-Objective Optimization: NSGA-II. In *Proceedings of the PPSN'2000*, pages 849–858, London, UK, UK, 2000. Springer-Verlag.

[12] J. A. F. Diniz-Filho et al. Planning for Optimal Conservation Geographical Genetic Variability within Species. *Conserv Genet*, 13:1085–1093, 2012.

[13] J. Engels, L. Visser, and International Plant Genetic Resources Institute. *A Guide to Effective Management of Germplasm Collections*. IPGRI Handbooks for Genebanks. International Plant Genetic Resources Institute, 2003.

[14] O. H. Frankel. *Genetic Manipulation: Impact on Man and Society*, chapter Genetic perspectives of germplasm conservation, pages 161–170. Cambridge University Press, 1984.

[15] A. Gupta et al. Cost of Conservation of Agrobiodiversity. Technical Report WP2002-05-03, Indian Institute of Management Ahmedabad, Research and Publication Department, 2002.

[16] J. Helly et al. Technical report, National Center for Ecological Analysis and Synthesis, Santa Barbara, California, 1995. Available at http://www.nceas.ucsb.edu/papers/compecol/compecol.pdf.

[17] T. Jansen and C. Zarges. Analyzing Different Variants of Immune Inspired Somatic Contiguous Hypermutations. *Theor Comput Sci*, 412:517–533, 2011.

[18] J. D. Knowles and D. W. Corne. Properties of an Adaptive Archiving Algorithm for Storing Nondominated Vectors. *IEEE T Evolut Comput*, 7(2):100–116, 2003.

[19] C. R. Margules and R. L. Pressey. Systematic Conservation Planning. *Nature*, 405(6783):243–253, 2000.

[20] D. P. McCarthy et al. Financial Costs of Meeting Global Biodiversity Conservation Targets: Current Spending and Unmet Needs. *Science*, 338(6109):946–949, Nov. 2012.

[21] E. Meir, S. Andelman, and H. P. Possingham. Does Conservation Planning Matter in a Dynamic and Uncertain World? *Ecol Lett*, 7(8):615–622, 2004.

[22] W. K. Michener et al. Ecological informatics: A long-term Ecological Research Perspective. *Japanese Journal of Ecology (Otsu)*, 51(3):291–303, 2001.

[23] P. S. Oliveto and D. Sudholt. On the Runtime Analysis of Stochastic Ageing Mechanisms. In *Proceedings of the 2014 Conference on Genetic and Evolutionary Computation*, GECCO '14, pages 113–120, New York, NY, USA, 2014. ACM.

[24] S. Schlottfeldt et al. A Multi-Objective Optimization Approach Associated to Climate Change Analysis to Improve Systematic Conservation Planning. In A. Gaspar-Cunha et al., editors, *EMO'2015*, volume 9019 of *Lect Notes Compt Sc*, pages 458–472. Springer International Publishing, 2015.

[25] J. R. Schott. Fault Tolerant Design Using Single and Multicriteria Genetic Algorithm Optimization. Master's thesis, Department of Aeronautics and Astronautics, Massachusetts Institute of Technology, Cambridge, Massachusetts, May 1995.

[26] T. N. Soares et al. Development of Microsatellite Markers for the Neotropical Tree Species *Dipteryx alata* (Fabaceae). *Am J Bot*, 99:e72–e73, 2012.

[27] A. Waldron et al. Targeting global conservation funding to limit immediate biodiversity declines. *Proceedings of the National Academy of Sciences*, 110(29):12144–12148, 2013.

[28] Y. Wang et al. Construction and Evaluation of a Primary Core Collection of Apricot Germplasm in China. *Sci Hort*, 128:311–319, 2011.

[29] P. Zhang et al. Population Structure and Genetic Diversity in a Rice Core Collection (*Oryza sativa* L.) Investigated with SSR Markers. *PLoS One*, 6(12):e27565+, Dec. 2011.

[30] E. Zitzler, K. Deb, and L. Thiele. Comparison of Multiobjective Evolutionary Algorithms: Empirical Results. *Evol Comput*, 8(2):173–195, 2000.

8.8 *Paper7*

Schlottfeldt, S.; Walter, M.E.M.T.; de Carvalho, A.C.P.L.F.; Telles, M.P.C.; Diniz-Filho, J.A.F. Challenges in Computational Sustainability: a New Multi-Objective Artificial Immune System Approach to Deal with Biodiversity Conservation. In Proceedings of the Workshop on Evolutionary Multi-Objective Optimization at the IEEE 2015 International Congress on Evolutionary Computation, CEC'2015, Sendai, Japan, 2015.

Challenges in Computational Sustainability: a New Multi-Objective Artificial Immune System Approach to Deal with Biodiversity Conservation

Shana Schlottfeldt
and Maria Emilia M. T. Walter
Department of Computer Science
University of Brasilia
Brasilia, Brazil
Email: shanass@unb.br

André P. L. F. de Carvalho
Department of Computer Science
SCC-ICMC-USP
São Paulo, Brazil

Mariana P. C. Telles
and José Alexandre F. Diniz-Filho
Institute of Biological Sciences
Federal University of Goiás
Goiânia, Brazil

***Abstract*—A key issue in the advancement of policies for sustainable development is related to biodiversity conservation, which aim is to minimize the costs of conservation, while ensuring the maximal biodiversity representation, which is an NP-hard problem. We proposed the use of a constrained multi-objective artificial immune algorithm (MAIS), based on principles of systematic conservation planning (SCP), incorporating allelic and habitat information to deal with the biodiversity conservation problem; therefore, optimization takes genotypic diversity and variability patterns into account. As a case study, we used *Eugenia dysenterica* molecular marker information. We were able to identify the best set of populations that should be protected to preserve species diversity. The proposed approach can be used to help construct conservation schemes with maximal genetic richness, and also be extended to *ex situ* conservation. This is the first time that an artificial immune system algorithm is applied to the SCP problem using genetic and habitat information as well.**

I. Introduction

The 1992 United Nations Convention on Biological Diversity (CBD), was the first document to define biodiversity in the context of social, economic, and other environmental issues and also the first global agreement on the conservation and sustainable use of all components of biodiversity, including genetic resources, species, and ecosystems [1].

In this context, the growing interest and concern regarding biodiversity is leading scientists to help develop effective strategies to meet conservation goals in the emerging interdisciplinary field of computational sustainability [2] and ecoinformatics [3]. The underlying principle of these strategies lies in the systematic conservation planning (SCP), a broadly accepted biodiversity-focused approach to determine the most cost effective way of investing in conservation actions. Computationally speaking, SCP is typically formalized by the minimum set covering problem [4], known to be NP-hard [5].

In short, SCP is the problem of finding a minimum set of elements (e.g., sites, populations, individuals, etc.) with the maximum representation of the features under study (e.g., genetic diversity). There are clearly two conflicting objectives, which makes SCP a natural candidate for multi-objective optimization (MOO). Moreover, other objectives (e.g., human use of land, habitat loss, region stability, etc.) can be incorporated to the SCP problem, increasing its complexity.

SCP is a biodiversity and sustainability real-world problem that demands integrated solutions within and among distinct disciplines – specially computer science and ecology – and rapid conversion of those data into information for decision-makers.

Although the inherent multi-objective nature of SCP, it is often dealt with using a monobjective approach by assigning weights to the problem objectives, resulting in a unique objective function. Often, the subjectivity associated to such approach can drive to distinct results for the same data set. Moreover, when two criteria represent distinctive value systems it can be impossible to combine and/or compare such criteria in a meaningful manner [6]. Commonly, it is used a greedy approach in which complementary sites are selected in a sequential order, until they reach the goal of representing all features. Alternatively, the adoption of an exact approach was discussed, e.g., integer linear programming [7]. However, as SCP is a NP-hard problem, even the available software packages computing exact algorithms are not able to solve some large data sets [8].

SCP is usually applied at species level (or hierarchically higher), but a lower level approach, using molecular markers to maximize genetic variation representation within populations has been successfully employed [9], [10]. Additionally, MOO was favorably applied to a problem of *in situ*[1] conservation [6], [11], in that study, authors employed the well known NSGA-II [12].

Here we propose a more sophisticated MOO approach using a constrained multi-objective artificial immune system algorithm (MAIS) associated to individual molecular variability information aiming to guide *in situ* and *ex situ*[2] conservation

[1]species conservation in its own habitat.

[2]species conservation outside of its own habitat, generally in genetic banks (e.g., germplasm collections).

planning. We represented the known alleles, and incorporated region stability and habitat loss information as well. Thus, optimization takes genotypic diversity, variability patterns, and habitat information into account. Our hypothesis is that by including these characteristics, solutions can better represent the genetic diversity, allowing to identify sets with a higher probability of persistence over time. This is an attempt to associate molecular analyses with macroecologic patterns, aiming to deliver a more effective conservation strategy.

This is the first study to apply an artificial immune system (AIS) multi-objective algorithm to an SCP problem of finding schemes of conservation using region stability, habitat loss, and genotypic information.

The remainder of the paper is structured as follows. Section II describes the material and methods adopted, in special, briefly describes the MAIS algorithm. In Section III, we discuss the results obtained so far. Conclusions and future work are presented in Section IV.

II. Material and Methods

A. Data

As a case study for our method, we used an *Eugenia dysenterica* data set composed of individual trees sampled in 23 local populations throughout *E. dysenterica*'s geographic range.

The *E. dysenterica* samples were genotyped for microsatellite *loci*, finding a total of 249 distinct alleles. These microsatellites (also known as simple sequence repeats – SSRs) are our molecular markers. Based on these data, we produced an allele-by-population matrix $A_{p\times k}$, where $p = 23$ (populations), $k = 249$ (alleles), and a_{ij} represents the occurrence of allele j in population i. We also generated stability and habitat loss indexes for the regions were the *E. dysenterica* populations were sampled (Fig. 1).

B. Problem Formalization

In essence, the problem consists in maximizing the representation of alleles while minimizing the number of populations required to represent them. Additional objectives can be incorporated to the main problem, e.g., simultaneously maximizing stability, and minimizing habitat loss.

A candidate solution for the problem is a vector $\overrightarrow{x} = x_1, ..., x_p$, where p is the number of populations, $x_i \in \{0, 1\}$, such that $x_i = 1$, if the population i is selected to compose the solution; or 0, otherwise.

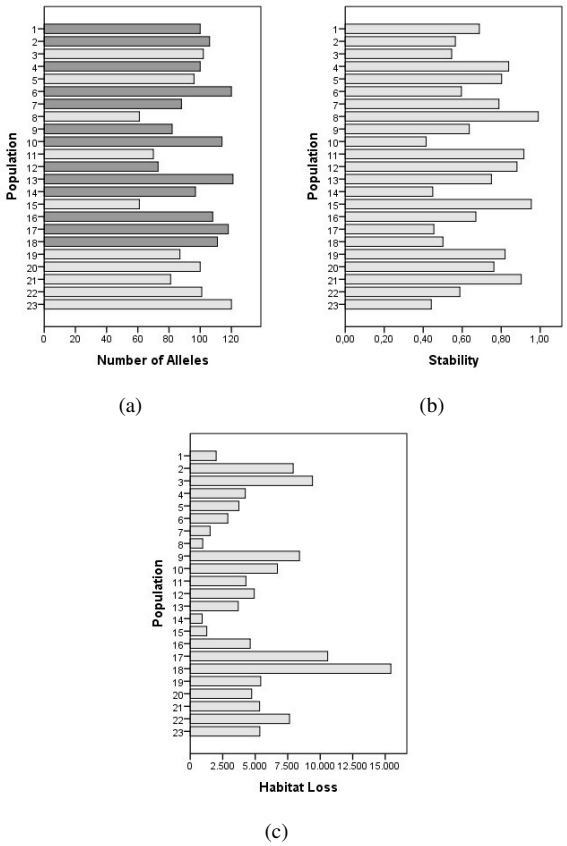
The aim is to obtain:

$$min\left(\sum_{i=1}^{p} x_i\right) \tag{1}$$

Subject to:

$$\forall j \in \{1, 2, ..., k\},\ \sum_{i=1}^{p} a_{ij}x_i \geq r_j \tag{2}$$

Where k =total number of alleles (i.e., $k = 249$) and the representation level $r_j = 1$.



(a) (b) (c)

Fig. 1. Alleles, stability and habitat loss values per population. (a) Number of distinct alleles (scaling 0-120). Populations containing rare alleles are indicated in dark grey. (b) Stability (scaling 0.00-1.00). (c) Habitat loss (scaling 0-15,000). It can be seen that less stable regions have higher habitat loss indices.

Concerning the MOO approach, there are four objectives to be optimized:

1. Minimize the number of selected populations;

$$min(f_1(\overrightarrow{x})) = min(populations(\overrightarrow{x})) \tag{3}$$

2. Maximize the representation of alleles;

$$\begin{aligned} max(f_2(\overrightarrow{x})) = max(alleles(\overrightarrow{x})) \Leftrightarrow \\ min(f_2'(\overrightarrow{x})) = min(lacking\_alleles(\overrightarrow{x})) \\ = min(k - alleles(\overrightarrow{x})) \end{aligned} \tag{4}$$

3. Maximize the region stability;

$$\begin{aligned} max(f_3(\overrightarrow{x})) = max(stability(\overrightarrow{x})) \Leftrightarrow \\ min(f_3'(\overrightarrow{x})) = min(-stability(\overrightarrow{x})) \end{aligned} \tag{5}$$

4. Minimize the habitat loss.

$$min(f_4(\overrightarrow{x})) = min(habitat\_loss(\overrightarrow{x})) \tag{6}$$

Based on the duality principle, w.l.o.g., we converted all objectives to their equivalent minimization (e.g., for Eq. 4, optimization consisted in minimizing the number of lacking alleles, which is the same as maximizing the number of alleles).

It was defined by the experts (biologists) that at least 95% from the total amount of alleles should be represented, meaning that solutions with more than 13 lacking alleles would not be acceptable in practice. Consequently, a constraint was defined as a penalty based on the objective function expressed in Eq. 4 as follows:

$$c_2(\overrightarrow{x}) = \begin{cases} f_2'(\overrightarrow{x}) - 13, & \text{if } f_2'(\overrightarrow{x}) > 13; \\ 0, & \text{otherwise.} \end{cases} \quad (7)$$

In this context, a solution that violates the constraint is said to be *unfeasible*, otherwise, it is *feasible*. The effect of the constraint is to exclude solutions that are non-dominated, but worthless to decision-makers. Therefore, we employed the *constrained-dominance* concept [12] in which a solution i is said to constrained-dominate a solution j ($i \preceq_{cd} j$), if any of the following conditions is true:

1. Solution i is feasible and solution j is not;
2. Solution i and j are both unfeasible, but solution i has smaller overall constraint violation;
3. Solutions i and j are feasible and solution i dominates[3] solution j.

C. MAIS: the Constrained Multi-Objective AIS Algorithm

The constrained multi-objective artificial immune system algorithm (MAIS), employed here, is based on the clonal selection principle [13], [14] and consider two entities: *antigens (Ag)* and *antibodies (Ab)*. The input is the antigen-problem, and the output is composed of antibodies-solutions that recognize-solve Ag.

Algorithm 1 MAIS

```
1:  P ← generateNewAb()
2:  Pm ← ∅
3:  while not(stopCondition()) do
4:      evaluate(P)
5:      Psel ← select(P)
6:      Pm ← updateMemory(Pm, Psel)
7:      Pc ← clone(Psel)
8:      Phyp ← hypermutate(Pc)
9:      Pmut ← mutate(P \ Psel)
10:     P ← P ∪ Phyp ∪ Pmut
11:     if (numberOfGenerations mod X) = 0 then
12:         Pnew ← generateNewAb()
13:         P ← P ∪ Pnew
14:     end if
15:     return2OriginalSize(P)
16: end while
17: Pm ← updateMemory(Pm, P)
18: return Pm
```

We used a secondary population (memory – Pm), which keeps the best Ab's found, and an *adaptive grid* [15], to maintain the spread of solutions in the memory.

