METODOLOGIA BASEADA EM MINERAÇÃO DE DADOS APLICADA NA ANÁLISE DA DISTRIBUIÇÃO ESPACIAL DE AGRUPAMENTOS DE ESPÉCIES VEGETAIS

Luís Alexandre Estevão da Silva

Tese de Doutorado apresentada ao Programa de Pós-graduação em Engenharia de Sistemas e Computação, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Doutor em Engenharia de Sistemas e Computação.

Orientadores: Geraldo Zimbrão da Silva
Jano Moreira de Souza

Rio de Janeiro
Dezembro de 2013

METODOLOGIA BASEADA EM MINERAÇÃO DE DADOS APLICADA NA ANÁLISE DA DISTRIBUIÇÃO ESPACIAL DE AGRUPAMENTOS DE ESPÉCIES VEGETAIS

Luís Alexandre Estevão da Silva

TESE SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA (COPPE) DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE DOUTOR EM CIÊNCIAS EM ENGENHARIA DE SISTEMAS E COMPUTAÇÃO.

Examinada por:

_____
Prof. Geraldo Zimbrão da Silva, D.Sc.

_____
Prof. Jano Moreira de Souza, Ph.D.

_____
Prof. Geraldo Bonorino Xexéo, D.Sc.

_____
Profa. Ana Maria de Carvalho Moura, D.Ing.

_____
Profa. Marinez Ferreira de Siqueira, D.Sc.

RIO DE JANEIRO, RJ - BRASIL
DEZEMBRO DE 2013

Silva, Luís Alexandre Estevão da

Metodologia Baseada em Mineração de Dados Aplicada na Análise da Distribuição Espacial de Agrupamentos de Espécies Vegetais / Luís Alexandre Estevão da Silva. – Rio de Janeiro: UFRJ/COPPE, 2013.

XVI, 130 p.: il.; 29,7 cm.

Orientador: Geraldo Zimbrão da Silva

Jano Moreira de Souza

Tese (doutorado) – UFRJ/ COPPE/ Programa de Engenharia de Sistemas e Computação, 2013.

Referências Bibliográficas: p. 107-121.

1. Mineração de dados. 2. Regras de Associação. 3. Biodiversidade. I. Silva, Geraldo Zimbrão. II. Universidade Federal do Rio de Janeiro, COPPE, Engenharia de Sistemas e Computação. III. Título.

Dedico este trabalho ao meu filho Gustavo, a Adriana, aos meus pais e ao meu irmão.

AGRADECIMENTOS

A Deus, Meu Senhor, por ter permitido que eu chegasse até este momento em minha vida. Sem suas bênçãos e consolações nos momentos mais difíceis, **com certeza**, não teria concluído o doutorado. Que todas as glórias sejam dele!

A minha esposa Adriana e ao meu filhão Gustavo por estarem ao meu lado durante todo esse tempo. A paciência e o incentivo de vocês foram fundamentais para que eu pudesse ter a tranquilidade necessária para a concretização da tese. Peço perdão pelas ausências.

Aos meus queridos pais José e Irene, pernambucanos, que como muitos outros corajosos, decidiram sair de sua terra natal em busca de melhores condições de vida para sua família. Obrigado por seu amor, exemplo de vida e disposição. Ao meu irmão Fábio pela amizade e carinho.

A minha tia Maria do Carmo e ao meu avô Antônio (In memoriam) pelo carinho e incentivo em meus estudos.

Aos meus sogros, Sr. Ivan e Dona Darci, pelo incentivo, carinho e ajuda.

Aos meus orientadores, Prof. Geraldo Zimbrão e Prof. Jano Moreira de Souza, por terem aceitado serem meus orientadores, pelo conhecimento transmitido ao longo desses anos, pela paciência, incentivo e compreensão no desenvolvimento de um doutorado em tempo parcial.

A Profa. Ana Maria de Carvalho Moura, minha orientadora no mestrado no Instituto Militar de Engenharia, com quem muito aprendi tanto em termos de conhecimento em banco de dados, quanto em conduta profissional e ética.

A Profa. Marinez Ferreira de Siqueira pela participação em minhas bancas, pelo conhecimento e pelo grande incentivo.

Aos Prof. Luís Alfredo Vidal de Carvalho e ao Prof. Geraldo Bonorino Xexéo, por terem aceitado o convite para participarem de minha banca.

Ao Instituto de Pesquisas Jardim Botânico do Rio de Janeiro pela licença parcial para a realização do doutoramento. Ao meu Diretor, Dr. Rogério Gribel, pela liberação parcial e compreensão. Ao Dr. Fábio Scarano, pelo incentivo inicial. Aos companheiros de trabalho, pelo apoio, amizade e motivação: Dalcin, Bruno Kurtz, Profa. Ariane Luna, Rafaela Forzza, Begonha, Gustavo Martinelli, Denise Pinheiro, Fabiana Filardi, Ricardo Avancini, Paula, Cyl, Ronaldo Marquete, Rejan Guedes, Erika Medeiros, Rafael, Carol, Vítor, Paulo Rogério e Alexandre Christo; pelas palavras de apoio. A Flávia Pinto pela amizade e ajuda no trabalho.

Aos amigos da Universidade Católica de Petrópolis pelo incentivo e compreensão, principalmente neste último ano. Aos Professores Giovane Quadrelli, Alexandre Sheremetief, Fabini Hoelz, Fábio Licht, Bruno Guingo, Demerson, Maria Cristina, Cristina Bernardes, Henriete, Marcelo Maller, Carlos Campeão, Fontanella, Tepedino e Bruno Tamancoldi.

Aos membros da 2ª Igreja Batista em Juscelino. Grato pelas orações!

Aos amigos da Universidade Estácio de Sá pelo incentivo e amizade: Cláudio Gimenez, Dimas, Sandro, Eppinghauss, Regina, Jorge França, Lázaro e Monteiro.

Aos amigos de doutorado: Carlos Eduardo, Luís Orleans, Ricardo Barros, Stainam, Wagner, Filipe Braida, Eduardo Ogasawara e Sérgio Manuel Serra.

Aos meus amigos de mestrado no IME: Hélio Perez e Franz Novillo Torrico.

As secretárias da linha de Banco de dados, Ana Rabello e Patrícia Leal. Aos funcionários da secretaria do PESC, em especial, a Solange e ao Gutierrez.

Aos meus alunos e ex-alunos, ao longo dos meus 20 anos de magistério completados neste ano. Fica o incentivo para vocês. Não é nada fácil, mas a sensação de realização de um sonho é algo que não se pode descrever!

Aos companheiros do Exército e demais amigos.

Por fim, a todos aqueles que tem um sonho e que sabem que a chama continua no peito, mesmo diante das dificuldades. O sonho é somente seu!

Pedi e vos será dado; buscai e achareis; batei e vos será aberto; pois todo o que pede recebe; o que busca acha e ao que bate se lhe abrirá

Mateus 7,7

Resumo da Tese apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Doutor em Ciências (D.Sc.)

METODOLOGIA BASEADA EM MINERAÇÃO DE DADOS APLICADA NA ANÁLISE DA DISTRIBUIÇÃO ESPACIAL DE AGRUPAMENTOS DE ESPÉCIES VEGETAIS

Luís Alexandre Estevão da Silva

Dezembro/2013

Orientadores: Geraldo Zimbrão da Silva
Jano Moreira de Souza

Programa: Engenharia de Sistemas e Computação

Com o objetivo de investigar padrões de coocorrências de espécies vegetais, esta tese apresenta a proposta de uma metodologia da mineração de dados para a aplicação em bancos de dados de inventários florísticos em parcelas. A tese avaliou de forma ampla as possibilidades de uso da mineração de dados corroborando nas pesquisas ecológicas com dados de múltiplas espécies, obtendo conhecimento de coocorrências, mesmo em grandes volumes de dados. Para tal, a metodologia proposta identifica padrões de coocorrências com o uso da análise de associação da mineração de dados aplicada a conjuntos frequentes de espécimes. Foi criado um repositório de regras para a persistência do conhecimento gerado, possibilitando a avaliação dos resultados experimentais obtidos por meio de métricas adequadas a esse tipo de aplicação. Estudos de casos foram realizados tornando possível o reconhecimento de padrões de correlações positivas e negativas entre pares e grupos de espécies. A posterior aplicação da análise de agrupamentos foi realizada para identificar a distribuição espacial das espécies presentes nos padrões observados. A tese comprovou por meio da metodologia que a mineração de dados pode ser um recurso útil na Ecologia.