[3]A vector $a = (a_1, ..., a_k)$ is said to dominate another vector $b = (b_1, ..., b_k)$ ($a \preceq b$) if and only if a is partially less than b, i.e., $\forall i \in \{1, ..., k\}, a_i \leqslant b_i \wedge \exists j \in \{1, ..., k\} : a_j < b_j$ [5].

MAIS steps are as follow (Algorithm 1): randomly generate an initial population (P) (line 1) – at this moment, the memory (Pm) is empty (line 2). Evaluate P based on constrained-dominance (line 4). At the end of this step, P is sorted according to the following hierarchy of solutions:

1. Feasible non-dominated;
2. Feasible dominated;
3. unfeasible non-dominated;
4. unfeasible dominated.

Dominated solutions are sorted in ascending order, according to the number of solutions that dominates them. unfeasible solutions are sorted in ascending order as well, but according to the value of constraint violation.

Select the best Ab's (i.e., all feasible non-dominated Ab; if the number of feasible non-dominated individuals is less than 10% of the population size, then select Ab's following the constrained-dominance hierarchy until reaching a number of individuals equal to the 10% of the population size) to be cloned (line 5). Copy the best Ab's obtained in the previous step into Pm (line 6). Entrance into memory is regulated using the adaptive grid. For each Ab selected in line 5, the constrained-dominance is verified against those that are already in the memory:

1. If the selected Ab is dominated by any Ab already present in memory, the new Ab is discarded;
2. All of the Ab's belonging to the memory that are dominated by the new Ab are removed. Then, the possibility of the new Ab composing the memory is verified:
   (a) If the memory is not full, the Ab is allowed to enter;
   (b) Otherwise, if the new Ab belongs to the most populated region, it is not allowed to enter;
   (c) Otherwise, it enters the memory, but an individual from the most populated cell is removed leaving space for the new Ab (the memory size is maintained).

For Ab's selected in line 5, clone them proportionally to their distance to the k-nearest neighbor, obtaining Pc (line 7). Hypermutate Pc inversely proportional to the hierarchy defined by constrained-dominance, i.e., feasible non-dominated Ab's suffer less mutation than hierarchically worse solutions and so on (line 8). At line 9, a uniform mutation is applied to Ab's that were not selected in line 5. At each set of X generations (here, $X = 25$), a number of new Ab's is generated and added to the main population (as a way of generating diversity, explore the objective space, and rescue the search from local maxima) (lines 10-14). Return the population P to its original size (select as many individuals as the original population size using the criterion of hierarchies taken by the constrained-dominance) (line 15). Repeat the process from line 3 until a stop criterion is achieved (here, number of evaluations $= 100,000$). At the end of execution, return Pm, the set of the best Ab's found.

D. Experiments

The objective of this study is two fold: (1) to present a new method to deal with the SCP problem of finding the minimal set of populations to represent a species genetic diversity; and (2) to verify if the use of additional objectives is somehow beneficial when dealing with the SCP problem.

1) Experiment 1 – Two Objectives Optimization: The purpose of this first experiment was to find the smallest set of *E. dysenterica* local populations that should be preserved in order to represent the species genetic diversity aiming its conservation. At least 95% of the 249 alleles should be represented at least once. We optimized Eqs. 3 and 4, subject to Eq. 7 constraint.

2) Experiment 2 – Three Objectives Optimization: The purpose of this experiment was to find the smallest set of *E. dysenterica* local population needed to represent at least 95% of the studied alleles, but optimizing the region stability as well. We optimized Eqs. 3–5, subject to Eq. 7 constraint.

3) Experiment 3 – Four Objectives Optimization: The purpose of this experiment was the same as the previouly stated experiments, but optimizing simultaneously region stability and habitat loss as additional objectives. We optimized Eqs. 3–6, subject to Eq. 7 constraint.

4) The null model: A null model was generated to find out if the same results would be found without the use of the MOO approach. We randomly generated solutions and computed their values for region stability, habitat loss, number of lacking alleles, and number of selected populations, following the computations performed by MAIS.

E. Experimental Setup

1) Aleatory Uncertainties: To determine the number of runs required to mitigate aleatory uncertainty inherent to the algorithm stochasticity , we used Spartan (Simulation Parameter Analysis R Toolkit Application) [16]. We analyzed 20 subsets sample sizes of 1, 5, 10, 50, and 100 runs each, requiring, therefore, 3,320 individual runs (for each experiment). For all experiments, it was found that 100 runs were sufficient to reduce the effect magnitude of aleatory uncertainty on results to less than "small" (the desired level) (Fig. 2).

2) Parameter Settings: MAIS parameter values were set to: pop. size $= 500$; secondary pop. size $= 500$; clone rate = inversely proportional to k-nearest neighbor; hypermutation rate = proportional to the hierarchy of constrained-dominance; uniform mutation = $1/L$ (where L is the number of populations); number of new Ab's created at each 25 generations $= 20\%$ of pop. size; number of fitness evaluations $= 100,000$.

3) Computer Infrastructure: The experiments were performed on a computer cluster consisting of 49 computers equipped with Intel(R) Core(TM) i5-2500 CPU 3.30GHz, 8GB RAM; and 9 computers with Intel(R) Core(TM) Duo CPU E7500 2.93GHz, 2GB RAM.

III. Results and Discussion

We performed *Experiment 1* aiming at selecting the smallest set of populations capable of representing the *E. dysenterica* genetic diversity (alleles). This also allowed us to establish a lower bound for Experiments 2 and 3.

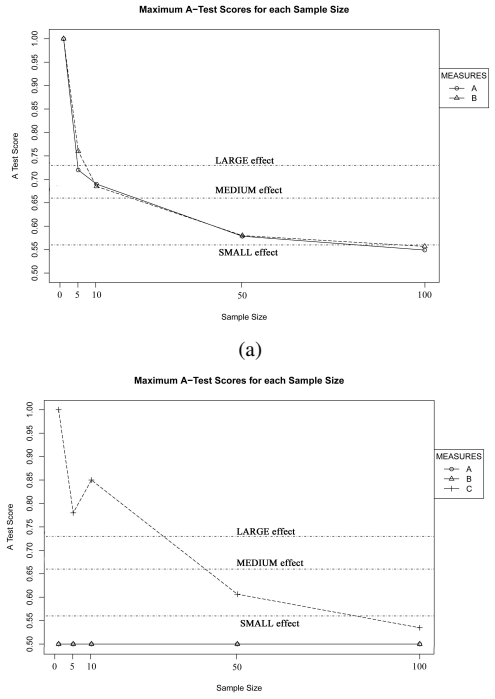


Fig. 2. Spartan's Technique 1 applied to: (a) Experiment 1, and (b) Experiment 2. This technique identifies which simulation results can be attributed to the dynamics of the modeled system, rather than artifacts of uncertainty or parametrization, or simulation stochasticity. At 100 runs, stochasticity over objectives attains a small effect. A = lacking alleles, B = number of selected populations, and C = stability.

Empirical attainment function (EAF) [5] is a quality indicator used for stochastic algorithmic evaluation. It is computed from the combined collection of approximation sets. Let $b_1(z)...b_n(z)$ be n runs of the optimizer, then the EAF is defined as $EAF : \mathbb{R}^d \longmapsto [0, 1]$ with

$$EAF = \frac{1}{n} \sum_{i=1}^{n} b_i(z) \tag{8}$$

It offers a useful description of the solution distribution location. Differences in the frequency with which certain goals are met by the respective algorithms are represented graphically. The intensity of the shading correspond to the frequency of the solution. The EAF surface for Experiment 1 can be seen in Fig. 3a. We found that the minimum number of sites required to represent all of the alleles was 15.

In *Experiment 2* and *Experiment 3*, the aim was to select the smallest set of populations capable of representing the most amount of alleles (preferably all of them), but at the same time optimizing additional objectives.

Similarly to Experiment 1, we found that it is possible to preserve the allele diversity (249 alleles) by selecting 15 populations (Fig. 4).

The portfolio of solutions increased significantly, which was expected, since it is known in the literature that as

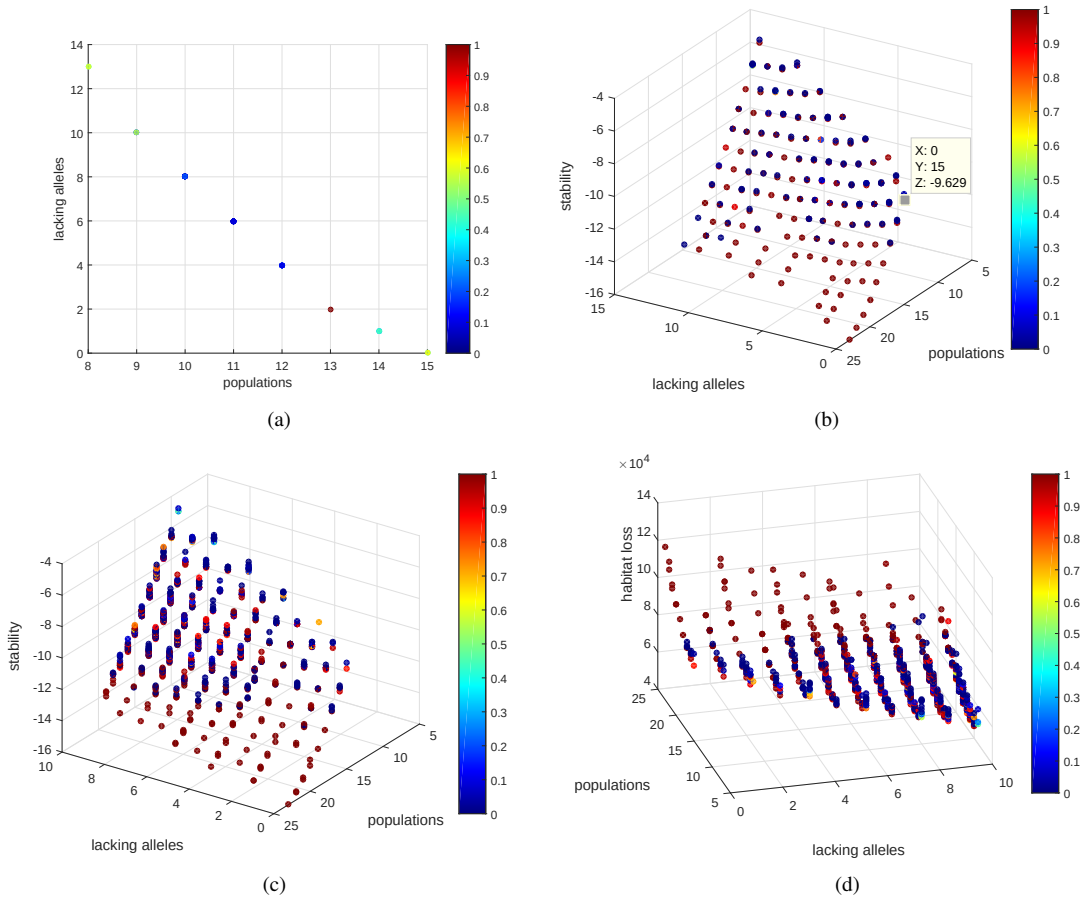


Fig. 3. EAF surfaces obtained after 100 independent runs for: (a) Experiment 1; (b) Experiment 2; and (c-d) Experiment 3. Experiments optimizing more objectives (b-d) were able to provide decision-maker with a more diversified portfolio of solutions distributed along the constrained objective space. Experiments 2 and 3 improved the results, refininig them.

the number of objectives increases, the number of solutions enlarges exponentially [5]. Almost all new combination of populations will give a different result with all alleles being represented, so it is included in the portfolio.

Even if the aim is to obtain a minimum set, this method identifies a portfolio of solutions, indicating sets with individuals that fulfill the objectives, providing decision-makers with additional alternatives for achieving their conservation targets (Fig. 3b-d). It is worth noting that there is no hierarchy among results, i.e., all of the solutions are equally good in the considered context and in the absence of additional preferences. The proposed method is important to define strategies to provide a set of genetically diverse material while selecting the most representative samples/populations.

For each experiment, the output of 100 individual runs were unified, and the frequency each population appeared in the union set was calculated. This frequency indicates the relative importance of a population in order to fulfill the optimization objectives. This frequency can be taken as an estimator of population *irreplaceability*, a measure that indicates the proportion a population contributes to the overall solution, e.g., if this measure converges to 1, the population tends to be irreplaceable, i.e., if it is lost, the conservation goal may not be accomplished (Fig. 5).

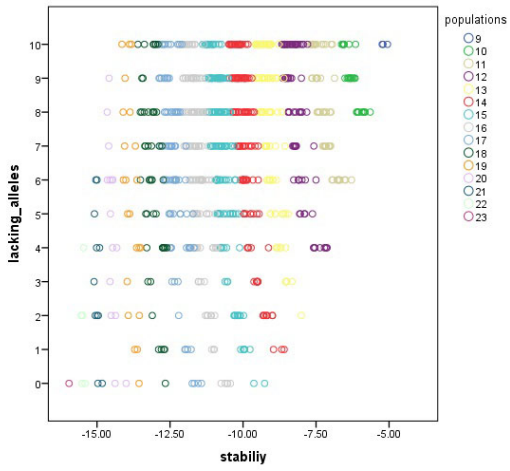


Fig. 4. Scatterplot for Experiment 3 evidencing the diversity of solutions. It can be seen that the minimum set representing all of the alleles has 15 populations, and representing at least 95% of alleles has 9 populations.

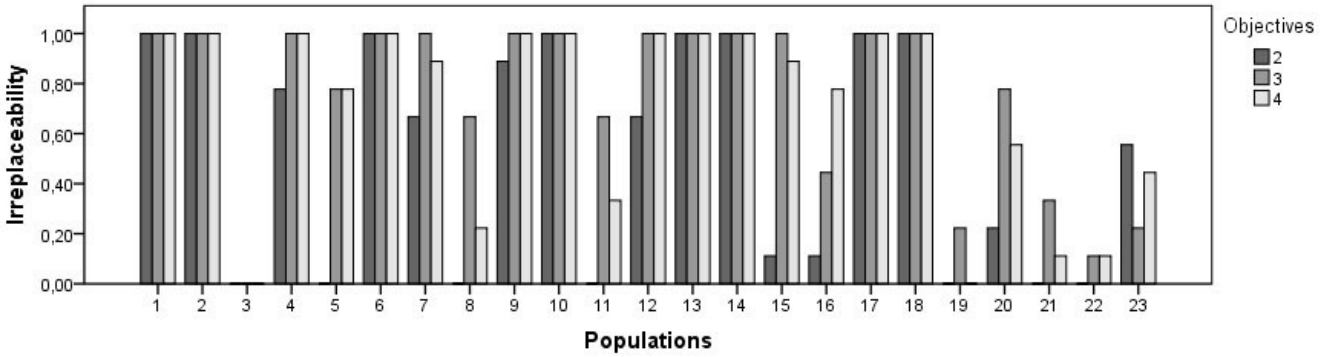


Fig. 5. Population irreplaceability for Experiment 1–3. Irreplaceability scales from 0 to 1. If this measure converges to 1, the population tends to be irreplaceable, if it lost, the conservation goal may not be accomplished. Populations presenting rare alleles tend to have higher irreplaceability scores.

There are 31 rare alleles that appear in only one population – they are distributed along 13 different populations –, and 3 rare alleles that appear in only two populations. The populations that have these alleles tend to be irreplaceable, hence, if at least one of them is not selected, the conservation goal may not be achieved. This is the case of populations 6, 10, 13, 17, and 18, the most frequent in solutions (associated to the presence of the rarest species, they have the greatest diversity of alleles), which can be noted comparing Fig. 5 to Fig. 1a.