Abstract of Thesis presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Doctor of Science (D.Sc.)

METHODOLOGY BASED ON DATA MINING APPLIED TO THE ANALYSIS OF SPATIAL DISTRIBUTION OF PLANT SPECIES GROUPINGS

Luís Alexandre Estevão da Silva

December/2013

Advisors: Geraldo Zimbrão da Silva
Jano Moreira de Souza

Department: Computer Science and Engineering

Aiming to investigate patterns of co-occurrences of plant species, this thesis presents a proposed methodology for data mining application in databases floristic inventories in plots. The thesis evaluated comprehensively the possibilities of using data mining in supporting ecological research with data from multiple species, obtaining knowledge of co-occurrences, even in large volumes of data. For such, the proposed methodology identifies patterns of co-occurrences with the use of association analysis of data mining applied to frequent sets of specimens. It was created a repository of rules for the persistence of the knowledge generated, allowing the evaluation of the experimental results using metrics appropriate for this type of application. Case studies were conducted making it possible to recognize patterns of positive and negative correlations between pairs and groups of species. The subsequent application of cluster analysis was performed to identify the spatial distribution of species present in the observed patterns. The thesis demonstrated through the methodology that data mining can be a useful resource in Ecology.

Índice

Índice de Figuras

Índice de Tabelas

Lista de Termos e Abreviações

AgNes - *Aglomerative Nesting*

ANOVA – Análise de Variância

BCI – *Barro Colorado Island*

BIRCH - *Balanced Iterative Reducing and Clustering using Hierarchies*

CLAra - *Clustering Large Applications*

CLARanS - *Clustering Large Applications based upon Randomized Search*

Clique - *Clustering High-Dimensional Space*

COMBO - *Species Combinations Score*

CURe - *Clustering Using Representatives*

DAP – Diâmetro da Altura do Peito

DBSCAN - *Density-Based Spatial Clustering of Applications with Noise*

Denclue - *Density-based Clustering*

DiAna - *Divisia Analysis*

ECLAT - Equivalence CLASS Transformation

ECOSIM - *Null Modeling Software for Ecologists*

EM - *Expectation Maximization*

ETL - *Extract, transform, load*

FDP – *Forest Dynamic Plot*

GARP - *Genetic Algorithm for Rule-set Prediction*

GBIF – *Global Biodiversity Information Facility*

JABOT – Sistema de Informações de Coleções Científicas do Instituto de Pesquisas

Jardim Botânico do Rio de Janeiro

KDD - *Knowledge Discovery in Databases*

LHS – *Left Hand Side*

OPtiCS - *Ordering Points to Identify the Clustering Structure*

PAM - *Partitioning Around Medoids*

PPA - Análise de Padrões de Pontos

RHS – *Right Hand Side*

RLD - *Relative Linkage Disequilibrium*

SAM - *Spatial Analysis in Macroecology*

SiBBR - Sistema de Informação sobre a Biodiversidade Brasileira

SOM – *Self-Organizing Map*

SpeciesLink - Sistema de Informação Distribuído para Coleções Biológicas

Sting - *Statistical Information Grid Approach*

Twinspan - *Two-Way Indicator Species Analysis*

UML – *Unified Modeling Language*

UPGMA - *Unweighted Pair Group Method with Arithmetic Averages*

WaveCluster - *Clustering Using Walete Transformation*

WPGMA - *Weighted Pair Group Method with Arithmetic Means*

Capítulo 1 – Introdução

1.1 Motivação

As mudanças ocorridas na natureza pela ação do homem vêm provocando o aumento das taxas de extinção de espécies de uma forma preocupante, o que pode ser comprovado numericamente pela redução da cobertura original de vários *hotspots* da biodiversidade (BROOKS *et al.*, 2002). Diversas são as razões das modificações realizadas nos ecossistemas, destacando-se as mudanças climáticas (WALTHER *et al.*, 2002). Essa preocupação com o meio ambiente e a necessidade de pesquisas para a melhor compreensão dos sistemas de biodiversidade fez com que ao longo dos anos, bases de dados cada vez maiores fossem criadas, tais como o *Global Biodiversity Information Facility* – GBIF (TELENIUS, 2011), a rede *SpeciesLink*[1] e o sistema *Jabot*[2] do Instituto de Pesquisas Jardim Botânico do Rio de Janeiro. Projetos visando o inventário de espécies também foram iniciados, como por exemplo o *Systematics Agenda 2000* (CLARIDGE, 1995) que objetiva a descoberta, descrição e inventário global de espécies, sintetizando resultados, permitindo uma classificação preditiva e, possibilitando o desenvolvimento de um sistema de informação adequado para gerir a informação resultante. No Brasil, o *Sistema de Informação sobre a Biodiversidade Brasileira* - SiBBr[3] começou a ser desenvolvido para a integração de fontes nacionais e internacionais subsidiando tomadores de decisão na preservação da flora.

Com base nos dados armazenados em sistemas com essas características, pesquisadores têm aplicado recursos tecnológicos para fins de avaliação da ocorrência de espécies como uma tentativa de determinar regiões onde elas têm as melhores

[1] www.splink.cria.org.br
2 www.jbrj.gov.br/jabot
3 www.sibbr.gov.br

condições de ocorrer e, consequentemente, tomar providências para a conservação desses biomas. Tais recursos possibilitaram acelerar as pesquisas e análises sobre a biodiversidade por meio de predições baseadas no uso de modelos computacionais (SIQUEIRA, 2005). O uso da computação também é justificado com o intuito de auxiliar na promoção do monitoramento de espécies, tendo em vista a escassez de recursos financeiros e de pessoal (PEIXOTO, MORIM, 2002, HOPKINS, FRECKLETON, 2002, DREW, 2011); acrescentando que em algumas regiões os levantamentos florísticos, tão necessários para o conhecimento da distribuição da vegetação de uma região, ainda nem foram realizados ou não foram concluídos.

Na busca por respostas que levem a uma melhor compreensão da natureza, são realizados levantamentos florísticos nas regiões selecionadas para os estudos. Porém, os levantamentos são difíceis de serem realizados, principalmente: pela carência de taxonomistas, biólogos que realizam a correta identificação das espécies, pelas dificuldades de acesso a determinadas regiões e obtenção de recursos financeiros para a realização das excursões de campo, entre outros fatores. Observa-se então, como consequência, que muitos dos levantamentos são apenas parciais, tendo em vista que foram realizados com dados de ocorrências de espécies de apenas parte da vegetação de uma região, ou seja, implicando em um mapeamento incompleto para o país detentor de uma das maiores biodiversidades do planeta, algo em torno de 10% da biota do mundo (MACHADO *et al.*, 2004, FORZZA *et al.*, 2012). Considera-se ainda importante mencionar que os taxonomistas, normalmente, realizam os levantamentos observando apenas as espécies necessárias aos estudos relacionados com suas atividades de pesquisas, por isso algumas regiões ainda não se encontram devidamente levantadas e, as análises quanto à diversidade dessas regiões ainda não disponibilizadas. Desta forma, nos estudos ecológicos, normalmente são usados índices de medição da biodiversidade

(MAGURRAN, 2011) para efetuar a avaliação da riqueza de ocorrência das espécies vegetais a partir de amostras da vegetação de uma região.

No levantamento do estado-da-arte das técnicas usadas nos estudos ecológicos, poucos trabalhos foram observados em relação à manipulação eficiente e extração automática de conhecimento de bancos de dados da biodiversidade, podendo isto ser considerado como uma dificuldade para o desenvolvimento das pesquisas na flora, especificamente por causa da manipulação de grandes bases de dados. A questão então se direciona em como transformar em *conhecimento*, os milhões de registros que estão armazenados nos bancos de dados da biodiversidade?

Em relação às iniciativas de uso de técnicas para agilizar as pesquisas, a Biologia tem buscado utilizar um número maior de ferramentas computacionais para tal propósito, e um exemplo disto é a *Modelagem de Distribuição de Espécies* (PEARSON, 2007). Por meio desta técnica, diversos trabalhos objetivando a definição de modelos preditivos da flora têm sido realizados. No entanto, os maiores avanços nessas pesquisas foram obtidos nos modelos criados para verificar a distribuição potencial de cada uma das espécies isoladamente, surgindo então a busca por novas abordagens nas análises de padrões de ocorrências de múltiplas espécies. O interesse para as pesquisas no nível de múltiplas espécies foi motivado pelas possibilidades de reconhecimento dos padrões que podem ser extraídos considerando a complexidade das análises que envolvem, normalmente, o uso de uma grande quantidade de variáveis bióticas e abióticas. Esse interesse pode ser comprovado pelo aumento do número de trabalhos realizados em aplicações para a modelagem espacial da biodiversidade no nível de comunidades em diversas categorias, tais como: análise da riqueza de espécies; estudo da caracterização dos diversos tipos de comunidades, tanto quanto a dissimilaridade de composição de

espécies, entre outras possibilidades (FERRIER, GUISAN, 2006, ARAUJO, BASELGA, 2010).

Ainda como motivação para o desenvolvimento desta tese foram verificados outros pontos de dificuldade das pesquisas ecológicas apresentados em AUSTIN (2007) e listadas a seguir: os problemas da definição da escala adequada para o estudo, a seleção de variáveis bióticas, a seleção de preditores ambientais, a adequação dos métodos em relação as teorias ecológicas, os tipos de respostas em face da adequação ambiental, a comparação e avaliação dos métodos e o uso de dados artificiais.

Por outro lado, na área de banco de dados, muitos avanços foram realizados com o uso da Mineração de Dados (HAN *et al*., 2011), principalmente por causa do contínuo aumento do volume de dados e do potencial de conhecimento presente nessas bases. Os recursos da técnica podem ser especialmente úteis quanto à necessidade de manipulação de grandes quantidades de dados, pois permite a extração de padrões por meio da aplicação de diversos tipos de algoritmos, para posterior avaliação por especialistas da área fim. Na mineração de dados, esses algoritmos são organizados em diferentes categorias, tais como: classificação, associação, *clusterização* (análise de agrupamentos), análise de sequências e redes neurais. Dentre as classes de algoritmos apresentadas, consideramos que o uso da análise de associação e da análise de agrupamentos, de uma forma conjunta, pode promover uma melhoria no processo de extração do conhecimento de coocorrências de espécies de uma forma alternativa aos métodos existentes (descritos no Capítulo 2). Os algoritmos de associação e agrupamentos serão avaliados no Capítulo 3.

Considera-se assim, que de uma forma complementar, a dificuldade no entendimento da complexidade da biodiversidade poderá ser reduzida pelo *insights*

obtidos com o uso de algoritmos e de uma metodologia adequada às especificidades das aplicações ecológicas pois, ferramentas de mineração de dados aplicam técnicas estatísticas a grandes quantidades de dados armazenados, a fim de procurar por padrões nesses dados (DATE, 2000). Na própria Ecologia podemos verificar a busca por *insights* como em GERING, CRIST (2002). A análise com a mineração de dados permite sintetizar complexos e volumosos bancos de dados, tornando-os mais interpretáveis para cientistas e tomadores de decisão, em consonância com as medidas tomadas para melhorar o gerenciamento diante do desafio do aumento constante da quantidade de dados dos bancos de dados da biodiversidade. O uso da técnica poderá facilitar o desenvolvimento de trabalhos que poderão ser úteis tanto para os responsáveis pela gestão da flora quanto para a realização de pesquisas científicas.

Portanto, como motivações para a presente pesquisa foram listadas a dificuldade no estudo de múltiplas espécies e a extração de conhecimento dos bancos de dados da biodiversidade. Sendo levantada, a hipótese de que os recursos da mineração de dados podem apoiar de forma eficiente e complementar a análise de distribuição de espécies, reconhecendo padrões de coocorrências entre pares e grupos de espécies. Adicionalmente, apresentamos a biologia para a mineração de dados, como um campo promissor de realizações de pesquisas que podem servir como pré-hipóteses a serem avaliadas por especialistas da área.

1.2 Definição do Problema

Apesar de toda a evolução alcançada na Biologia, ainda existem perguntas cujas respostas não são conclusivas. Dentre essas questões, talvez a mais óbvia para um ecólogo seja a compreensão do motivo pelo qual o número de espécies varia no tempo e no espaço (TOWNSEND *et al*., 2009). Assim, para o ecólogo, observar os padrões

ocorridos na natureza é uma forma de buscar as razões pelas quais as espécies ocorrem. Contudo, para aumentar ainda mais a complexidade dos estudos na área, o autor afirma que "[...] *os padrões na riqueza das espécies têm sido modificados de várias formas pelas atividades humanas, tais como o desenvolvimento no uso do solo, a poluição e a introdução de espécies exóticas*". Essas observações demonstram que ao mesmo tempo que há a necessidade de uma melhor compreensão dos relacionamentos e padrões presentes na biodiversidade, suportando e complementando os estudos florísticos realizados pelos taxonomistas, há também uma necessidade cada vez maior de agilizar esse processo diante das constantes modificações ocorridas na natureza pela ação do homem. Neste sentido, podemos compreender as razões que levaram a comunidade científica a utilizar modelos baseados em algoritmos computacionais. Este fato pode ser comprovado pelo crescente uso nas últimas décadas das técnicas de modelagem de distribuição de espécies, envolvendo a análise ambiental para avaliar regiões onde as espécies têm condições ambientais semelhantes àquelas onde foram localizadas.

Porém, a modelagem encontra dificuldades quanto à manipulação de grandes volumes de dados e na definição de modelos eficientes nas pesquisas com múltiplas espécies pois, como citado anteriormente, os maiores avanços foram realizados nos modelos criados em nível de espécies isoladas (FERRIER, GUISAN, 2006). Dificuldades que podem ser compreendidas pela complexidade que envolvem os processos que governam a composição das comunidades (SWENSON, 2013) em relação ao pouco que é conhecido sobre os fatores responsáveis por tais estruturas, especialmente em florestas tropicais (URIARTE *et al*., 2004). Havendo assim também, uma necessidade de conhecimento no nível de comunidade. Isto pode ser verificado em ARAUJO, BASELGA (2010), que diz que apesar da modelagem em nível de múltiplas

espécies ter sido apresentada como uma área de grande interesse, seus resultados ainda necessitam de melhoramentos.

Ainda em relação ao estudo em nível de comunidades, destacam-se também as pesquisas de coocorrências de espécies, especialmente pelas possibilidades relativas a avaliação de hipóteses quanto as interações entre espécies, como por exemplo, a competição e mutualismo. Alguns importantes estudos foram realizados para as análises de coocorrências de espécies, como por exemplo, Modelos Nulos (GOTELLI, 2000, GOTELLI *et al.*, 2009, GOTELLI *et al.*, 2007, DINIZ *et al.*, 2002) comprovando a necessidade por esse tipo de pesquisa.