In order to compare the coincidence in frequency of population selection, we converted the irreplaceability values of each experiment in percentage values and compared them in pairs (Fig. 6). On the one hand, Experiment 1 (optimization of 2 objectives) compared to Experiments 2 and 3 only showed some coincidence in less the 50% of population selection (Fig. 6a-b). On the other hand, optimization of 3 and 4 objectives were almost synergistic, in the sense that the frequence they selected the same populations was very similar (Fig. 6c), this is an evidence that the use of stability and habitat loss as additional objectives guide selection in the same direction; it worth note that they promoted a better spread of selection among populations (which has a positive impact in increasing diversity and persistence over time).

We verified that there was a significant improvement in the retention of alleles in selected populations in Experiments 2 and 3, suggesting that the use of additional objectives (in particular stability and habitat loss) is adequate to be applied for defining a conservation selection policy.

The proposed approach is straightforward. Once the matrix based on the molecular markers and the additional objective indexes are generated, there is no need of expertise to proceed the selection of populations.

The use of constraint allowed us to concentrate the exploration of the objective space in a more rational and efficient way, privileging feasible solutions over unfeasible ones, since the former are more valuable to decision-makers than the later.

In Eq. 2, experts defined $r_j = 1$, $j \in \{1, ..., 55\}$, i.e., all of the alleles should be represented at least once. Nevertheless, it could be settled a different representation level $r_j \geq n > 1$, where n is the minimum number of times an allele should be present in solution. By doing so, allele representation in selected populations is improved, benefiting persistence throughout time.

Comparing Fig. 5 to Fig. 1, it can be seen that regions with higher stability an less habitat loss were adequately prioritized using optimization of 3 and 4 objective, e.g., population #23, although it has higher number of distinct alleles, it was less frequently selected in Experiments 2 and 3 than in Experiment 1 since it has low stability and high habitat loss values.

The current study makes a significant contribution by applying a multi-objective optimization method to a real-world problem, revealing important relationships among objectives that are common to conservation scenarios of a practical SCP problem.

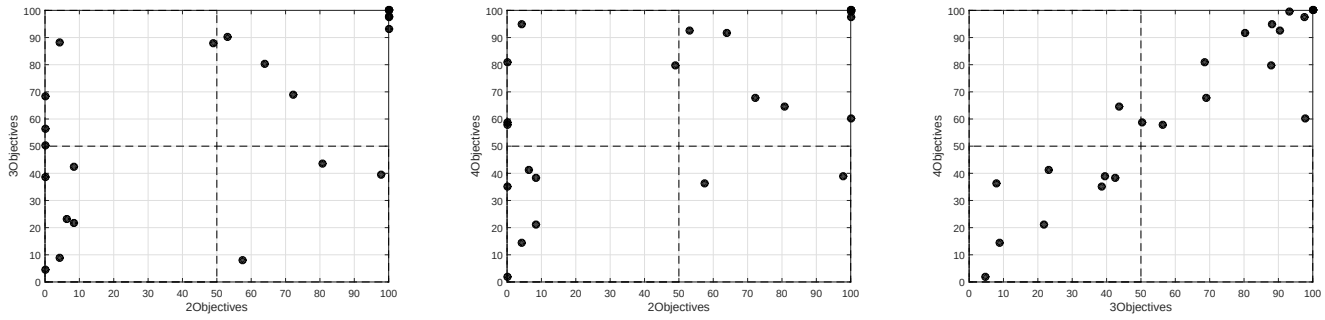
In the context of the case study carried out, most *E. dysenterica* diversity is found in nature, and many such populations are increasingly threatened by habitat reduction. In nature, there are potential useful populations of *E. dysenterica*, yet, for practical purposes, only a fraction of this material can be afforded conservation in protected areas. Since some of these areas can be considered not stable or have habitat loss that can compromise the persistence of the species, the proposed approach can be used to select samples for *ex situ* conservation, e.g., maintenance in germplasm collections[4], with the advantage that previous approaches generally indicate a population to be preserved, the proposed method can be applied using a higher level of granularity, i.e., at individuals level, indicating exactly which individual within the population should be sampled for or kept in the germplasm collection.

Null model randomly generated solutions had worse values compared to solutions found with MAIS, thus, assuring that solutions obtained with MAIS have not emerged by chance.

IV. Conclusions and Future Work

The SCP problem for biodiversity conservation is NP-hard, nevertheless, decision-makers still need a solution for it. This

[4]collections of living tissues from which new plants can be grown, they play a significant role among strategies for conservation once they preserve the diversity of common, rare, threatened, or endangered species, which are made available for further study or habitat restoration projects [10]



(a) (b) (c)

Fig. 6. Bi-dimensional comparison between frequency of population selection in the experiments. The frequency was converted in percentage for graph visualization purposes. (a) Experiment 1 (2 objectives) vs Experiment 2 (3 objectives); (b) Experiment 1 vs Experiment 3 (4 objectives); (c) Experiment 2 vs Experiment 3. Reference lines are displayed at 50% of each axis. Similar values (both axis with high values or both axis with low values) are placed in 1st and 3rd quadrants; likewise, dissimilar values are placed in 2nd and 4th quadrants. Experiment 1 compared to Experiments 2 and 3 only showed some coincidence in less the 50% of population selection, while Experiments 2 and 3 were almost synergistic, in the sense that the frequency they selected the same populations was very similar (placed in 1st and 3rd quadrants).

is a relevant problem in ecology with real impact on resources availability (whether financial or genetic).

This paper is pioneer in showing how principles of SCP and the use of a constrained multi-objective AIS algorithm associated to molecular marker and habitat information can be applied to successfully help select minimum local populations sets having maximum allelic richness.

Having established that the proposed approach is viable, a future work is to apply it to other kinds of molecular markers (e.g, single nucleotide polymorphism – SNP, and diversity arrays technology – DArT) in order to verify its feasibility for those kind of data.

We did not incorporate domain information in MAIS, so there is still room for improvement. MAIS uses standard operators for hypermutation and cloning. Our future work will focus on the use of more complex operators (e.g., contiguous hypermutation, aging), which have shown interesting results in the literature [17], [18] and have not reported use with multi-objective approaches.

ACKNOWLEDGMENT

SS wishes to thank Alex Schlottfeldt for the invaluable aid in running the experiments and the support from CNPq throughout a Science without Borders scholarship. GENPAC has been supported by CNPq/MCT/CAPES. Work by MEMTW, ACPLFC, MPCT, RDL and JAFDF have been continuously supported by productivity fellowships from CNPq.

REFERENCES

[1] CDB. Convention on Biological Diversity Site. [Online]. Available: http://www.cbd.int

[2] C. P. Gomes, "Computational Sustainability: Computational methods for a sustainable environment, economy, and society," *The Bridge*, vol. 39, pp. 5–13, 2009.

[3] R. Reddy, *Ecoinformatics: Tools and Techniques*. SBS Pub. & Dist., 2009.

[4] M. Cabeza and A. Moilanen, "Design of Reserve Networks and the Persistence of Biodiversity," *Trends Ecol Evol*, vol. 16, no. 5, pp. 242–248, 2001.

[5] C. A. Coello Coello, G. B. Lamont, and D. A. Van Veldhuizen, *Evolutionary Algorithms for Solving Multi-Objective Problems*, 2nd ed., D. E. Goldberg and J. R. Koza, Eds. New York: Springer-Verlag, 2007, ISBN 978-0-387-33254-3.

[6] S. Schlottfeldt, J. Timmis, M. E. M. T. Walter, A. C. P. L. F. Carvalho, J. A. F. Diniz-Filho, L. M. Simon, R. D. Loyola, and M. P. C. Telles, "A Multi-Objective Optimization Approach Associated to Climate Change Analysis to Improve Systematic Conservation Planning," in *Proceedings of the 8th International Conference on Evolutionary Multi-Criterion Optimization, EMO'2015, Part II*, ser. Lect Notes Comput Sc, A. Gaspar-Cunha, A. Henggeler, and C. Coello, Eds., vol. 9019, 2015, p. 591.

[7] S. Sarkar, "Complementarity and the Selection of Nature Reserves: Algorithms and the Origins of Conservation Planning, 1980-1995," *Arch Hist Exact Sci*, vol. 66, pp. 397–426, 2012.

[8] R. L. Pressey, H. P. Possingham, and C. R. Margules, "Optimality in Reserve Selection Algorithms: When Does it Matter and How Much?" *Biol Conserv*, vol. 76, no. 3, pp. 259–267, 1996.

[9] J. A. F. Diniz-Filho *et al.*, "Planning for Optimal Conservation Geographical Genetic Variability within Species," *Conserv Genet*, vol. 13, pp. 1085–1093, 2012.

[10] S. Schlottfeldt, M. E. M. Walter, A. C. P. L. F. Carvalho, T. N. Soares, M. P. C. Telles, R. D. Loyola, and J. A. F. Diniz-Filho, "Multi-objective optimization for plant germplasm collection conservation of genetic resources based on molecular variability," *Tree Genet Genomes*, vol. 11, no. 2, 2015. [Online]. Available: http://dx.doi.org/10.1007/s11295-015-0836-3

[11] S. Schlottfeldt, J. Timmis, M. E. M. T. Walter, A. C. P. L. F. Carvalho, J. A. F. Diniz-Filho, L. M. Simon, R. D. Loyola, and M. P. C. Telles, "Multi-objective Optimization Applied to Systematic Conservation Planning and Spatial Conservation Priorities under Climate Change," in *Proceedings of the 2014 Conf. on Genetic and Evolutionary Computation Companion*, ser. GECCO Comp'14. New York, NY, USA: ACM, 2014, pp. 177–178.

[12] K. Deb, S. Agarwal, A. Pratap, and T. Meyarivan, "A Fast Elitist Non-Dominated Sorting Genetic Algorithm for Multi-Objective Optimization: NSGA-II," in *Proceedings of the 6th International Conference on Parallel Problem Solving from Nature*, ser. PPSN VI. London: Springer-Verlag, 2000, pp. 849–858.

[13] C. A. Coello Coello and N. C. Cortés, "Solving Multiobjective Optimization Problems Using an Artificial Immune System," *Genet Program Evol M*, vol. 6, no. 2, pp. 163–190, 2005.

[14] L. N. de Castro and J. Timmis, *Artificial Immune Systems: A New Computational Intelligence Approach*. Springer-Verlag, New York, 2002.

[15] J. D. Knowles and D. W. Corne, "Properties of an Adaptive Archiving Algorithm for Storing Nondominated Vectors," *IEEE T Evolut Comput*, vol. 7, no. 2, pp. 100–116, 2003.

[16] K. Alden, M. Read, J. Timmis, P. S. Andrews, H. Veiga-Fernandes, and M. Coles, "Spartan: A Comprehensive Tool for Understanding Uncertainty in Simulations of Biological Systems," *PLoS Comput Biol*, vol. 9, no. 2, pp. e1 002 916+, 2013.

[17] T. Jansen and C. Zarges, "Analyzing Different Variants of Immune Inspired Somatic Contiguous Hypermutations," *Theor Comput Sci*, vol. 412, pp. 517–533, 2011.

[18] P. S. Oliveto and D. Sudholt, "On the Runtime Analysis of Stochastic Ageing Mechanisms," in *Proceedings of the 2014 Conference on Genetic and Evolutionary Computation*, ser. GECCO '14. New York, NY, USA: ACM, 2014, pp. 113–120.

Capítulo 9

Conclusões e Trabalhos Futuros

9.1 Conclusões

Nesta tese:

- Aplicou-se conceitos de MOO ao problema SCP, o que permitiu trabalhar com instâncias do problema SCP com mais de duas dimensões, possibilitando maior flexibilidade pela inclusão de objetivos adicionais, bem como acrescentando mais complexidade e aumentando, assim, o poder de decisão do método computacional;
- Utilizou-se um algoritmo inspirado em AIS que foi capaz de encontrar o menor conjunto de populações locais e o menor conjunto de indivíduos que deveriam ser conservados para representar a diversidade genética da espécie, tomando por base informação alélica proveniente de análise molecular em nível populacional como unidade básica de investigação;
- O método proposto foi estendido, incorporando análise dinâmica de biodiversidade para prover os decisores com informação acerca da projeção das decisões atuais de conservação face a cenários futuros de mudança climática, possibilitando rever tais decisões no presente com base em uma decisão informada.

  Além disso, com a abordagem utilizada, foi possível identificar as relações existentes entre objetivos de maneira a identificar quais objetivos são conflitantes, quais são os mais importantes, quais são necessários e quais são redundantes. Por exemplo, no problema da análise dinâmica da biodiversidade, pôde-se ratificar o conflito entre remanescentes de vegetação (VR) e ocupação humana (H_O), bem como a identificação da ausência de conflito entre VR e evapotranspiração anual (AET) [206].

9.2 Contribuições

Até onde sabemos, esta tese foi pioneira em:

- aplicar MOO ao problema SCP usando alelos de análise molecular em nível populacional como unidade básica de análise;
- aplicar MOO associado à análise de mudanças climáticas ao problema SCP;
- aplicar AIS para resolver o problema SCP;
- resolver um problema de Ecologia usando métodos computacionais e dados de espécies nativas, apoiando a implantação, no Brasil, da denominada área de Ecoinformática.

Os principais resultados obtidos constam de **sete *papers***:

- **Cinco apresentados e publicados em congressos**:

  1. *Multi-Objective Optimization in Systematic Conservation Planning to Represent Genetic Variability within Species. 8th International Conference on Ecological Informatics*, ISEI'2012;
  2. *Multi-Objective Optimization Applied to Systematic Conservation Planning and Spatial Conservation Priorities Under Climate Change. 2014 Conference Companion on Genetic and Evolutionary Computation Companion*, GECCO'2014. **Estrato Qualis A1**;
  3. *A Multi-Objective Optimization Approach Associated to Climate Change Scenario to Improve Systematic Conservation Planning and Spatial Conservation Priorities Setting. 8th International Conference on Evolutionary Multi-Criterion Optimization*, EMO'2015. **Estrato Qualis A2**;
  4. *Challenges in Computational Sustainability: a New Multi-Objective Artificial Immune System Approach to Deal with Biodiversity Conservation. Workshop on Evolutionary Multi-Objective Optimization at the IEEE 2015 International Congress on Evolutionary Computation*, CEC'2015. **Estrato Qualis A2**;
  5. *Using Multi-Objective Artificial Immune Systems to Find Core Collections Based on Molecular Markers. 2015 Conference on Genetic and Evolutionary Computation*, GECCO'2015. **Estrato Qualis A1**.

- E **dois publicados em periódicos**:

    1. *Multi-Objective Optimization for Plant Germplasm Collection Conservation of Genetic Resources Based on Molecular Variability. Tree Genetics & Genomes.* Fator de Impacto 2.435, **Estrato Qualis A2**.
    2. *Multiobjective Optimization for Conservation of Genetic Resources Based on Molecular Variability. Genetics and Molecular Research.* **Estrato Qualis B2**.

Os resultados inéditos apresentados no Capítulo 7 ainda serão submetidos a publicação.

Quando a autora iniciou seu Doutorado, a área específica de Ecoinformática era incipiente, sobretudo no Brasil. Quase não havia produção e mesmo chamadas em conferências de Ciência da Computação para trabalhos na área, tampouco havia periódicos específicos, motivo pelo qual *papers* que integram esta tese foram submetidos a revistas em **estratos superiores do Qualis Capes com enfoque em Biodiversidade e Interdisciplinaridade**. Em que pese a Ecoinformática ainda não ter atingido o patamar que áreas multidisciplinares (e.g., Bioinformática) já alcançaram, ao longo do desenvolvimento desta tese, as primeiras chamadas específicas para Ecologia em conferências de Ciência da Computação começaram a ser feitas. É o caso, no Brasil, do *Workshop de Computação Aplicada à Gestão do Meio Ambiente e Recursos (WCAMA)*, evento satélite do *Congresso da Sociedade Brasileira de Computação (CSBC)*, em sua VI edição em 2015.