Adicionalmente aos problemas apresentados anteriormente, há a necessidade da definição de uma metodologia adequada para a implementação de um modelo que permita a extração do conhecimento relacionado a coocorrências de espécies dos bancos de dados. Assim como, propor mais uma opção de *software* que possibilite um suporte aos ecólogos e taxonomistas em suas pesquisas, tendo em vista que há um limitado número de opções de programas atualmente, como por exemplo o Twinspan (HILL, 1979) e o Ecosim (ENTSMINGER, 2012). Esses *softwares* serão discutidos no fim do Capítulo 2.

A seguir, resumidamente, estão listados os problemas identificados nas pesquisas relacionadas com a extração de padrões em bases de dados de múltiplas espécies:

- Dificuldade de avaliação de padrões de ocorrências de espécies em razão das constantes mudanças ocorridas na biodiversidade pela ação do homem, caracterizando a necessidade da definição de um modelo de extração de conhecimento automático para

agilizar na busca da compreensão da natureza; principalmente diante de grandes volumes de dados;

- Dificuldade na extração de padrões e obtenção de conhecimento em bases de dados de múltiplas espécies, pois a modelagem de distribuição de espécies em nível de comunidade ainda apresenta resultados insatisfatórios, caracterizando desta forma a necessidade de mais métodos com o objetivo de possibilitar uma melhor compreensão da natureza, subsidiando as análises sobre as hipóteses ecológicas existentes;

- Necessidade da oferta de novas opções de *softwares* para permitir a análise de coocorrências de espécies vegetais por ecólogos, tendo em vista a quantidade limitada de opções atuais de programas com esse fim e que, em suas concepções, não foram projetados para análises em grandes bases de dados.

1.3 Hipótese

A pesquisa realizada nesta tese tem por base a hipótese de que o desenvolvimento de uma metodologia, definida com o uso da mineração de dados e aplicada com algoritmos da análise de associação e de agrupamentos, pode identificar padrões espaciais de coocorrências de espécies vegetais em amostras de dados de parcelas. Sendo que esse conhecimento pode ser utilizado: no levantamento de espécies indicadoras da ocorrência de outras espécies, assim como na geração de conhecimento associado ao habitat preferencial das espécies envolvidas, entre outras possibilidades de pesquisas.

Para a formulação da hipótese, foram estudadas algumas premissas ecológicas, como os fundamentos observados nas análises filogenéticas e funcionais e ainda, o conceito de espécies indicadoras, que são aquelas espécies que, devido às suas

preferências de nicho, podem ser utilizadas como indicadores ecológicos de tipos de comunidades, condições de habitat ou mudanças ambientais (CACERES *et al*., 2010).

Na Ecologia, os estudos orientados para a determinação dos processos que caracterizam a diversidade e a composição das comunidades têm sido constantes. Essencialmente, é assumido que espécies que ocorrem em uma mesma região devem ser, em média, mais similares filogeneticamente do que as espécies distantes geograficamente (SWENSON, 2013). Desta forma, foi usado como método para a identificação das coocorrências de espécies, a análise de associação, pois foi considerado que essa categoria de algoritmos da mineração de dados pode contribuir consideravelmente ao tipo de análise pretendida; fazendo uma alusão às análises realizadas com ocorrências de itens frequentes (AGRAWAL *et al*., 1993). Em nossa pesquisa, os itens frequentes são os espécimes das árvores identificadas em levantamentos florísticos.

Como resultado da análise de associação, espera-se que padrões de coocorrências de espécies sejam identificados. Após esta etapa e, para melhor identificação dos agrupamentos formados, foi considerada necessária a aplicação da análise de agrupamentos (clusterização) na pesquisa, tendo em vista que, na própria Biologia, a ideia de particionar o problema por meio de agrupamentos vem sendo usada há muito tempo. Segundo Valentim (VALENTIN, 2000), "*os ecólogos tem uma tendência normal em procurar agrupar amostras de mesmas características bióticas e abióticas, ou associar espécies em comunidades, de acordo com o objetivo do seu trabalho*". Outros fatores que corroboram quanto a importância da análise de agrupamentos de espécimes faz alusão a identificação das regiões com condições ambientais semelhantes ou ecorregiões (OLSON *et al.*, 2001), ou ainda micro-habitats, dependendo da escala utilizada. Considera-se ainda que pela sua própria natureza, o

levantamento de agrupamentos pode ser especialmente útil para a obtenção de conhecimento em base de dados com relacionamentos complexos entre dados abióticos, bióticos, espaciais e, a identificação da relativa importância dos conjuntos de variáveis candidatas explanatórias (HOCHACHKA *et al.*, 2007). No Capítulo 4, é realizada a descrição detalhada da metodologia da aplicação proposta, com destaque para os componentes da análise de associação e da análise de agrupamentos e, as justificativas teóricas para a seleção desses tipos de análises da mineração de dados.

1.4 Objetivo da Tese

Em função do problema apresentado, esta tese tem como objetivo principal a proposta de uma metodologia, baseada na mineração de dados, para a definição de um processo de extração de conhecimento de padrões espaciais de coocorrências de pares ou grupos de espécies vegetais em bases de dados de inventários florísticos em parcelas. Com a implementação da mesma sendo realizada com o uso dos algoritmos da análise de associação e de agrupamentos. A proposta inclui, além da execução das fases de captura e transformação dos dados para um formato necessário à aplicação do algoritmo de associação; as etapas de aplicação do algoritmo de associação e do algoritmo de clusterização. A implementação contempla ainda a geração e visualização dos resultados obtidos. A proposta inova quanto à aplicação da mineração de dados no processo de levantamento de coocorrências entre espécies e em especial, nos padrões de grupos de espécies.

O objetivo principal pode ser dividido em 4 objetivos específicos:

Objetivo 1 – Desenvolver um método para transformar os dados de parcelas em um formato transacional; formato este que possibilita a aplicação do algoritmo da análise de associação. O método tem ainda por objetivo ser flexível o suficiente para ser

aplicado a múltiplas distâncias entre os pontos de ocorrência dos espécimes envolvidos na análise;

Objetivo 2 – Desenvolver um método para aplicar a análise de agrupamentos a partir dos resultados obtidos com a análise de associação, identificando coocorrências positivas e negativas entre pares ou agrupamentos de espécies;

Objetivo 3 – Desenvolvimento de uma base de dados para servir de repositório das regras obtidas com a análise de associação; preservando e facilitando o acesso ao conhecimento obtido;

Objetivo 4 – Avaliar, a partir de um conjunto de métricas da análise de associação, aquelas mais adequadas para a identificação das coocorrências de espécies em pares e grupos.

1.5 Delimitação de Escopo

Este trabalho tem em seu escopo a definição de uma metodologia apropriada para aplicação em dados de levantamentos florísticos de parcelas, caracterizando assim um escopo em escala local. O escopo tem como justificativa a definição da parcela como a unidade de estudo florístico, sendo esta estrutura uma forma abrangente e precisa de amostragem para um grande número de ocorrências de espécimes. Por parcela ou *plots* podemos considerar uma região dividida em parcelas menores ou *subplots* de mesma dimensão, por exemplo, com parcelas de 20 metros^2.

1.6 Metodologia de Pesquisa

A fundamentação desta tese foi realizada na área de Banco de Dados. O trabalho adota o método de pesquisa quantitativo e a estratégia de pesquisa está baseada em estudos de casos. A seguir são listadas as fases da metodologia de pesquisa aplicada na

tese e na Figura 1 são resumidas as etapas realizadas, sendo que na mesma são associados os objetivos de cada uma das fases.