O presente trabalho foi iniciado no contexto da rede GENPAC (Genética Geográfica e Planejamento Regional para Conservação de Recursos Naturais do Cerrado), tendo esse projeto previsão de finalização em dezembro de 2015. O GENPAC foi financiada pelo MCTI-CNPq-FAPs com **recursos da ordem de R$4.500.000,00 (quatro milhões e quinhentos mil reais)**.

Há que se destacar que o presente trabalho permitiu a entrada da Universidade de Brasília (UnB) na proposta de constituição do Instituto Nacional de Ciência & Tecnologia (INCT) em Ecologia, Evolução e Conservação da Biodiversidade (EECBIO) que visa reunir especialistas nas principais áreas de pesquisa em biodiversidade, consolidando uma rede de pesquisa e de formação de recursos de excelência na área.

O EECBIO tem previsão de gerir **recursos da ordem de R$10.000.000,00 (dez milhões de reais)** ao longo de seis anos e será uma referência internacional em análise da biodiversidade e sua conservação, com um forte componente teórico e metodológico, mostrando a importância de interação entre pesquisadores de diferentes áreas a fim de inovar e definir novas direções de pesquisa e intervenção.

9.3 Trabalhos Futuros

1. **Abordagem numerosos-objetivos**. É sabido de uma série de estudos empíricos que os algoritmos que têm sido utilizados para MOO, como NSGA-II e SPEA2, têm dificuldade em obter frentes próximas à de Pareto se o número de objetivos é grande [26]. Desenvolver algoritmos que são eficientes e encontram boas aproximações à frente de Pareto, inclusive com grande quantidade de objetivos, é o foco da otimização *many-objective* [1] [26] (aqui livremente traduzida para *numerosos-objetivos*[2]). Os casos tratados nesta tese são discretos, porém o maior impacto da quantidade de objetivos se sente em otimizações com dados contínuos [88]. Seria interessante investigar qual o comportamento de MAIS frente a numerosos-objetivos (e.g., Cenário 2 do Problema 3 [206]) e neste caso testar seu desempenho com o recém-proposto NSGA-III [88, 128].

2. **Otimização em dois níveis (*bilevel optimization*)** [90, 214]. Na *bilevel optimization*, a tarefa de otimização é feita em dois níveis. Os problemas resolvidos com tal tipo de otimização contêm, geralmente, uma otimização aninhada a uma restrição de um outro problema de otimização. A otimização dita "externa", é referida como tarefa de alto nível (*upper level task*) e o problema de otimização "interno" é referido como tarefa de baixo nível (*lower level task*). A estrutura aninhada do problema geral requer que a solução do problema da alto nível seja factível, apena se ela é uma solução ótima para o problema da baixo nível. Nos estudos de caso desenvolvidos nesta tese, em que pese a inclusão de objetivos adicionais ter trazido benefícios à análise e ter representado um avanço em relação aos métodos tradicionais empregados, observou-se que dentre as soluções encontradas, ainda que por premissa não houvesse prioridade entre os objetivos (sendo todos considerados igualmente importantes), na prática, havia uma maior relevância conferida à minimização dos alelos faltantes. Seria interessante testar se a abordagem *bilevel optimization* é aplicável. Tomando-se o Problema 2 como exemplo, entende-se que a otimização dos alelos e dos indivíduos comporiam a tarefa de baixo nível, e a otimização dos demais objetivos (hetereozigose, HWE, frequência de alelos) faria parte da tarefa de alto nível.

3. **Escalabilidade e *big data***. Há diversos contextos em que SCP lida com um volume de dados ainda maior do que os trabalhados nesta tese, os chamados *big*

[1] Monobjetivo: 1 objetivo; multiobjetivo: 2 a 3 objetivos; numerosos-objetivos: mais de 3 objetivos [88].

[2] Optou-se pela tradução numerosos-objetivos, pois segundo o Novo Dicionário Eletrônico Aurélio [121] muitos = em excesso (o que não é o caso); bastante = que basta, é suficiente (o que não é o caso); abundante = demasiado, supérfluo (o que não é o caso), numeroso = grande número (o que se considerou mais adequado).

data. Além disso, independentemente do algoritmo MOO utilizado, a quantidade de execuções empreendidas envolve um esforço computacional considerável. Junte-se a isso o fato de que as soluções são conjuntos de vetores, que muitas vezes, para serem utilizados como informação útil, necessitam de um pós-processamento a fim de apresentar os resultados aos DM de uma maneira que permita a tomada de decisão. Estudar maneiras de representar internamente os dados, processá-los, tratá-los e apresentá-los ao decisor são desafios que continuam em aberto.

É o caso do estudo que se pretende empreender a partir dos resultados obtidos nesta tese, de trabalhar dados da distribuição geográfica global (no mundo todo) de espécies de mamíferos não marinhos e conflitos de conservação envolvendo produção agrícola, integração política (países, blocos econômicos) e diminuição de probreza[3].

No presente estudo, o cálculo do *fitness* não representou um gargalo, mas uma ideia que pode mostrar-se interessante ao lidar-se com grande volume de dados é a possibilidade de estimar o valor de *fitness* em vez de calculá-lo efetivamente.

4. **Estudar formas de apresentar os resultados**. É intutitivo pensar que representar superfícies resultantes de otimizações multiobjetivo (ou com numerosos-objetivos), onde se tem um grande número de dimensões, é um desafio [88]. O simples fato de não ser uma otimização monobjetivo implica que múltiplos pontos serão gerados, o que representa uma maior quantidade de opções para os decisores. Ao mesmo tempo que isso é uma vantagem, pode causar dificuldades para o decisor quanto à escolha de uma solução.

   Associada à própria dificuldade de apresentação dos dados inerente à abordagem multi-objetivo, a apresentação dos resultados corresponde a um conhecimento a ser construído conjuntamente com os biólogos dado que, muitas vezes, pela novidade na utilização do método, o próprio significado dos achados não está claro para os especialistas, e.g., a posição que uma solução ocupa ao longo da frente de Pareto podem ser o indicativo de estratégias diferentes de conservação (*in situ vesus ex situ*), mas não se sabe ainda ao certo. Assim, conforme citado no item anterior, desenvolver formas de apresentação da informação de maneira a facilitar a interpretação e utilização dos dados gerados é um desafio em aberto.

5. **Comparações com algoritimos monobjetivo**. Verificou-se que a abordagem MOO era viável e permitia não apenas lidar com dados de uma granularidade mais fina (em nível molecular), mas inserir objetivos antes não contemplados na abordagem monobjetivo, propiciando uma flexibilidade que as técnicas utilizadas até então

[3]Os dados foram gentilmente cedidos pelo Prof. Dr. Ricardo Dobrovolski, do Departamento de Zoologia da Universidade Federal da Bahia (UFBA).

não permitiam. Apesar de terem sido feitas comparações entre o algoritmo proposto nesta tese (MAIS) e algoritmos representantes do estado-da-arte em MOO (NSGA-II e SPEA2), não foram empreendidas comparações entre a abordagem MOO e a monobjetivo. Seria interessante comparar os resultados obtidos entre tais abordagens.

6. **Calibragem (*tunning*) de algoritmos**. Em que pesem as comparações entre os algoritmos terem levado em consideração configurações de parâmetros determinadas tanto empiricamente como com suporte na literatura [48, 89], há uma vertente de pesquisa que aponta não ser justo, em termos de comparação de desempenho, usar parâmetros comuns entre diferentes algoritmos. Segundo essa linha, dever-se-ia fazer a calibragem de cada um dos algoritmos a fim de ter dados que permitissem compará-los de uma maneira adequada [237]. Apesar de ser importante considerar o custo-benefício de se fazer tal calibragem em termos de ganho de desempenho, dispêndio de recursos computacionais e resultados, sugere-se que, a depender do objetivo do experimento, este seria um ponto a ser investigado.

7. **Configuração automática dos parâmetros do algoritmo**. Esta é uma área de pesquisa que está crescendo e vem apresentando resultados interessantes. A identificação da configuração ideal de parâmetros de um algoritmo é parte do desenvolvimento da sua aplicação em uma dada classe de instâncias de problemas [124]. Nesta tese, um estudo da configuração de parâmetros para MAIS foi feita empregando-se as Técnicas 2 e 3 de *Spartan*. Mas ainda assim, todo o processo foi manual. Tendo em mente o *trade-off* apresentado no item anterior, investigar a aplicação de um processo automático de configuração de parâmetros aos problemas tratados nesta tese pode ser útil. Uma abordagem intermediária corresponderia ao ajuste de parâmetros ao longo da execução do algoritmo.

8. **Aperfeiçoar MAIS**. Não era uma preocupação, neste primeiro momento, tornar o algoritmo mais eficiente, mas testar sua efetividade nos problemas tratados e observar seu comportamento, suas funcionalidades. Não foi incorporado conhecimento específico do problema no algoritmo, de maneira que ainda há espaço para melhora:

   (a) MAIS usa operadores padrões para hipermutação e clonagem. Seria interessante usar operadores mais complexos (e.g., *contiguous hypermutation*, *aging*, etc.), que mostraram resultados interessantes na literatura [131, 170], mas que até então não têm registro na literatura de uso com abordagens MOO;

   (b) adaptar MAIS para manter a diversidade no espaço de busca e não apenas no espaço de objetivos, a fim de permitir que soluções geneticamente diferen-

tes, mas que dão origem a valores iguais para as funções objetivo coexistam. Nenhum dos algoritmos utilizados nesta tese permite tal situação;

(c) Testar novas formas de inicialização dos dados, como inicializar todos Ab em zero em vez de valores aleatórios. Considerando-se o Problema 2, de um universo de 642 opções em que o mínimo encontrado tem 9 indivíduos [207], seria melhor partir, desde logo, de um valor menor de indivíduos selecionados, lembrando que devem ser envidados esforços para evitar a convergência prematura;

(d) Testar o conceito de ortogonalidade entre os objetivos. Em um espaço vetorial de n dimensões, pode-se escolher conjuntos de n vetores de maneira que cada par de vetores é um par de vetores perpendiculares. Como graficamente nos problemas tratados, o formato da PF_{known} está bastante relacionado à minimização da distância à origem dos eixos, o conceito de ortogonalidade pode ter uma aplicação interessante ao guiar a otimização.

9. **Combinar técnicas de busca**. Em muitos problemas de otimização, pontos na PF_{known} estão agrupados em determinadas regiões do espaço de objetivos. É possível direcionar computacionalmente tais pontos usando mecanimos que explorem certas propriedades do espaço de busca, por exemplo, usando um ou alguns poucos objetivos, é possível utilizar uma técnica de busca local para mover um ponto para mais próximo da PF_{true} e com isso conseguir uma melhor distribuição de pontos em PF_{known}. Abordagens para a busca local no espaço de decisão poderiam ser *depth-first search (hill-climbing), simulated annealing* ou busca tabu. A combinação de um algoritmo de busca global com técnicas de busca local dá origem aos chamados algoritmos *híbridos* ou *meméticos* [48].

10. **Estudar o número de gerações necessárias para a convergência de MAIS**. No que diz respeito ao número de gerações necessárias para se atingir a convergência, Cutello et al. [62] demonstraram um limite superior para garantir a visita do AIS ao ótimo global em t gerações com probabilidade δ. Há que se destacar que o limite encontrado depende de fatores como a cardinalidade do alfabeto K, do comprimento das soluções γ, do tamanho da população n. Os autores encontraram para um alfabeto binário ($K = 2$), $\gamma = 100$, $n = 1.000$ e $\delta = 0.9$, o valor $t = 10^{30} \approxeq 2^{100}$. Por sua vez, Clark, Hone e Timmis [44], mostraram, para um modelo específico de AIS, que em $t = 2^{20}$ gerações a probabilidade de se obter uma solução ótima é de ~ 0.85; sendo que com $t = 2^{40}$ gerações, a probabilidade cresceu para 1. Seria interessante empreender estudo para encontrar um limite superior para o número de gerações necessária para garantir a visita de MAIS ao ótimo global sob a probabilidade δ. Tal estudo seria relevante visto que, neste estudo, a terminação do

algoritmo foi determinada em termos do número de avaliações das funções objetivo (para efeito prático de comparação com os demais algoritmos) e não de convergência do algoritmo.

11. **Disponibilizar o método para uso público**. Em especial, com uma *interface* amigável, de fácil utilização, preferencialmente em ambiente *web*, a fim de que não seja necessário o suporte de um profissional de computação para a execução dos experimentos pelos biólogos.

Referências

[1] Dicionário Priberam da Língua Portuguesa. online, 2008–2013. Disponível em <http://www.priberam.pt/dlpo/>. Acessado em jun. 2015. 72

[2] Decision Point: Special Marxan Issue. Online, October 2010. Australia. 43, 45, 49

[3] A. K. Abbas and A. H. Lichtman. *Cellular and Molecular Immunology*. Sauders-Elsevier, USA, 2002. 73

[4] R. Abell. Conservation Biology for the Biodiversity Crisis: A Freshwater Follow-Up. *Conserv Biol*, 16(5):1435–1437, 2002. 1

[5] P. R. Ackery and R. I. Vane-Wright. *Milkweed Butterflies, Their Cladistics and Biology: Being an Account of the Natural History of the Danainae, a Subfamily of the Lepidoptera, Nymphalidae*. Publication (British Museum (Natural History))). British Museum (Natural History), 1984. 49

[6] K. Alden, P. S. Andrews, H. Veiga-Fernandes, J. Timmis, and M. Coles. Utilising a Simulation Platform to Understand the Effect of Domain Model Assumptions. *Nat Comput*, pages 1–9, 2014. 108

[7] K. Alden, M. Read, J. Timmis, P. S. Andrews, H. Veiga-Fernandes, and M. Coles. Spartan: A Comprehensive Tool for Understanding Uncertainty in Simulations of Biological Systems. *PLoS Comput Biol*, 9(2):e1002916+, February 2013. 108, 109, 113, 114, 140

[8] K. Alden, J. Timmis, and M. C. Coles. Easing Parameter Sensitivity Analysis of Netlogo Simulations Using Spartan. In *Proceedings of the 14th International Conference on the Synthesis and Simulation of Living Systems*. MIT Press, 2014. 108

[9] S. J. Andelman and W. F. Fagan. Umbrellas and Flagships: Efficient Conservation Surrogates or Expensive Mistakes? *P Natl Acad Sci USA*, 97(11):5954–5959, May 2000. 15

[10] P. S. Andrews, S. Stepney, and J. Timmis. Simulation as a Scientific Instrument. In *Proceedings of CoSMoS 2012*, pages 1–10. Luniver Press, 2012. 59

[11] J. A. Ardron, H. P. Possingham, and C. J. Klein, editors. *Marxan Good Practices Handbook*. Pacific Marine Analysis and Research Association (PacMARA), Victoria, BC, Canada, July 2010. 44, 49

[12] J. F. Arthur, M. Hachey, K. Sahr, M. Huso, and A. R. Kiester. Finding all Optimal Solutions to the Reserve Site Selection Problem: Formulation and Computational Analysis. *Environ Ecol Stat*, 4:153–165, 1997. 19, 21, 24, 49

[13] T. Bäck. *Evolutionary Algorithms in Theory and Practice: Evolution Strategies, Evolutionary Programming, Genetic Algorithms*. Oxford University Press, Oxford, UK, 1996. 54, 59, 60

[14] T. Baeck, D. B. Fogel, and Z. Michalewicz. *Evolutionary Computation 1: Basic Algorithms and Operators (Evolutionary Computation)*. TF-TAYLOR, May 2000. 62

[15] I. R. Ball. *Mathematical Applications for Conservation Ecology: The Dynamics of Tree Hollows and the Design of Nature Reserves*. PhD thesis, University of Adelaide, Depts. of Applied Mathematics, Environmental Science and Management, 2000. 14, 23, 24, 26, 44, 49

[16] Ian R. Ball, Hugh P. Possingham, and Matthew E. Watts. *Spatial Conservation Prioritization: Quantitative Methods and Computational Tools*, chapter Marxan and Relatives: Software for Spatial Conservation Prioritization, pages 185–195. Oxford University Press, 2009. 42