1ª fase - Seleção do tema e formulação do problema – nesta etapa foi feito o estudo dos conceitos ecológicos necessários à fundamentação da pesquisa, a formulação da hipótese; além do planejamento da tese;

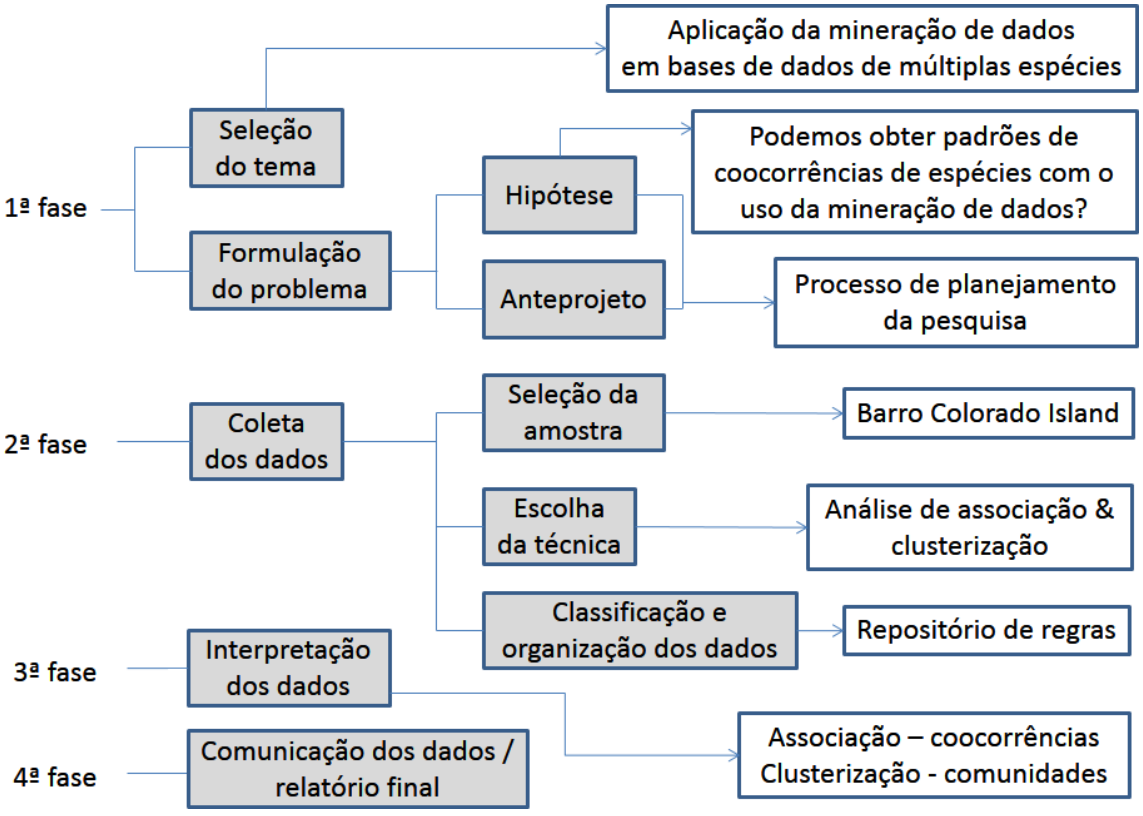


Figura 1. Metodologia de pesquisa usada na tese - adaptada de BARROS, LEHFELD (2007)

2ª fase – Foi realizada nesta etapa uma escolha criteriosa da base de dados a ser analisada. Para isso, foi realizado o contato para uso permissionado com o *Smithsonian Institute*[4] para o uso dos dados do Projeto de Floresta Dinâmica (FDP) da Ilha de Barro Colorado (BCI) no Panamá (HUBBELL *et al.*, 2005). Uma base de amplo conhecimento da comunidade ecológica, tendo seus dados classificados como de alta qualidade como resultado dos censos realizados a cada 5 anos no *plot*. Nesta etapa

[4] http://www.si.edu/

também foi realizada a avaliação das técnicas (algoritmos) a serem aplicadas sobre os dados e, a definição do repositório de regras;

3ª fase – Os resultados foram avaliados em termos das coocorrências identificadas entre espécies e importantes para a definição de uma etapa posterior de uso da clusterização, considerando facilitar a determinação das comunidades ou micro-habitats por meio dos agrupamentos espaciais formados;

4ª fase – Apresentação dos resultados, tanto para a análise de associação quanto para a análise de agrupamentos.

1.7 Organização da Tese

Este capítulo apresentou a introdução do trabalho, destacando a motivação baseada na aplicabilidade da solução, a definição do problema, a hipótese considerada para a proposta, os objetivos da pesquisa, a delimitação do escopo e a metodologia utilizada.

No Capítulo 2 são apresentadas as metodologias mais usadas atualmente para a avaliação da distribuição de espécies vegetais com o objetivo de demonstrar a necessidade na busca por este conhecimento, suas características e limitações, principalmente, quanto a manipulação de grandes bancos de dados. São descritos também os fundamentos teóricos do trabalho por meio do estudo da análise de distribuição espacial de espécies vegetais, permitindo associar à teoria a proposta em desenvolvimento.

No Capítulo 3 é apresentada a mineração de dados e as categorias da análise de associação e da análise de agrupamentos. É realizada uma comparação de algoritmos, de

modo a avaliar e justificar o uso dessas categorias, assim como os algoritmos selecionados para a aplicação na proposta apresentada.

No Capítulo 4 é apresentada a proposta da metodologia para a solução do problema, no qual os algoritmos da mineração de dados são aplicados de modo a extrair os padrões de coocorrências a partir de banco de dados de múltiplas espécies. No capítulo, todas as etapas da metodologia são descritas em detalhe.

No Capítulo 5 são apresentados os testes e as avaliações dos estudos de casos realizados para a análise par-a-par e de agrupamentos de espécies, com correlação positiva e negativa.

No Capítulo 6 são apresentadas as considerações finais, destacando as contribuições obtidas com a presente pesquisa e as sugestões de trabalhos futuros que podem ser desenvolvidos a partir da tese.

No Apêndice A é apresentada a permissão de uso dos dados de Barro Colorado. Nos Apêndices B, C e D são apresentados estudos de casos para coocorrências positivas e negativas em pares de espécies e, em grupos, respectivamente.

No Apêndice E é apresentada parte do código fonte usado na codificação da metodologia.

Um resumo da estrutura da tese é apresentado na Figura 2.