[17] A. Banks, J. Vincent, and C. Anyakoha. A Review of Particle Swarm Optimization. II: Hybridisation, Combinatorial, Multicriteria and Constrained Optimization, and Indicative Applications. *Nat Comput*, 7(1):109–124, March 2008. 4

[18] B. Barreto, G. Oliveira, M. Pinto, L. Bini, J. Diniz-Filho, and D. Blamires. Riqueza de Espécies de Emberizídeos e Conflitos de Conservação no Cerrado Brasileiro. *Acta Sci Biol Sci*, 30(1), 2008. 14

[19] D. N. Barton, G. Rusch, J. O. Gjershaug, D. P. Faith, and L. Paniagua. TARGET as a Tool for Prioritising Biodiversity Conservation Payments on Private Land - a Sensitivity Analysis. Technical Report SNR 4859-2004, Norwegian Institute for Water Reserach (NIVA), 2004. 42, 49

[20] A. Belaj, M. del C. Dominguez-Garcia, S. G. Atienza, N. M. Urdiroz, R. de la Rosa, Z. Satovic, A. Martin, A. Kilian, I. Trujillo, V. Valpuesta, and C. Del Rio. Core Collection of Olive (*Olea europaea L.*) Based on Molecular Markers (DArTs, SSRs, SNPs) and Agronomic Traits. *Tree Genet Genomes*, 8(2):365–378, April 2012. 106

[21] Biogeography & Conservation Lab of the Natural History Museum of London - NHM. Biodiversity and WorldMap: Measuring the variety of nature & selecting priority areas for conservation. Online, 2012. Disponível em <http://www.nhm.ac.uk/research-curation/research/projects/worldmap/>. Acessado em set. 2012. 42, 43

[22] S. L.. Bonfim. Viabilidade Econômica-Financeira de Extração Sustentada de Múltiplos Produtos em Floresta Estacional Semidecídua Secundária na Microrregião do Entorno de Brasília. Master's thesis, Departamento de Engenharia Florestal, Universidade de Brasília (UnB), 2010. 6

[23] L. Bradstreet. *The Hypervolume Indicator for Multi-Objective Optimisation: Calculation and Use*. PhD thesis, The University of Western Australia, 2011. 120

[24] Brasil. Lei nº 12.651, de 25 de maio de 2012. Disponível em https://www.planalto.gov.br/ccivil_03/_ato2011-2014/2012/lei/l12651.htm. Acessada em dez. 2014. 5

[25] K. Bringmann and T. Friedrich. Approximating the Volume of Unions and Intersections of High-Dimensional Geometric Objects. *Comput Geom Theory Appl*, 43(6-7):601–610, August 2010. 120

[26] D. Brockhoff. *Many-Objective Optimization and Hypervolume-Based Search*. PhD thesis, ETH Zurich, 2009. 120, 237

[27] T. Brooks, G. A. B. da Fonseca, and A. S. L. Rodrigues. Species, Data and Conservation Planning. *Conserv Biol*, 18(6):1682–1688, 2004. 8

[28] T. M. et al. Brooks, R. A. Mittermeier, da G. A. B. Fonseca, J. Gerlach, M. Hoffmann, J. F. Lamoreux, C. G. Mittermeier, J. D. Pilgrim, and A. S. L. Rodrigues. Global Biodiversity Conservation Priorities. *Science*, 313(5783):58–61, July 2006. 1, 2

[29] A.H.D. Brown. Core Collections: a Practical Approach to Genetic Resources Management. *Genome*, 31(2):818–824, 1989. 106, 107

[30] M. Cabeza and A. Moilanen. Design of Reserve Networks and the Persistence of Biodiversity. *Trends Ecol Evol*, 16(5):242–248, May 2001. 3, 16, 18

[31] J. D. Camm, S. Polasky, A. Solow, and B. Csuti. A Note on Optimal Algorithms for Reserve Site Selection. *Biol Conserv*, 78(3):353–355, 1996. 19, 22, 49

[32] F. Campelo, F. G. Guimar aes, R. R. Saldanha, H. Igarashi, S. Noguchi, D. A. Lowther, and J. A. Ramirez. A Novel Multiobjective Immune Algorithm Using Nondominated Sorting. In *11th International IGTE Symposium on Numerical Field Calculation in Electrical Engineering*, Seggauberg, Austria, 2004. 81, 83

[33] L. R. Carrazza and J. C. C. D'Ávila. *Manual Tecnológico de Aproveitamento Integral do Fruto do Baru (Dipteryx alata)*. Instituto Sociedade, População e Natureza (ISPN), 2010. 5, 100, 101, 102

[34] P. A. D. Castro. *Sinergia entre Sistemas Imunológicos Artificiais e o Modelos Gráficos Probabilísticos*. PhD thesis, Departamento de Engenharia de Computação e Automação Industrial (DCA), Faculdade de Engenharia Elétrica e de Computação (FEEC), Universidade Estadual de Campinas (Unicamp), 2009. 81, 83, 84

[35] P. A. D. Castro and F. J. Von Zuben. MOBAIS: A Bayesian Artificial Immune System for Multi-Objective Optimization. In Peter J. Bentley, Doheon Lee, and Sungwon Jung, editors, *Proceedings of the 7th International Conference on Artificial Immune Systems, ICARIS'08*, volume 5132 of *Lect Notes Comput Sc*, pages 48–59, Phuket, Thailand, 10-13 August 2008. Springer-Verlag. 81, 83, 84

[36] P. A. D. Castro and F. J. Von Zuben. BAIS: A Bayesian Artificial Immune System for the Effective Handling of Building Blocks. *Inf. Sci.*, 179(10):1426–1440, 2009. 84

[37] Jun Chen and Mahdi Mahfouf. A Population Adaptive Based Immune Algorithm for Solving Multi-Objective Optimization Problems. In *Proceedings of the 5th International Conference on Artificial Immune Systems, ICARIS'06*, volume 4163 of *Lect Notes Comput Sc*, pages 280–293, Oeiras, Portugal, 2006. Springer-Verlag. 81, 83

[38] R. Church and C. ReVelle. The Maximal Covering Location Problem. *Papers of the Regional Science Association*, 32:101–118, 1974. 19, 24

[39] R. L. Church, W. J. Okin, M. Figueroa, and K. Barber. Integrating the Biodiversity Management Area Selection Model into a Multi-Use Forest Programming Model: The Case of RELMDSS. In *Seventh Symposium on Systems Analysis in Forest Resources*, Bellaire, MI, May 1997. 26

[40] R. L. Church, D. M. Stoms, and F. W. Davis. Reserve Selection as a Maximal Covering Location Problem. *Biol Conserv*, 76(2):105–112, 1996. 19, 22

[41] M. Ciarleglio. *Modular Abstract Self-Learning Tabu Search (MASTS) Metaheuristic Search Theory and Practice*. PhD thesis, University of Texas at Austin, Texas, May 2008. 42, 49

[42] M. Ciarleglio, J. W. Barnes, and S. Sarkar. ConsNet: New Software for the Selection of Conservation Area Networks with Spatial and Multi-Criteria Analyses. *Ecography*, 32:205–209, 2009. 40, 42, 46

[43] M. Ciarleglio, J. W. Barnes, and S. Sarkar. ConsNet: A Tabu Search Approach to the Spatially Coherent Conservation Area Network Design Problem. *J Heuristics*, 16:537–557, 2010. 40, 42, 46, 55

[44] E. Clark, A. Hone, and J. Timmis. A Markov Chain Model of the B-cell Algorithm. In C. Jacob, M. L. Pilat, P. J. Bentley, and J. Timmis, editors, *Proceedings of the 4th International Conference on Artificial Immune Systems, ICARIS'05*, volume 3627 of *Lect Notes Comput Sc*, pages 318–330, Banff, Canada, August 2005. Springer-Verlag. 85, 86, 240

[45] E. B. Clark. *A Framework for Modelling Stochastic Optimisation Algorithms with Markov Chains*. PhD thesis, Department of Electronics, University of York, November 2008. 86

[46] G. P. Coelho and F. J. Von Zuben. Omni-aiNet: An immune-inspired Approach for Omni Optimization. In *Proceedings of the 5th International Conference on Artificial Immune Systems, ICARIS'06*, volume 4163 of *Lect Notes Comput Sc*, pages 294–308, Oeiras, Portugal, September 2006. Springer-Verlag. 81, 83, 84

[47] C. A. Coello Coello and N. C. Cortés. Solving Multiobjective Optimization Problems Using an Artificial Immune System. *Genetic Programming and Evolvable Machines*, 6(2):163–190, 2005. 50, 62, 65, 81, 82, 94, 97

[48] C. A. Coello Coello, G. B. Lamont, and D. A. Van Veldhuizen. *Evolutionary Algorithms for Solving Multi-Objective Problems*. Springer-Verlag, New York, 2nd edition, September 2007. ISBN 978-0-387-33254-3. 4, 51, 56, 60, 62, 64, 116, 117, 124, 239, 240

[49] I. R. Cohen. Immune System Computation and the Immunological *Homunculus*. In O. Nierstrasz, J. Whittle, D. Harel, and G. Reggio, editors, *MoDELS*, volume 4199 of *Lect Notes Comput Sc*, pages 499–512. Springer-Verlag, 2006. 72

[50] R. G. Collevatti, J. S. Lima, T. N. Soares, and M. P. de C. Telles. Spatial Genetic Structure and Life History Traits in Cerrado Tree Species: Inferences for Conservation. *Nat Conservacao*, 8(1):54–59, 2010. 8, 102

[51] Convención sobre la Diversidad Biológica (CDB). Guías Breves de las Metas de Aichi para la Diversidad Biológica. Online, 2011. Disponível em <https://www.cbd.int/nbsap/training/quick-guides/>. Acessado em ago. 2012. 2

[52] Convención sobre la Diversidad Biológica (CDB). Decisión Adoptada por la Conferencia de las Partes de la Convención sobre la Diversidad Biológica Durante su Decima Reunión: X/2. El Plan Estratégico para la Diversidad Biológica 2011-2020 y las Metas de Aichi para la Diversidad Biológica. Online, June 2012. Disponível em <http://www.cms.int/about/nbsap/cbd_cop10_decision_s.pdf>. Acessado em set./2012. 2

[53] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein. *Introduction to Algorithms*. The MIT Press, 2nd edition, 2001. 3, 21, 24

[54] G. C. Correa, R. V. Naves, M. R. Rocha, L. J. Chaves, and J. D. Borges. Determinações Físicas em Frutos e Sementes de Baru (*Dipteryx alata* Vog.), Cajuzinho (*Anacardium othonianum* Rizz.) e Pequi (*caryocar Brasiliense* Camb.), Visando Melhoramento GenÉtico. *Biosci. J*, 24(4):42–47, 2008. 101, 102

[55] N. C. Cortés. *Sistema Inmune Artificial para Solucionar Problemas de Optimización*. PhD thesis, Centro de Investigación y de Estudios Avanzados del Instituto Politécnico Nacional – México, 2004. 57, 81, 82, 83

[56] N. C. Cortés, D. Trejo-Pérez, and C. A. Coello Coello. Handling Constraints in Global Optimization Using an Artificial Immune System. In C. Jacob, M. L. Pilat, P. J. Bentley, and J. Timmis, editors, *Proceedings of the 4th International Conference on Artificial Immune Systems, ICARIS'05*, volume 3627 of *Lect Notes Comput Sc*, pages 234–247, Banff, Canada, August 2005. Springer-Verlag. 51

[57] R. H. Crozier. Preserving the Information Content of Species: Genetic Diversity, Phylogeny, and Conservation Worth. *Annu Rev Ecol Syst*, 28:243–268, 1997. 8

[58] V. Cutello, G. Narzisi, G. Nicosia, and M. Pavone. Clonal Selection Algorithms: A Comparative Case Study Using Effective Mutation Potentials. In C. Jacob, M. L. Pilat, P. J. Bentley, and J. Timmis, editors, *Proceedings of the 4th International Conference on Artificial Immune Systems, ICARIS'05*, volume 3627 of *Lect Notes Comput Sc*, pages 13—-28, Banff, Canada, August 2005. Springer-Verlag. 57, 85

[59] V. Cutello and G. Nicosia. An Immunological Approach to Combinatorial Optimization Problems. In Francisco J. Garijo, José C. Riquelme, and Miguel Toro, editors, *Advances in Artificial Intelligence, IBERAMIA 2002*, volume 2527 of *Lect Notes Comput Sc*, pages 361–370, Berlin, Heidelberg, November 2002. Springer Berlin Heidelberg. 85

[60] V. Cutello, G. Nicosia, and M. Pavone. A Hybrid Immune Algorithm with Information Gain for the Graph Coloring Problem. In E. Cantú-Paz, J. A. Foster, K. Deb, L. D. Davis, R. Roy, U.-M. O'Reilly, H.-G. Beyer, R. Standish, G. Kendall, S. Wilson, M. Harman, J. Wegener, D. Dasgupta, M. A. Potter, A. C. Schultz, K. A. Dowsland, N. Jonoska, and J. Miller, editors, *Proceedings of the Genetic and Evolutionary Computation Conference, GECCO'03*, volume 2723 of *Lect Notes Comput Sc*, pages 171–182, Chicago, IL, USA, 12-16 July 2003. Springer Berlin Heidelberg. 85

[61] V. Cutello, G. Nicosia, and M. Pavone. An Immune Algorithm with Stochastic Aging and Kullback Entropy for the Chromatic Number Problem. *J Comb Optim*, 14(1):9–33, 2007. 4

[62] V. Cutello, G. Nicosia, M. Romeo, and P. S. Oliveto. On the Convergence of Immune Algorithms. In *Foundations of Computational Intelligence, 2007. FOCI 2007. IEEE Symposium on*, pages 409–415, April 2007. 57, 85, 86, 87, 240

[63] L. P. Czeder. Composição Nutricional e Qualidade Protéica da Amêndoa de Baru (*Dipteryx alata* Vog.) de Planta de Três Regiões do Cerrado do Estado de Goiás. Master's thesis, Escola de Agronomia e Engenharia de Alimentos, Universidade Federal de Goiás (UFG), 2009. 101

[64] V. G. da Fonseca, C. M. Fonseca, and A. O. Hall. Inferential Performance Assessment of Stochastic Optimisers and the Attainment Function. In E. Zitzler, K. Deb, L. Thiele, C. A. Coello Coello, and D. Corne, editors, *Proceedings of the 1st International Conference on Evolutionary Multi-Criterion Optimization, EMO'2001*, volume 1993 of *Lect Notes Comput Sc*, pages 213–225, London, UK, 2001. Springer-Verlag. 116, 117

[65] Ministério da Relações Exteriores and Ministério do Meio Ambiente. Portal da Convenção sobre Diversidade Biológica (CDB). Disponível em <http://www.cdb.gov.br/CDB>. Acessado em jun./2012. 2

[66] Ministério da Relações Exteriores / Ministério do Meio Ambiente. Glossário de Termos para a MOP3/COP8, 2006. Disponível em <http://www.cbd.int/cepa/toolkit/2008/doc/CBD-Toolkit-Glossaries.pdf>. Acessado em jul./2012. 1

[67] D. Dasgupta. *Artficial Immune Systems and Their Applications*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 1998. 79

[68] I. Dawson and J. Were. Collecting Germplasm from Trees – Some Guidelines. *Agroforestry Today*, 9(2):6–9, 1997. 106

[69] S. P de Almeida. *Cerrado: Aproveitamento Alimentar*. Embrapa - CPAC, Planaltina, 1998. 101, 102

[70] A.C.P.L.F. de Carvalho, A.C.B. Delbem, R.A.F. Romero, E.V. Sim oes, and G.P. Telles. Computação Bioinspirada. In *XXIII Jornada de Atualização em Informática (JAI) doXXIV Congresso da Sociedade Brasileira de Computação*, Salvador, Bahia, 2004. 60