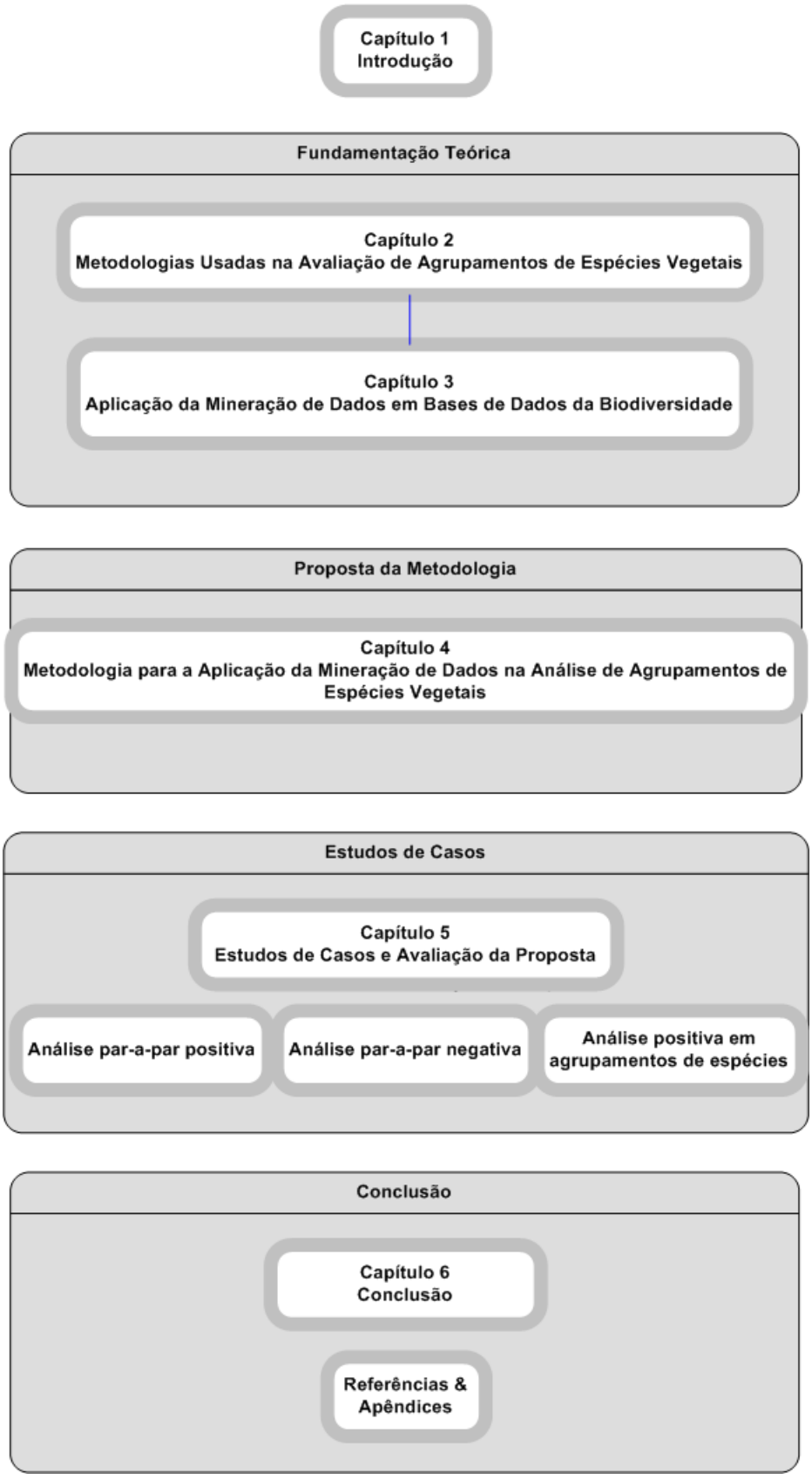


Figura 2. Resumo da organização da tese

Capítulo 2 – Metodologias Usadas na Avaliação de Agrupamentos de Espécies Vegetais

2.1 Fundamentação Ecológica sobre Comunidades

A natureza sempre foi um vasto campo de pesquisas para o Homem e há muito tempo a busca por padrões vem sendo tentada para uma melhor compreensão da biodiversidade. Inúmeros pesquisadores avaliaram a distribuição espacial das espécies pelo planeta, e ainda em 1807, já havia sido observado que tanto a composição quanto a estrutura da vegetação apresentam mudanças em relação aos gradientes latitudinais (BORGES, 2011).

Por meio desses estudos, trabalhos vêm sendo realizados a partir dos dados abióticos relacionando-os com a ocorrência de espécies, comprovando que essa ocorrência está diretamente associada à heterogeneidade espacial (TOWNSEND *et al.*, 2009). Exemplos de dados abióticos normalmente utilizados nas pesquisas são: os gradientes latitudinais (um dos mais conhecidos), os gradientes altitudinais, a variação climática, os tipos de solos, a luz, os recursos hídricos, entre outros. Esse conjunto de fatores possibilita verificar que regiões onde há uma grande riqueza de recursos, tendem a possuir também uma maior riqueza de espécies. Ou seja, as diferenças de temperatura, umidade e evapotranspiração, entre outros fatores, determinam com frequência comunidades inteiramente diferentes. O presente estudo objetiva avaliar os padrões existentes em bases de dados de múltiplas espécies vegetais, mais especificamente em relação às suas coocorrências, sem considerar as hipóteses que justificam essas coocorrências.

Tendo em vista que o conhecimento a ser explorado no estudo baseia-se em comunidades vegetais, buscamos na literatura a fundamentação teórica para o devido suporte. O conceito de comunidade é importante na Biologia e define-se uma *Comunidade Biótica* como qualquer conjunto de populações que vivem numa área determinada ou habitat físico. Essas estruturas são entendidas como conjuntos mais ou menos homogêneos de plantas pertencentes a táxons (nomes científicos) distintos, que coexistem e interagem em um espaço ou território determinado, denominado biótopo. Em segundo lugar, as comunidades estão sujeitas a um conjunto de fatores mesológicos de natureza biótica e abiótica designado por habitat (CAPELO, 2003).

No estudo das comunidades vegetais, importantes conceitos são estudados, tais como a composição estrutural, interações competitivas, o habitat, a diversidade e os padrões de espécies. A seguir analisaremos, resumidamente, as características principais das comunidades e os conceitos mais relacionados com nosso estudo. As comunidades podem então ser analisadas quanto à:

2.2.1 Composição Estrutural

A composição de uma comunidade e seus relacionamentos, pelo menos as maiores comunidades (GOMES, 2004), é formada por conjuntos de espécies que controlam em maior ou menor grau o consumo de recursos dentro da comunidade, afetando o ambiente de todas as outras espécies. Essas espécies são designadas *espécies dominantes*. O grau em que o domínio está centrado em uma ou várias espécies pode ser determinado por um *índice de dominância*. Ou seja, uma comunidade basicamente caracteriza-se quanto a sua composição por apresentar um conjunto de espécies raras com baixa abundância e poucas espécies dominantes com alta abundância.

Como compartilham do mesmo espaço físico, podem disputar esses recursos de uma forma diferenciada ou não. Cada espécie está sujeita no ambiente aos componentes abióticos e bióticos de formas singulares, exercendo influência direta sobre seu desenvolvimento e até sobre a sua existência. A competição nem sempre leva à exclusão da espécie, mas pode interferir gerando a diminuição do número de indivíduos.

2.2.2 Interações Competitivas

As interações entre as espécies são normalmente mais intensas em grandes escalas espaciais, enquanto filtros ambientais predominam em escalas espaciais pequenas. Isso porque as interações competitivas e os filtros ambientais conduzem a predições opostas em relação às diferenças funcionais e distâncias filogenéticas entre espécies. As interações competitivas tendem a limitar a similaridade funcional das espécies coocorrentes, ao contrário dos filtros ambientais, como o solo, que tendem a reunir espécies que possuem similaridades em termos funcionais. Em relação à filogenia, os filtros ambientais tendem a reunir espécies filogeneticamente aparentadas, ao passo que as interações competitivas tendem a reunir espécies filogeneticamente distintas. Com isso, podemos compreender que os fatores ecológicos ou geográficos predominam na estruturação de comunidades vegetais (ODUM, 2007). É possível presumir portanto que tais fatores podem ser considerados como determinantes para a existência de coocorrências de espécies em seus habitats, tanto positivamente quanto negativamente.

2.2.3 Habitat

A análise de uma comunidade parte da caracterização do ambiente natural, demarcado pela identificação da região geográfica onde as espécies se desenvolvem. No estudo da Biologia de Comunidades dois índices principais são apresentados: a

diversidade *Alfa*, que é definida como um grande número de espécies ocorrendo em escalas locais em florestas tropicais e; a diversidade *Beta* (WHITTAKER, 1960), que descreve como a composição de espécies muda de uma área para outra, sendo esta última ainda pouco estudada. Na avaliação das causas da diversidade das comunidades, dois processos ecológicos são geralmente usados para explicar as diversidades Alfa e Beta das comunidades: as interações competitivas e os filtros ambientais.

2.2.4 Diversidade Funcional

Medidas de diversidade filogenética também podem ser usadas para analisar os processos ecológicos que são os responsáveis pela estruturação de uma comunidade. Ecologicamente, as comunidades são assembléias de espécies coocorrentes que apresentam interação entre as espécies que a compõem. Elas são o resultado não apenas de processos ecológicos, mas também da competição que ocorre entre as espécies, da aplicação dos filtros ambientais e também dos processos de evolução aos quais foram submetidas essas espécies (CIANCIARUSO *et al.*, 2009, STEINMANN *et al*., 2009).

Apesar do longo tempo dos estudos ecológicos, ainda não há uma explicação conclusiva sobre a relação produtividade e riqueza de espécies, sendo que avaliar os padrões de riqueza continua sendo um grande desafio para a Biologia moderna (TOWNSEND et al., 2009). Os resultados do estudo realizado na tese podem interessar, além da Biologia, a Biogeografia que é um outro ramo da Biologia que também busca explicações para a distribuição dos seres vivos no planeta. Nessa área de pesquisa, fatores espaço-temporais e históricos são considerados como causas para a formulação de hipóteses explanatórias. A Tabela 1, resume os importantes conceitos teóricos sobre comunidades que possibilitaram os alicerces para o desenvolvimento desta pesquisa.

Tabela 1 – Principais conceitos sobre comunidades vegetais

Conceito	Definição
Agrupamentos	Organismos na natureza apresentam distribuição agregada e outros tipos de padrões espaciais
Caracterização	As comunidades podem ser caracterizadas pelo índice de abundância
Competição	Os filtros ambientais tendem a reunir espécies filogeneticamente aparentadas
Confiança	Fontes mais seguras do que espécies isoladas
Coocorrência	Existe certa probabilidade de que dadas espécies ocorram juntas tendo em vista os mesmos recursos que necessitam para suas existências
Grupos funcionais	Comunidades são assembléias de espécies coocorrentes que apresentam interação entre as espécies
Limites	Comunidades são definidas com precisão e distinguem-se umas das outras
Localização	Qualquer conjunto de populações que existe em uma área ou habitat físico

Desta forma, diversas metodologias têm sido pesquisadas na busca por hipóteses sobre como as espécies variam espacialmente na natureza. Considerando o uso de recursos computacionais, cada vez mais disponíveis, grandes avanços foram obtidos nesse sentido. Em termos de escalas, os estudos podem ser divididos em:

- Escala continental - para as análises macroecológicas e de mudanças climáticas;
- Regional - para a compreensão de uma determinada região ecológica ou trabalhos relacionados com o planejamento conservacional;
- Local – para a identificação de habitats e micro-habitats

2.2.5 Modelagem de Distribuição de Espécies

Segundo Elith "*Modelos de distribuição de espécies são ferramentas numéricas que combinam a observação da ocorrência de espécies ou a abundância com estimativas ambientais*" (ELITH, LEATHWICK, 2009). Essas ferramentas suportam

uma grande diversidade de aplicações e estão em uso na Biologia, com diversos objetivos diferentes, há pelo menos duas décadas.

Na modelagem de distribuição de espécies, os seguintes passos são normalmente realizados: (1) localização dos pontos de ocorrência e não ocorrência, se disponíveis; (2) levantamento dos valores dos preditores, tais como o clima; (3) modelagem dos valores ambientais para os pontos de presença ou outra medida como a abundância de espécies; (4) o modelo gerado é então usado para predizer os possíveis pontos de presença da espécie em estudo na(s) área(s) de interesse ou em outro clima, para o caso de estudos de mudanças climáticas (HIJMANS, ELITH, 2011). Obviamente, todos os métodos de modelagem apresentam melhores resultados em face da qualidade dos dados usados (GRAHAM *et al.*, 2007) e, também com a escala utilizada e, se existe uma quantidade razoável de dados para o desenvolvimento das pesquisas (WISZ *et al.*, 2008). Existem dois tipos de modelos de distribuição de espécies: o primeiro trata dos métodos que usam dados de ausência e presença, de modo a examinar a associação estatística com o ambiente e; o segundo, utiliza apenas dados de presença. Os métodos que usam somente dados de presença são mais recentes e são capazes de lidar com conjuntos de dados incompletos (GIOVANELLI *et al*., 2010, ELITH *et al*., 2006).

Diversas são as aplicações da modelagem. Na Tabela 2 são apresentados alguns tipos de pesquisas realizadas com a modelagem para fins de conservação de espécies. Essas aplicações podem servir de base para a definição dos usos da metodologia proposta na tese, principalmente de uma forma complementar ao processo de modelagem atual ou as outras formas de estudos em comunidades vegetais.

Tabela 2 – Usos da modelagem para a conservação de espécies. Fonte (PEARSON, 2007)

Tipos de uso
Levantamento de novos sítios para encontrar populações de espécies conhecidas
Levantamento de campos necessários para encontrar populações de espécies desconhecidas
Projeção dos impactos potenciais das alterações climáticas
Predição do impacto de espécies invasoras
Exploração de mecanismos de especiação
Apoio à seleção de prioridades em conservação e na determinação de reservas
Delimitação da ocorrência de espécies
Avaliação dos impactos da mudança de cobertura do solo sobre a distribuição das espécies
Validação da teoria ecológica
Comparação da paleodistribuição e filogeografia
Definição de guias para a reintrodução de espécies em perigo
Avaliação do risco de doenças

Quanto à classificação em níveis de aplicação, a modelagem pode ser classificada no nível de espécies simples ou múltiplas espécies (comunidades). Segundo GUISAN, FERRIER (2006), a modelagem em nível de comunidade pode conferir benefícios significativos para aplicações que envolvem uma grande quantidade de espécies, especialmente quanto as espécies raras (com baixa quantidade), pois enquanto na modelagem em nível de espécies essas são retiradas da amostra, por razões estatísticas, muitas estratégias de modelagem de comunidades utilizam todos os dados disponíveis. Os dados das espécies comuns podem auxiliar no suporte à modelagem das espécies menos frequentes. A estratégia pode ainda, por produzir informação sobre padrões em senso coletivo, sintetizar dados complexos de grande quantidade de espécies de uma forma simples para cientistas e tomadores de decisão.