[71] L. N. de Castro. The Clonal Selection Algorithm with Engineering Applications. In *Proceedings of the Annual Conference on Genetic and Evolutionary Computation, GECCO'00*, pages 36–37. Morgan Kaufmann, 2000. 76

[72] L. N. de Castro. Immune Engineering: A Personal Account. In *In Proceedings of the II Workshop on Computational Intelligence and Semiotics*, page CD ROM. IEEE Press, 2002. 79, 81, 94

[73] L. N. de Castro. Immune, Swarm, and Evolutionary Algorithms Part I: Basic Models. In *Workshop on Artificial Immune Systems 3*, pages 1464–1468. ICONIP Conference (International Conference on Neural Information Processing), 2002. 4

[74] L. N. de Castro. Immune, Swarm, and Evolutionary Algorithms Part II: Philosophical Comparisons. In *Workshop on Artificial Immune Systems 3*, pages 1469–1473. ICONIP Conference (International Conference on Neural Information Processing), 2002. 4

[75] L. N. de Castro. *Fundamentals of Natural Computing: Basic Concepts, Algorithms, and Applications (Computer and Information Sciences)*. Chapman & Hall/CRC, June 2006. 57, 59, 72

[76] L. N. de Castro. Fundamentals of Natural Computing: an Overview. *Phys Life Rev*, 4(1):1–36, March 2007. 57

[77] L. N. de Castro and J. Timmis. An Artificial Immune Network for Multimodal Function Optimization. In *Proceedings of the 2002 Congress on Evolutionary Computation, CEC '2002*, volume 1, pages 699–704, 2002. 81, 82

[78] L. N. de Castro and J. Timmis. *Artificial Immune Systems: A New Computational Intelligence Approach*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2002. 74, 75, 76, 77, 79

[79] L. N. de Castro and J. Timmis. Artificial Immune Systems: A Novel Paradigm to Pattern Recognition. In *University of Paisley*, pages 67–84. Springer Verlag, University of Paisley, UK, 2002. 75, 76, 80

[80] L. N. de Castro and F. J. Von Zuben. Artificial Immune Systems: Part I – Basic Theory and Applications – RT DCA 0199. Technical report, Universidad de Campinas, 1999. 74, 80, 94

[81] L. N. de Castro and F. J. Von Zuben. Artificial Immune Systems: Part II – A Survey of Applications – RT DCA 0200. Technical report, Universidad de Campinas, 2000. 80, 94

[82] L. N. de Castro and F. J. Von Zuben. Learning and Optimization Using the Clonal Selection Principle. *IEEE T Evolut Comput*, 6(3):239–251, 2002. 4, 76, 78, 79, 81, 85

[83] L. N. de Castro and F. J. Von Zuben. The Clonal Selection Algorithm with Engineering Applications. In *Proceedings of the Annual Conference on Genetic and Evolutionary Computation, GECCO'02*, pages 36–37. Morgan Kaufmann, 2002. 4, 85

[84] F. O. de França, F. J. Von Zuben, and L. N. de Castro. An Artificial Immune Network for Multimodal Function Optimization on Dynamic Environments. In *Proceedings of the 7th Annual Conference on Genetic and Evolutionary Computation, GECCO'05*, pages 289–296, New York, NY, USA, 2005. ACM. 81, 82

[85] K. A. De Jong. *Evolutionaty Computation: a Unified Approach.* MIT Press, Cambridge, MA, 2006. 59, 60

[86] K. Deb. Multi-Objective Genetic Algorithms: Problem Difficulties and Construction of Test Problems. *Evol Comput*, 7:205–230, 1998. 54, 95

[87] K. Deb. *Multiobjective Optimization*, chapter Introduction to Evolutionary Multiobjective Optimization, pages 59–96. Springer-Verlag, 2008. 51

[88] K. Deb and H. Jain. An Evolutionary Many-Objective Optimization Algorithm Using Reference-Point-Based Nondominated Sorting Approach, Part I: Solving Problems With Box Constraints. *IEEE Trans Evol Comput*, 18(4):577–601, Aug 2014. 237, 238

[89] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan. A Fast and Elitist Multiobjective Genetic Algorithm: NSGA-II. *IEEE T Evolut Comput*, 6(2):182–197, April 2002. 63, 64, 65, 67, 68, 117, 239

[90] K. Deb and A. Sinha. An Efficient and Accurate Solution Methodology for Bilevel Multi-objective Programming Problems Using a Hybrid Evolutionary-local-search Algorithm. *Evol Comput*, 18(3):403–449, September 2010. 237

[91] K. Deb and S. Tiwari. Omni-Optimizer: A Procedure for Single and Multi-Objective Optimization. In C. A. Coello Coello, A. Hernández Aguirre, and E. Zitzler, editors, *Proceedings of the 3rd International Conference on Evolutionary Multi-Criterion Optimization, EMO'2005*, volume 3410 of *Lect Notes Comput Sc*, pages 47–61, Guanajuato, México, March 2005. Springer-Verlag. 81, 83, 84

[92] J. A. F. Diniz-Filho, D. B. Melo, G. de Oliveira, R. G. Collevatti, T. N. Soares, J. C. Nabout, J. de S. Lima, R. Dobrovolski, L. J. Chaves, R. V. Naves, R. D. Loyola, and M. P. de C. Telles. Planning for Optimal Conservation Geographical Genetic Variability within Species. *Conserv Genet*, 13:1085–1093, 2012. 8, 9, 10, 103

[93] J. A. F. Diniz-Filho and M. P. C. Telles. Spatial Autocorrelation Analysis and the Identification of Operational Units for Conservation in Continuous Populations. *Conserv Biol*, 16(4):924–935, 2002. 9

[94] J.A.F. Diniz-Filho and M.P.C. Telles. Optimization Procedures for Establishing Reserve Networks for Biodiversity Conservation Taking into Account Population Genetic Structure. *Genet Mol Biol*, 29:207–214, 2006. 8

[95] J.L. Doob. *Stochastic Processes*. Wiley Publications in Statistics. John Wiley & Sons, 1953. 86

[96] A. L. Drummond. Compósitos Poliméricos Obtidos a Partir do Óleo de Baru - Síntese e Caracterização. Master's thesis, Instituto de Química, Universidade de Brasília (UnB), 2008. 100, 102

[97] N. Dudley, editor. *Lignes Directrices pour l'Application des Catégories de Gestion aux Aires Protégés*. Union Internationale pour la conservation de la natures et de ses resources, Gland, Suisse, 2008. 13, 14

[98] A. E. Eiben and J. E. Smith. *Introduction to Evolutionary Computing*. Springer, Berlin, 2003. 60, 61

[99] M. Emmerich, N. Beume, and B. Naujoks. An EMO Algorithm Using the Hypervolume Measure as Selection Criterion. In C. A. Coello Coello, A. A. Hernández, and E. Zitzler, editors, *Proceedings of the 3rd International Conference on Evolutionary Multi-Criterion Optimization, EMO'2005*, volume 3410 of *Lect Notes Comput Sc*, pages 62–76. Springer Berlin Heidelberg, 2005. 120

[100] J. Engels, L. Visser, and International Plant Genetic Resources Institute. *A Guide to Effective Management of Germplasm Collections*. IPGRI Handbooks for Genebanks. International Plant Genetic Resources Institute, 2003. 106

[101] Mark Erickson, Alex Mayer, and Jeffrey Horn. The Niched Pareto Genetic Algorithm 2 Applied to the Design of Groundwater Remediation Systems. In E. Zitzler, K. Deb, L. Thiele, C. A. Coello Coello, and D. Corne, editors, *Proceedings of the 1st International Conference on Evolutionary Multi-Criterion Optimization, EMO'2001*, volume 1993 of *Lect Notes Comput Sc*, pages 681–695, London, UK, 2001. Springer-Verlag. 64

[102] J. Ferreira, L. E. O. C. Arag ao, J. Barlow, P. Barreto, E. Berenguer, M. Bustamante, T. A. Gardner, A. C. Lees, A. Lima, J. Louzada, R. Pardini, L. Parry, C. A. Peres, P. S. Pompeu, M. Tabarelli, and J. Zuanon. Brazil's Environmental Leadership at Risk. *Science*, 346(6210):706–707, 2014. 4

[103] P.A.V. Ferreira. *Otimização Multiobjetivo: Teoria e Aplicações, Tese de Livre Docência*. PhD thesis, Faculdade de Engenharia Elétrica e de Computação, Universidade Estadual de Campinas, 1999. 3

[104] M. Fleischer. The Measure of Pareto Optima Applications to Multi-objective Metaheuristics. In *Proceedings of the 2nd International Conference on Evolutionary Multi-criterion Optimization, EMO'2003*, Lect Notes Comput Sc, pages 519–533, Berlin, Heidelberg, 2003. Springer-Verlag. 118

[105] C. M. Fonseca, V. G. da Fonseca, and L. Paquete. Exploring the Performance of Stochastic Multiobjective Optimisers with the Second-Order Attainment Function. In C. A. Coello Coello, A. Hernández Aguirre, and E. Zitzler, editors, *Proceedings of the 3rd International Conference on Evolutionary Multi-Criterion Optimization, EMO'2005*, volume 3410 of *Lect Notes Comput Sc*, pages 250–264, Guanajuato, México, March 2005. Springer-Verlag. 117

[106] C. M. Fonseca and P. J. Fleming. Genetic Algorithms for Multiobjective Optimization: Formulation, Discussion and Generalization. In *Proceedings of the 5th International Conference on Genetic Algorithms*, pages 416–423. Morgan Kaufmann, 1993. 63

[107] C. M. Fonseca and P. J. Fleming. An Overview of Evolutionary Algorithms in Multiobjective Optimization. *Evol Comput*, 3:1–16, 1995. 3, 4, 52, 62

[108] S. Forrest, S. A. Hofmeyr, and A. Somayaji. A Sense of Self for Unix Processes. In *Proceedings of the 1996 IEEE Symposium on Security and Privacy*, pages 120–128. IEEE Computer Society Press, 1996. 79

[109] O. H. Frankel. *Genetic Manipulation: Impact on Man and Society*, chapter Genetic perspectives of germplasm conservation, pages 161–170. Cambridge University Press, 1984. 106

[110] F. Freschi and M. Repetto. Multiobjective Optimization by a Modified Artificial Immune System Algorithm. In C. Jacob, M. L. Pilat, P. J. Bentley, and J. Timmis, editors, *Proceedings of the 4th International Conference on Artificial Immune Systems, ICARIS'05*, volume 3627 of *Lect Notes Comput Sc*, pages 248–261, Banff, Canada, August 2005. Springer-Verlag. 81, 83

[111] E. T. Game and H. S. Grantham. *Marxan User Manual: For Marxan version 1.8.10*. University of Queensland, St. Lucia, Queensland, Australia, and Pacific Marine Analysis and Research Association, Vancouver, British Columbia, Canada, February 2008. 42, 45

[112] J. Garson, A. Aggarwal, and S. Sarkar. *ResNet Manual Verv 1.2*. Biodiversity and Biocultural Conservation Laboratory, Section of Integrative Biology, University of Texas at Austin, Austin, TX, 2002. Disponível em <http://uts.cc.utexas.edu/ consbio/Cons/program.html>. Acessado em set. 2012. 42, 45, 46, 49

[113] GENPAC. Rede 09 – GENPAC - Genética Geográfica e Planejamento Regional para Conservação de Recursos Naturais no Cerrado. Online, 2012. Disponível em <http://redeprocentrooeste.org.br/portal/redes/rede-09/folders/folder-01—portugues.htm>. Acessado em ago. 2012. 7, 8

[114] W. Gong and Z. Cai. A Multiobjective Differential Evolution Algorithm for Constrained Optimization. In *Proceedings of the 2008 Congress on Evolutionary Computation, CEC '2008*, pages 181–188, Hong Kong, June 2008. IEEE Service Center. 4

[115] N. J. Gotelli and G. R. Graves. *Null Models in Ecology*. Smithsonian Institution Press, Washington and London, 1996. 108

[116] H. S. Grave. The Duty of Scientific Men in Conservation. *Science*, 1379(53):505–509, June 1921. 1

[117] C.E.V. Grelle, M.L. Lorini, and M.P. Pinto. Reserve Selection Based on Vegetation in the Brazilian Atlantic Forest. *Nat Conservacao*, 8:46–53, 2010. 8

[118] C. R. Groves, D. B. Jensen, L. L. Valutis, K. H. Redford, M. L. Shaffer, M. J. Scott, J. V. Baumgartner, J. V. Higgins, M. W. Beck, and M. G. Anderson. Planning for Biodiversity Conservation: Putting Conservation Science into Practice. *BioScience*, 52(6):499–512, June 2002. 14

[119] A. P. Guerreiro. Efficient Algorithms for the Assessment of Stochastic Multiobjective Optimizers. Master's thesis, Universidade Técnica de Lisboa, Lisboa, November 2011. 117

[120] A.K. Gupta, S. Sanjeev, C. Vikas, Gosh S. B., S. Riya, and J. Neeru. Cost of Conservation of Agrobiodiversity. IIMA Working Papers WP2002-05-03, Indian Institute of Management Ahmedabad, Research and Publication Department, 2002. 9, 106

[121] Aurélio Buarque de Holanda-Ferreira. Novo dicionário eletrônico aurélio versão 6.0.1. versão eletrônica 6.0.1, 2009. 237

[122] J. Horn, N. Nafpliotis, and D. E. Goldberg. A Niched Pareto Genetic Algorithm for Multiobjective Optimization. In *Proceedings of the 1st IEEE Conference on Evolutionary Computation, IEEE World Congress on Computational Intelligence*, pages 82–87, 1994. 63

[123] C. J. Humphries, Paul H. Williams, and R. I. Vane-Wright. Measuring Biodiversity Value for Conservation. *Annu Rev Ecol Syst*, 26:93–111, 1995. 8

[124] F. Hutter, H. H. Hoos, K. Leyton-Brown, and T. Stützle. Paramils: An automatic algorithm configuration framework. *J. Artif. Int. Res.*, 36(1):267–306, September 2009. 239

[125] IBGE. IBGE Lança o Mapa de Biomas do Brasil e o Mapa de Vegetação do Brasil, em Comemoração ao Dia Mundial da Biodiversidade. Online, 2004. Disponível em <http://www.ibge.gov.br/home/presidencia/noticias/noticia_visualiza.php ?id_noticia=169>. Acessado em ago. 2012. 5

[126] IBGE-MMA. Mapas de Bioma e Vegetação. Online, 2004. Disponível em <ftp://ftp.ibge.gov.br/Cartas_e_Mapas/Mapas_Murais/>. Acessado em ago.2012. 5

[127] Instituto Brasileiro do Meio Ambiente e dos Recursos Naturais Renováveis (IBAMA). *Planejamento Sistemático da Conservação: Material Didático*. Ibama, 2010. 4, 14, 16, 17

[128] H. Jain and K. Deb. An Evolutionary Many-Objective Optimization Algorithm Using Reference-Point Based Non-Dominated Sorting Approach. Part II: Handling Constraints and Extending to an Adaptive Approach. *IEEE Trans Evol Comput*, 18(4):602–622, 2014. 237

[129] W. Jakob, M. Gorges-Schleuter, and C. Blume. Application of Genetic Algorithms to Task Planning and Learning. In R. Männer and B. Manderick, editors, *Proceedings of the International Conference on Parallel Problem Solving From Nature, PPSN'1992*, pages 293–302. Elsevier, 1992. 54

[130] C. A. Janeway and P. Travers. *Immunobiology: The Immune System in Health and Disease*. Garland, New York, 2007. 72, 75, 77

[131] T. Jansen and C. Zarges. Analyzing Different Variants of Immune Inspired Somatic Contiguous Hypermutations. *Theor Comput Sci*, 412:517–533, 2011. Theoretical Aspects of Artificial Immune Systems. 239

[132] J. Justus and S. Sarkar. The Principle of Complementarity in the Design of Reserve Networks to Conserve Biodiversity: A Preliminary History. *J Biosci*, 27(4 Suppl 2):421–435, July 2002. 28, 42, 49

[133] A. Juutinen, M. Mönkkönen, and M. Ollikainen. Do Environmental Diversity Approaches Lead to Improved Site Selection? A Comparison with the Multi-Species Approach. *Forest Ecol Manag*, 255(11):3750–3757, June 2008. 8

[134] N. Karmakar. A New Polynomial-Time Algorithm for Linear Programming. *Combinatorica*, 4(4):373–395, 1984. 24

[135] C. Kelley, J. Garson, A. Aggarwal, and S. Sarkar. Place Prioritization for Biodiversity Reserve Network Design: A Comparison of the SITES and ResNet Software Packages for Coverage and Efficiency. *Divers Distrib*, 8:297–306, 2002. 15, 16, 42, 44, 46

[136] J. Kelsey and J. Timmis. Immune Inspired Somatic Contiguous Hypermutation for Function Optimisation. In E. Cantú-Paz, J. A. Foster, K. Deb, L. D. Davis, R. Roy, U.-M. O'Reilly, H.-G. Beyer, R. Standish, G. Kendall, S. Wilson, M. Harman, J. Wegener, D. Dasgupta, M. A. Potter, A. C. Schultz, K. A. Dowsland, N. Jonoska, and J. Miller, editors, *Proceedings of the Genetic and Evolutionary Computation Conference, GECCO'03*, volume 2723 of *Lect Notes Comput Sc*, pages 207–218, Chicago, IL, USA, 12-16 July 2003. Springer Berlin Heidelberg. 57, 85

[137] J. Kelsey, J. Timmis, and A. Hone. Chasing Chaos. In *Proceedings of the 2003 Congress on Evolutionary Computation, CEC '2003*, volume 1. IEEE Computer Society, USA, 2003. 85

[138] J. B. Kirkpatrick. An Interative Method for Establishing Priorities for the Selection of Nature Reserves – An Example from Tasmania. *Biol Conserv*, 25(2):127–134, 1983. 8, 22, 27, 49

[139] J. B. Kirkpatrick and M. J. Brown. *Nature Conservation: Cost Effective Biological Surveys and Data Analysis*, chapter Planning for Species Conservation. Commonwealth Scientific & Industrial Research (CSIRO), Dickson, Australia, 1991. 27

[140] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by Simulated Annealing. *Science*, 220(4598):671–680, May 1983. 26

[141] K. Klamroth and J. Tind. Constrained Optimization Using Multiple Objective Programming. *J Global Optim*, 37(3):325–355, March 2007. 4

[142] J. D. Knowles and D. W. Corne. Local Search, Multiobjective Optimization and the Pareto Archived Evolution Strategy. In *Ashikaga Institute of Technology*, pages 209–216, 1999. 64, 95, 96

[143] J. D. Knowles and D. W. Corne. Approximating the Nondominated Front Using the Pareto Archived Evolution Strategy. *Evol Comput*, 8(2):149–172, 2000. 63, 64, 96

[144] J. D. Knowles and D. W. Corne. Properties of an Adaptive Archiving Algorithm for Storing Nondominated Vectors. *IEEE T Evolut Comput*, 7(2):100–116, 2003. 96, 118, 119

[145] M. Köppen. On the Benchmarking of Multiobjective Optimization Algorithm. In V. Palade, R. J. Howlett, and L. C. Jain, editors, *Proceedings of the 7th International Conference on Knowledge-Based Intelligent Information and Engineering Systems, KES'2003. Part I*, volume 2773 of *Lect Notes Comput Sc*, pages 379–385, Oxford, UK, September 2003. Springer Berlin Heidelberg. 55, 56

[146] M. Köppen, D. H. Wolpert, and W. G. Macready. Remarks on a Recent Paper on the "No Free Lunch"Theorems. *IEEE T Evolut Comput*, 5(3):295–296, June 2001. 55

[147] F. W. W. Larsen, J. Bladt, and C. Rahbek. Indicator Taxa Revisited: Useful for Conservation Planning? *Diversity Distrib*, 15(1):70–79, 2009. 15

[148] A. T. Lombard. Introduction to an Evaluation of the Protection Status of South Africa's Vertebrates. *S Afr J Zool*, 30(3):63–70, 1995. 36

[149] A. López-Jaimes and C. A. Coello Coello. Multi-Objective Evolutionary Algorithms: A Review of the State-of-the-Art and some of their Applications in Chemical Engineering. In Rangaiah Gade Pandu, editor, *Multi-Objective Optimization Techniques and Applications in Chemical Engineering*, chapter 3, pages 61–90. World Scientific, Singapore, 2009. ISBN 978-981-283-651-9. 64

[150] R. D. Loyola and T. M. Lewinsohn. Diferentes Abordagens para a Seleção de Prioridades de Conservação em um Contexto Macrogeográfico. *Megadiversidade*, 5(1-2):27–42, 2009. 8, 15

[151] G.-C. Luh, C.-H. Chueha, and W.-W. Liu. MOIA: Multi-Objective Immune Algorithm. *Eng Optimiz*, 35(2):143–164, 2003. 81, 83

[152] R. B. Machado, M. B. Ramos Neto, P. G. P. Pereira, E. F. Caldas, D. A. Gonçalves, N. S. Santos, K. Tabor, and M. Steininger. Estimativas de Perda da Área do Cerrado Brasileiro. Technical report, Conservação Internacional, Brasilia, DF, July 2004. 5

[153] R. M. Magalhães. *Obstáculos à Exploração do Baru (Dipteryx alata Vog.) no Cerrado Goiano: Sustentabilidade Comprometida?* PhD thesis, Centro de Desenvolvimento Sustentável, Universidade de Brasília (UnB), 2011. 6

[154] C. R. Margules and A.O. Nicholls. *Nature Conservation: The Role of Remnants of Native Vegetation*, chapter Assessing the conservation value of remnant habitat 'islands': Mallee patches on the Western Eyre peninsula, pages 89–102. Surrey Beatty and Sons, Chipping Norton, 1987. 8, 30, 49

[155] C. R. Margules and R. L. Pressey. Systematic Conservation Planning. *Nature*, 405(6783):243–253, May 2000. 14, 15, 16

[156] C. R. Margules, R. L. Pressey, and A. O. Nicholls. *Nature Conservation: Cost Effective Biological Surveys and Data Analysis*, chapter Selecting Nature Reserves. Commonwealth Scientific & Industrial Research (CSIRO), Dickson, Australia, 1991. 23, 30, 31, 32, 34

[157] C. R. Margules and S. Sarkar. *Systematic Conservation Planning.* Cambridge University Press, 2007. 23

[158] C.R. Margules, A.O. Nicholls, and R.L. Pressey. Selecting Networks of Reserves to Maximise Biological Diversity. *Biol Conserv*, 43(1):63–76, 1988. 8, 32, 34, 49

[159] S. Marino, I. B. Hogue, C. J. Ray, and D. E. Kirschner. A Methodology for Performing Global Uncertainty and Sensitivity Analysis in Systems Biology. *J Theor Biol*, 254(1):178–196, September 2008. 108, 113, 140

[160] A. Menchaca-Mendez and C. A. Coello Coello. A New Proposal to Hybridize the Nelder-Mead Method to a Differential Evolution Algorithm for Constrained Optimization. In *Proceedings of the 2009 Congress on Evolutionary Computation, CEC '2009*, pages 2598–2605, Trondheim, Norway, May 2009. IEEE Press. 4

[161] N. Mladenovic, J. Brimberg, P. Hansen, and J. Morenoperez. The P-Median Problem: A Survey of Metaheuristic Approaches. *Eur J Oper Res*, 179(3):927–939, June 2007. 21

[162] A. Moffett, J. Garson, and S. Sarkar. MultCSync: a Software Package for Incorporating Multiple Criteria in Conservation Planning. *Environ Modell Softw*, 20(10):1315–1322, 2005. 42, 46, 49

[163] A. Moilanen. Reserve Selection Using Nonlinear Species Distribution Models. *Am Nat*, 165(6):695–706, 2005. 49

[164] A. Moilanen. Landscape Zonation, Benefit Functions And Target-Based Planning: Unifying Reserve Selection Strategies. *Biol Conserv*, 134:571–579, 2007. 42

[165] A. Moilanen and H. Kujala. *Zonation: Spatial Conservation Planning Framework and Software v.1.0 – User Manual*. Metapopulation Research Group, Dept. Biological and Environmental Sciences, University of Helsinki, Helsinki, Finland, November 2006. 42, 47, 48, 49

[166] J. N. Muriuki, H. M. De Klerk, P. H. Williams, L. A. Bennun, T. M. Crowe, and E. V. Berge. Using Patterns of Distribution and Diversity of Kenyan Birds to Select and Prioritize Areas for Conservation. *Biodivers Conserv*, 6:292–210, 1997. 42

[167] N. Myers, R. A. Mittermeier, C. G. Mittermeier, G. A. B. da Fonseca, and J. Kent. Biodiversity Hotspots for Conservation Priorities. *Nature*, 403:853–858, 2000. 5

[168] A. O. Nicholls and C.R. Margules. An Updated Reserve Selection Algorithm. *Biol Conserv*, 64:165–169, 1993. 49

[169] N. C. Nussenzweig. Immune Receptor Editing: Revise and Select. *Cell*, 95(7):875–878, December 1998. 77

[170] P. S. Oliveto and D. Sudholt. On the Runtime Analysis of Stochastic Ageing Mechanisms. In *Proceedings of the 2014 Conference on Genetic and Evolutionary Computation, GECCO'14*, GECCO '14, pages 113–120, New York, NY, USA, 2014. ACM. 239

[171] D. M. Olson, E. Dinerstein, G. V.N. Powell, and E. D. Wikramanayake. Conservation Biology for the Biodiversity Crisis. *Conserv Biol*, 16(1):1–3, Feb. 2002. 2

[172] A. Osyczka. *Multicriteria Optimization for Engineering*. Academic Press, London, 1985. 50

[173] A. R. Pacheco. *Adubação de Mudas de Baru (Dipteryx alata Vog.) em Viveiro*. PhD thesis, Escola de Agronomia e Engenharia de Alimentos, Universidade Federal de Goiás (UFG), 2008. 6, 101

[174] E. Paes and P. B. Blinder. Modelos Nulos e Processos de Aleatorização: Algumas Aplicações em Ecologia de Comunidades. *Oecologia Brasiliensis*, 2(1):119–139, 1995. 108

[175] C. S. Pedamallu and L. Ozdamar. Investigating a Hybrid Simulated Annealing and Local Search Algorithm for Constrained Optimization. *Eur J Oper Res*, 185(3):1230–1245, March 2008. 4

[176] T. Pierrard and C. A. Coello Coello. A Multi-Objective Artificial Immune System Based on Hypervolume. In Carlos A. Coello Coello, Julie Greensmith, Natalio Krasnogor, Pietro Liò, Giuseppe Nicosia, and Mario Pavone, editors, *Proceedings of the 10th International Conference on Artificial Immune Systems, ICARIS'12*, volume 7597 of *Lect Notes Comput Sc*, pages 14–27. Springer-Verlag, 2012. 120

[177] H. P. Possingham, I. Ball, and S. Andelman. *Mathematical Methods for Identifying Representative Reserve Networks*, chapter 17, pages 291–305. Springer-Verlag, New York, 2000. 14, 15, 17, 21, 22, 24, 26, 27, 42, 49

[178] J. R. Prendergast, R. M. Quinn, and J. H. Lawton. The Gaps Between Theory and Practice in Selecting Nature Reserves. *Conserv Biol*, 13(2):484–492, June 1999. 8, 24, 42

[179] R. L. Pressey. The First Reserve Selection Algorithm: A Retrospective on Jamie Kirkpatrick's 1983 Paper. *Prog Phys Geog*, 26(3):434–441, 2002. 17, 27, 28, 30, 32

[180] R. L. Pressey. Conservation Planning and Biodiversity: Assembling the Best Data for the Job. *Conserv Biol*, 18(6):1677–1681, 2004. 8

[181] R. L. Pressey, C. J. Humphries, C. R. Margules, R. I. Vane-Wright, and P. H. Williams. Beyond Opportunism: Key Principles for Systematic Reserve Selection. *Trends Ecol Evol*, 8(4):124–128, April 1993. 22

[182] R. L. Pressey and A. O. Nicholls. *Nature Conservation: Cost Effective Biological Surveys and Data Analysis*, chapter Reserve Selection in the Western Division of New South Wales: Development of a New Procedure Based on Land System Mapping. Commonwealth Scientific & Industrial Research (CSIRO), Dickson, Australia, 1991. 27

[183] R. L. Pressey, H. P. Possingham, and J. R. Day. Effectiveness of Alternative Heuristic Algorithms for Identifying Indicative Minimum Requirements for Conservation Reserves. *Biol Conserv*, 80(2):207–219, 1997. 14, 18, 23, 25

[184] R. L. Pressey, H. P. Possingham, and C. R. Margules. Optimality in Reserve Selection Algorithms: When Does It Matter and How Much? *Biol Conserv*, 76(3):259–267, 1996. 21, 25

[185] R.L. Pressey, M. Watts, M. Ridges, and T. Barrett. *C-Plan Conservation Planning Software. User Manual.* NSW Department of Environment and Conservation, 2005 (última atualização em 9 ago. 2012). 42, 44, 49

[186] M. R. W. Rands, W. M. Adams, L. Bennun, S. H. M. Butchartand A. Clements, D. Coomes, A. Entwistle, I. Hodge, V. Kapos, J.P.W. Scharlemann, W.J. Sutherland, and B. Vira. Biodiversity Conservation: Challenges Beyond 2010. *Science*, 329:1298–1303, 2010. 1, 2, 14

[187] M. Read, P. Andrews, J. Timmis, and V. Kumar. Techniques for Grounding Agent-Based Simulations in the Real Domain: a Case Study in Experimental Autoimmune Encephalomyelitis. *Math Comp Model Dyn*, 18(1):67–86, 2012. 108, 109, 114, 138

[188] M. Read, P. S. Andrews, J. Timmis, and V. Kumar. A Domain Model of Experimental Autoimmune Encephalomyelitis. In *2nd Workshop on Complex Systems Modelling and Simulation*, pages 9–44, 2009. 108

[189] A. G. Rebelo and W.R. Siegfried. Protection of Fynbos vegetation: Ideal and Real-World Options. *Biol Conserv*, 54:15–31, 1990. 49

[190] Rede Pró-Centro-Oeste. Rede Centro-Oeste de Pós-Graduação, Pesquisa e Inovação. Online, 2012. Disponível em <http://redeprocentrooeste.org.br/>. Acessado em ago. 2012. 5, 6, 7

[191] J. Reese. Methods for Solving the P-Median Problem: An Annotated Bibliography. Technical report, Department of Mathematics, Trinity University, August 2005. 21

[192] C. ReVelle, M. Scholssberg, and J. Williams. Solving the Maximal Covering Location Problem with Heuristic Concentration. *Comput Oper Res*, 35(2):427–435, feb. 2008. 21

[193] J. F. Ribeiro, M. C. Oliveira, A. P. S. M. Gulias, J. M. F. Fagg, and F. G. Aquino. *Savanas: Desafios e Estratégias para o Equilíbrio entre a Sociedade, Agronegócio e Recursos Naturais*, chapter Usos Múltiplos da Biodiversidade no Bioma Cerrado: estratégia sustentável para a sociedade, o agronegócio e os recursos naturais, pages 337–360. Embrapa Cerrados, Planaltina, 2008. 6, 8, 101

[194] A. Rodrigues. Optimisation in Reserve Selection Procedures - Why Not? *Biol Conserv*, 107(1):123–129, September 2002. 49

[195] A. S. L. Rodrigues, K. J. Gaston, and R. D. Gregory. Using Presence-Absence Data to Establish Reserve Selection Procedures That Are Robust to Temporal Species Turnover. *Proc Biol Sci*, 1446(267):897–902, 2000. 24

[196] S. T. Rodrigues. Noções Básicas sobre Planejamento Sistemático da Conservação da Biodiversidade. In *Seminário ZEE-Zoneamento Ecológico-Econômico e Proteção da Biodiversidade*. Secretaria de Desenvolvimento Sustentável e de Biodiversidade e Florestas, Ministério do Meio Ambiente (MMA), 2006. 13

[197] R. Rosin-Arbesfeld, F. Townsley, and M. Bienz. The APC Tumour Suppressor Has a Nuclear Export Function. *Nature*, 406:1009–1012, 2000. 85

[198] S. Sarkar. Complementarity and the Selection of Nature Reserves: Algorithms and the Origins of Conservation Planning, 1980-1995. *Arch Hist Exact Sci*, 66:397–426, 2012. 21, 22, 24, 25, 26, 27, 28, 31, 32, 34, 36, 38, 40, 42

[199] S. Sarkar, A. Aggarwal, J. Garson, C. R. Margules, and J. Zeidler. Place Prioritization for Biodiversity Content. *J Bioscience*, 27(4 Suppl 2):339–346, July 2002. 27, 42

[200] S. Sarkar, T. Fuller, A. Aggarwal, A. Moffett, and C. D. Kelley. *Spatial Conservation Prioritization: Quantitative Methods and Computational Tools*, chapter The ConSNet software platform for systematic conservation planning. Oxford University Press, 2009. 42

[201] S. Sarkar, J. Garson, and A. Moffett. *MultCSync Manual v. 1.0*. Biodiversity and Biocultural Conservation Laboratory, Section of Integrative Biology, University of Texas at Austin, Austin, TX, July 2004. 42, 47, 48

[202] S. Sarkar and P. Illoldi-Rangel. Systematic Conservation Planning: An Updated Protocol. *Natureza Conservação*, 8(01):19–26, 2010. 15, 17, 22

[203] S. Sarkar, R. L. Pressey, D. P. Faith, C. R. Margules, T. Fuller, D. M. Stoms, A. Moffett, K. A. Wilson, K. J. Williams, P. H. Williams, and S. Andelman. Biodiversity Conservation Planning Tools: Present Status and Challenges for the Future. *Annu Rev Environ Resour*, 31, 2006. 28, 42

[204] J. D. Schaffer. Multiple Objective Optimization with Vector Evaluated Genetic Algorithms. In *Proceedings of the 1st International Conference on Genetic Algorithms*, pages 93–100, Hillsdale, NJ, USA, 1985. L. Erlbaum Associates Inc. 62, 63

[205] S. Schlottfeldt, J. Timmis, M. E. M. T. Walter, A. C. P. L. F. Carvalho, J. A. F. Diniz-Filho, L. M. Simon, R. D. Loyola, and M. P. C. Telles. Multi-objective Optimization Applied to Systematic Conservation Planning and Spatial Conservation Priorities Under Climate Change. In *Proceedings of the 2014 Conference Companion on Genetic and Evolutionary Computation, GECCO Comp '14*, pages 177–178, Vancouver, BC, Canada, 2014. ACM. 49

[206] S. Schlottfeldt, J. Timmis, M. E. M. T. Walter, A. C.P.L.F. Carvalho, J. A. F. Diniz-Filho, L. M. Simon, E. D. Loyola, and M. P. C. Telles. A Multi-Objective Optimization Approach Associated to Climate Change Scenario to Improve Systematic Conservation Planning and Spatial Conservation Priorities Setting. In A Gaspar-Cunha, C. A. Henggeler, and C. C. Coello, editors, *Evolutionary Multi-Criterion Optimization, EMO'2015*, volume 9019 of *Lect Notes Comput Sc*, pages 458–472, Guimarães, Portugal, 29 March - 1 April 2015. Springer International Publishing. 49, 234, 237

[207] S. Schlottfeldt, M. E. M. T. Walter, A. C. P. L. F. de Carvalho, T. N. Soares, M. P. C. Telles, R. D. Loyola, and J. A. F. Diniz-Filho. Multi-Objective Optimization for Plant Germplasm Collection Conservation of Genetic Resources Based on Molecular Variability. *Tree Genet Genomes*, 11(2), 2015. 240

[208] S. Schlottfeldt, M. E. M. T. Walter, A.C.P.L.F. de Carvalho, M. P. de C. Telles, and J. A. F. Diniz-Filho. Challenges in Computational Sustainability: a New Multi-Objective Artificial Immune System Approach to Deal with Biodiversity Conservation. In *Proceedings of the Workshop on New Directions of Evolutionary Computation and Intelligent Systems at the IEEE 2015 International Congress on Evolutionary Computation, CEC'2015*, 2015 IEEE CEC, 2015. 49

[209] S. Schlottfeldt, M. E. M. T. Walter, J. A. F. Diniz-Filho, and M. P. de C. Telles. Multiobjective Optimization in Systematic Conservation Planning to Represent Genetic Variability within Species. In *Proceedings of the 8th International Conference on Ecological Informatics–ISEI'2012*, 2012. 49

[210] S. Schlottfeldt, Y. Saéz, and P. Isasi. Sistemas Inmunológicos Artificiales aplicados al Problema de Optimización Multiobjetivo *Radio Network Design*. Technical Report UC3M-TR-CS-2009-01, Universidad Carlos III de Madrid, 2009. 4, 10, 83, 92, 138

[211] J. R. Schott. Fault Tolerant Design Using Single and Multicriteria Genetic Algorithm Optimization. Master's thesis, Department of Aeronautics and Astronautics, Massachusetts Institute of Technology, Cambridge, Massachusetts, May 1995. 120

[212] D. B. Segan, E. T. Game, M. E. Watts, R. R. Stewart, and H. P. Possingham. An Interoperable Decision Support Tool for Conservation Planning. *Environ Model Softw*, 26(12):1434–1441, December 2011. 23

[213] E. L. F. Senne and L. A. N. Lorena. Abordagens Complementares para Problemas de P-Medianas. *Produção*, 13:78 – 87, 00 2003. 21

[214] A. Sinha, P. Malo, P. Xu, and K. Deb. A Bilevel Optimization Approach to Automated Parameter Tuning. In *Proceedings of the 2014 Conference on Genetic and Evolutionary Computation, GECCO'14*, pages 847–854, New York, NY, USA, 2014. ACM. 110, 237

[215] T. N. Soares, L. J. Chaves, M. P. de C. Telles, J. A. F. Diniz-Filho, and L. V. Resende. Spatial Distribution of Intrapopulational Genetic Variability in *Dipteryx alata*. *Pesq Agropec Bras*, 43(9):1151–1158, September 2008. 8, 102

[216] T. N. Soares, D. B. Melo, L. V. Resende, R. P. Vianello, L. J. Chaves, R. G. Collevatti, and M. P. C. Telles. Development of Microsatellite Markers for the Neotropical Tree Species *Dipteryx alata* (*Fabaceae*). *Am J Bot*, 99:e72–e73, 2012. 102

[217] B. Soares-Filho, R. Raj ao, M. Macedo, A. Carneiro, W. Costa, M. Coe, H. Rodrigues, and A. Alencar. Cracking Brazil's Forest Code. *Science*, 344(6182):363–364, 2014. 5

[218] A. Somayaji, S. Hofmeyr, and S. Forrest. Principles of a Computer Immune System. In *In New Security Paradigms Workshop*, pages 75–82, 1997. 4, 79

[219] M. E. Soulé. Conservation: Tactics for a Constant Crisis. *Science*, pages 744–750, 1991. 1

[220] N. Srinivas and K. Deb. Multiobjective optimization using nondominated sorting in genetic algorithms. *Evol Comput*, 2:221–248, 1994. 63, 64

[221] S. Stepney, R. E. Smith, J. Timmis, and A. M. Tyrrell. Towards a Conceptual Framework for Artificial Immune Systems. In Giuseppe Nicosia, Vincenzo Cutello, Peter J. Bentley, and Jon Timmis, editors, *Proceedings of the International Conference on Artificial Immune Systems, ICARIS'04*, volume 3239 of *Lect Notes Comput Sc*, pages 53–64. Springer, 2004. 59

[222] S. Stepney, R. E. Smith, J. Timmis, A. M. Tyrrell, M. J. Neal, and A. N. W. Hone. Conceptual Frameworks for Artificial Immune Systems. *Int J Unconv Comput*, 1(3):315–338, July 2005. 59

[223] W. J. Sutherland, W. M. Adams, R. B. Aronson, R. Aveling, and et al. One Hundred Questions of Importance to the Conservation of Global Biological Diversity. *Conserv Biol*, 23(3):557–567, 2009. 1, 2, 14

[224] The Ramsar Convention on Wetlands. The Annotated Ramsar List: Brazil, 2011. Disponível em <http://www.ramsar.org/cda/en/ramsar-pubs-notes-annotated-ramsar-16692/main/ramsar/1-30-168-5E16692-4000-0–>. Acessada em ago. 2012. 5

[225] L. G. Underhill. Optimal and Suboptimal Reserve Selection Algorithms. *Biol Conserv*, 70(1):85–87, 1994. 22, 25, 49

[226] M. Valenzuela-Rendón. Reinforcement Learning in the Fuzzy Classifier System. Technical Report CIA-RI-031, Centro de Inteligencia Artificial, Instituto Tecnológico y de Estudios Superiores de Monterrey, Mexico, 1997. 4, 62

[227] D. A. van Veldhuizen and G. B. Lamont. Multiobjective Evolutionary Algorithms: Analyzing the State-of-the-Art. *Evol Comput*, 8(2):125–147, 2000. 51, 52, 54

[228] R. I. Vane-Wright, C. J. Humphries, and P. H. Williams. What to Protect? - Systematics and the Agony of Choice. *Biological Con*, 55:235–254, 1991. 28, 30, 42

[229] A. Vargha and H. D. Delaney. A Critique and Improvement of the "CL" Common Language Effect Size Statistics of McGraw and Wong. *J Educ Behav Stat*, 25(2):pp. 101–132, 2000. 109, 113

[230] J.-C. Vié, C. Hilton-Taylor, and S. N. Stuart. *Wildlife in a Changing World: An Analysis of the 2008 IUCN Red List of Threatened Species*. Red List Series. IUCN, 2009. 1, 14

[231] M. Villalobos-Arias, C. A. Coello-Coello, and O. Hernández-Lerma. Convergence analysis of a multiobjective artificial immune system algorithm. In G. Nicosia, V. Cutello, P. J. Bentley, and J. Timmis, editors, *ICARIS*, volume 3239 of *Lect Notes Comput Sc*, pages 226–235. Springer, 2004. 4, 62, 86

[232] M. Villalobos-Arias, C. A. Coello-Coello, and O. Hernández-Lerma. Asymptotic Convergence of Some Metaheuristics Used for Multiobjective Optimization. In Alden H. Wright, Michael D. Vose, Kenneth A. De Jong, and Lothar M. Schmitt, editors, *FOGA*, volume 3469 of *Lect Notes Comput Sc*, pages 95–111. Springer, 2005. 86

[233] Y. Wang, Z. Cai, G. Guo, and Y. Zhou. Multiobjective Optimization and Hybrid Evolutionary Algorithm to Solve Constrained Optimization Problems. *IEEE T Syst Man Cy B*, 37(3):560–575, June 2007. 4

[234] Y. Wang, J. Zhang, H. Sun, N. Ning, and L. Yang. Construction and Evaluation of a Primary Core Collection of Apricot Germplasm in China. *Sci Hortic-Amsterdam*, 128:311–319, 2011. 106

[235] M. E. Watts, I. R. Ball, R. S. Stewart, C. J. Klein, K. Wilson, C. Steinback, R. Lourival, L. Kircher, and H. P. Possingham. Marxan with Zones: Software for Optimal Conservation Based Land- and Sea-Use Zoning. *Environ Modell Softw*, 24(12):1513–1521, 2009. 42

[236] K.J. Wessels, S. Freitag, and A.S. van Jaarsveld. The Use of Land Facets as Biodiversity Surrogates During Reserve Selection at a Local Scale. *Biol Conserv*, 89(1):21–38, 1999. 15

[237] D. R. White and S. Poulding. A Rigorous Evaluation of Crossover and Mutation in Genetic Programming. In L. Vanneschi, S. Gustafson, A. Moraglio, I. De Falco, and M. Ebner, editors, *Genetic Programming*, volume 5481 of *Lect Notes Comput Sc*, pages 220–231. Springer Berlin Heidelberg, 2009. 239

[238] K. J. Willi, M. B. Araújo, K. D. Bennett, B. Figueroa-Rangel, C. A. Froyd, and N. Myers. How Can a Knowledge af the Past Help to Conserve the Future? Biodiversity Conservation and the Relevance of Long-Term Ecological Studies. *Phil Trans R Soc*, 362(1478):175–186, 2007. 14

[239] J. C. Williams and C. S. ReVelle. Applying Mathematical Programming to Reserve Selection. *Environ Model Assess*, 2(3):167–175, 1997. 18, 24, 49

[240] D. H. Wolpert and W. G. Macready. No Free Lunch Theorems for Optimization. *IEEE T Evolut Comput*, 1(1):67–82, April 1997. 55, 56

[241] P. Zhang, J. Li, X. Li, X. Liu, X. Zhao, and Y. Lu. Population Structure and Genetic Diversity in a Rice Core Collection (*tOryza sativa* L.) Investigated with SSR Markers. *PLoS ONE*, 6(12):e27565+, December 2011. 106

[242] E. Zitzler. A Tutorial on Evolutionary Multiobjective Optimization. Plenary lecture at the MOMH 2002 workshop in Paris, France, 2002. 69, 71, 72, 86, 116

[243] E. Zitzler, D. Brockhoff, and L. Thiele. The Hypervolume Indicator Revisited: On the Design of Pareto-compliant Indicators via Weighted Integration. In S. Obayashi, K. Deb, C Poloni, T. Hiroyasu, and T. Murata, editors, *Proceedings of the 4th International Conference on Evolutionary Multi-Criterion Optimization, EMO'2007*, volume 4403 of *Lect Notes Comput Sc*, pages 862–876, Matsushima, Japan, 2007. Springer-Verlag. 118, 120

[244] E. Zitzler, K. Deb, and L. Thiele. Comparison of Multiobjective Evolutionary Algorithms: Empirical Results. *Evol Comput*, 8(2):173–195, 2000. 64, 116, 117, 122, 131

[245] E. Zitzler, M. Laumanns, and L. Thiele. SPEA2: Improving the Strength Pareto Evolutionary Algorithm for Multiobjective Optimization. In K.C. Giannakoglou, D.T. Tsahalis, J. Periaux, K.D. Papaliliou, and T. Fogarty, editors, *Evolutionary Methods for Design, Optimisation and Control with Application to Industrial Problems. Proceedings of the EUROGEN2001 Conference, Athens, Greece, September 19-21, 2001*, pages 95–100, 2002. 63, 64, 69

[246] E. Zitzler and L. Thiele. An Evolutionary Algorithm for Multiobjective Optimization: The Strength Pareto Approach. Technical report, Computer Engineering and Communication Networks Lab (TIK), Swiss Federal Institute of Technology (ETH), 1998. 3, 54, 55

[247] E. Zitzler and L. Thiele. Multiobjective Optimization Using Evolutionary Algorithms – A Comparative Case Study. In *Conference on Parallel Problem Solving from Nature (PPSN V)*, pages 292–301, Amsterdam, 1998. 120

[248] E. Zitzler and L. Thiele. Multiobjective Evolutionary Algorithms: A Comparative Case Study and the Strength Pareto Approach. *IEEE T Evolut Comput*, 3(4):257–271, 1999. 3, 4, 63, 118

[249] N. Ziviani. *Projeto de algoritmos: com implementação em Pascal e C*. Cengage Learning, 3 ed. edition, 2011. 137