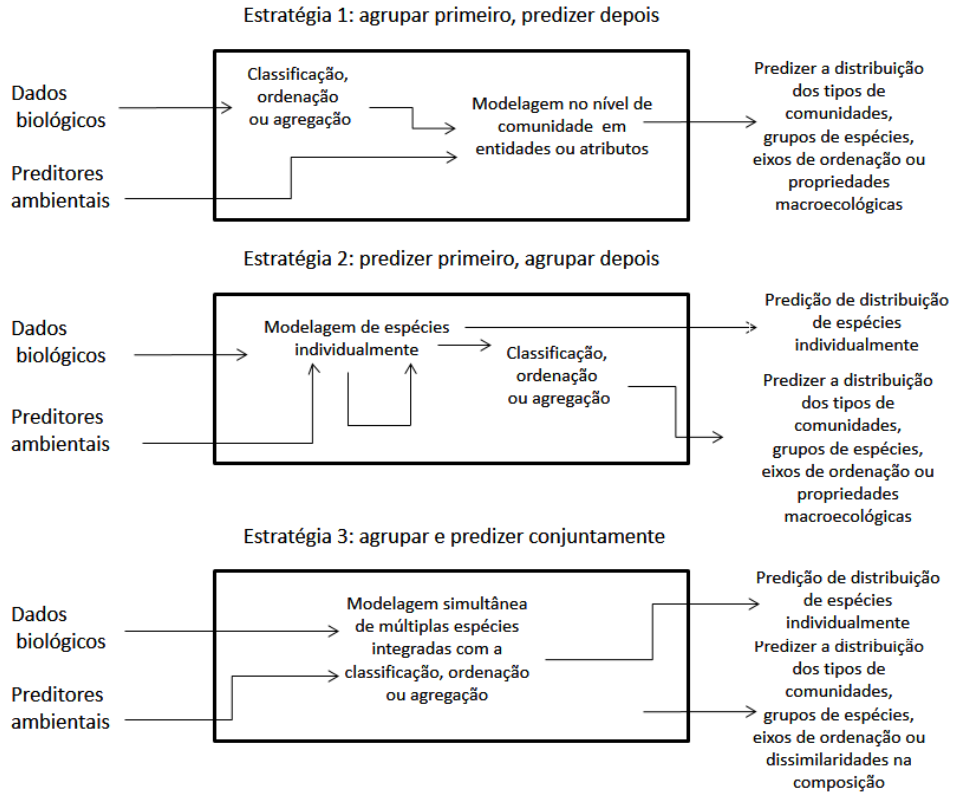


Figura 3. Estratégias de modelagem espacial em nível de comunidade

Porém, apesar das vantagens prometidas anteriormente, um trabalho realizado por (ARAÚJO, BASELGA, 2010), apresenta resultados insatisfatórios para as os testes com as três estratégias definidas por (FERRIER, GUISAN, 2006) e exibidas na Figura 3: agrupar primeiro, predizer depois; predizer primeiro, agrupar depois; e, agrupar e predizer conjuntamente. O que permite concluir que mais pesquisas necessitam ser realizadas na área. Na Tabela 3, são apresentados alguns tipos de usos quando a modelagem é realizada para comunidades de espécies.

2.2.6 Métodos de Agrupamentos

O uso de métodos de agrupamentos tem como objetivo reconhecer objetos com um grau de similaridade suficiente para reuni-los em um mesmo conjunto, ou ainda, de realizar o reconhecimento de subconjuntos descontínuos em um ambiente que é

algumas vezes discreto, como no caso da taxonomia e, quase sempre descontínuo como na Biologia (LEGENDRE *et al*., 2011). Os métodos de agrupamentos não são um método estatístico típico, pois não testam nenhuma hipótese, mas auxilia a levantar características implícitas nos dados; através das quais o pesquisador deverá realizar a avaliação se o agrupamento obtido é de interesse ou não ecologicamente.

Tabela 3 – Usos da modelagem em nível de comunidade. Fonte: (FERRIER, GUISAN, 2006)

Tipos de uso	Descrição
Tipos de comunidades	Cada "tipo comunidade" é definido como um grupo de locais (*grid* de células) que se assemelham umas com as outras em termos de composição de espécies. O agrupamento é normalmente obtido através de alguma forma de classificação
Grupos de espécies	Cada "grupo de espécies" é definido com uma distribuição preditiva similar, novamente alcançado através da classificação numérica, mas neste caso os objetos classificados são espécies ao invés de locais de ocorrência
Eixos de variação composicional	Um conjunto de eixos contínuos (ou gradientes), representando dimensões de um espaço reduzido, resume o padrão de composição das espécies. Estes eixos são derivados mais comumente através de alguma forma de ordenação
Níveis de dissimilaridade de composição entre os pares de células	O nível previsto de dissimilaridade na composição da comunidade entre todos os possíveis pares de *grids* de células em uma região
Propriedades macroecológicas	A propriedade mais comumente modelada é a riqueza de espécies, de todo um grupo (por exemplo, todas as plantas vasculares) ou de um subgrupo funcional (por exemplo, anuais e árvores). Muitas outras propriedades macroecológicas (por exemplo, média, limites, tamanho e endemismo) podem ser potencialmente modeladas

De modo a comparar os métodos usados na Biologia e na Mineração de Dados, realizamos um levantamento da forma como os métodos de agrupamentos são utilizados

e quais desses são usados. Na Biologia, a análise de agrupamentos é aplicada seguindo os passos abaixo (VALENTIN, 2000):

a) Coleta de dados;
b) Escolha do modo de análise: modo Q (objetos) ou modo R (descritores);
c) Escolha do coeficiente de associação (similaridade, distância ou dependência);
d) Escolha do método de agrupamento, que depende de critérios baseados no menor grau de distorção e, sua capacidade de evidenciar melhor a estrutura dos dados (existência de grupos);
e) Elaboração e interpretação de dendograma.

Os métodos de agrupamentos podem ser classificados em seis categorias, da seguinte forma (LEGENDRE, LEGENDRE, 1998):

- **Sequenciais** – os objetos são reunidos um após o outro, respeitando uma determinada sequência de operações, ou simultâneos, que são executados em uma única sequência e são menos frequentes;
- **Aglomerativos** – os objetos são inicialmente isolados, depois são progressivamente reunidos em grupos sucessivos até formar um único grupo, ou **divisivos** – inicia-se a análise com um único grupo, o qual é dividido em subgrupos até chegar aos objetos individuais;
- **Monotéticos** – baseados em um único descritor por vez, ou politéticos, baseados em vários descritores;
- **Hierárquicos** – os elementos são organizados em uma série hierárquica ou não, que procuram maximizar a homogeneidade intragrupo, sem considerar a hierarquia entre grupos;

- **Probabilísticos** – recomendados para o agrupamento de espécies, eles são porém menos usados por conta da complexidade dos cálculos necessários. Esses algoritmos são utilizados, algumas vezes, para verificar a associação entre espécies.

Dos métodos listados, ainda segundo Valentim, os mais usados na Biologia são:

- **Método da ligação simples** (sequencial), que utiliza dendogramas e, que por reunir um objeto ao elemento mais próximo do grupo anteriormente formando, faz com que os objetos intermediários sejam rapidamente aglomerados a esses. Esse método deve ser evitado nos estudos ecológicos em que amostras intermediárias são geralmente numerosas;
- **Método por ligações completas**, conhecido também como o do "vizinho mais distante", é o oposto ao anterior. A união de dois grupos depende dos elementos mais distantes. É recomendado em Biologia quando se deseja descobrir fortes descontinuidades;
- **Método pela associação média** ou UPGMA (SNEATH, SOKAL, 1973), calcula-se a média aritmética da similaridade (ou distância) entre o objeto que se quer incluir em um determinado grupo e cada objeto desse grupo;
- **Método dos pesos proporcionais**, como é comum na Biologia, grupos de amostras de regiões diferentes e de tamanhos diferentes são usados, resultando em regiões mais amostradas que outras. Assim (SOKAL, MICHENER, 1958), definiram o método de aglomeração por pesos proporcionais – WPGMA (*Weighted Pair Group Method With Arithmetic Means*);
- **Método pela variância mínima**, também chamada de método de Ward (WARD, 1963, LATTIN et al., 2011), nele um grupo será reunido a outro se essa união

proporcionar o menor aumento da variância intragrupo. O método é altamente eficiente na formação de grupos.

Dos métodos listados, os métodos UPGMA e o de Ward são os mais indicados para aplicação na Biologia, mais recentemente, o método de clusterização por meio de particionamento K-Means (BOCARD *et al*., 2011). Este último já se encontra disponível para uso na Ecologia através do *package* Vegan do R (OKSANEN, 2011) e no programa *Spatial Analysis in Macroecology* - SAM (RANGEL *et al*., 2010).

2.2.7 Coocorrências de Espécies

Uma questão muito importante na Biologia é se existem regras gerais para explicar a composição para determinar a estrutura de comunidades. Nesse sentido, diversas pesquisas são realizadas e uma delas é de especial interesse neste trabalho, a associação entre espécies ou associação interespecífica. As comunidades biológicas naturais apresentam, como citado anteriormente, um grande número de espécies raras com baixa abundância e poucas espécies dominantes com alta abundância. Para o estudo de coocorrências de espécies, a presença de um grande número de valores nulos é prejudicial para o estabelecimento de associações biológicas baseadas nos cálculos de coeficiente de dependência paramétricos. Entretanto, podemos definir as associações biológicas por meio da coocorrência de espécies, e não de correlações entre abundâncias (VALENTIN, 2000). Para cada par de espécies que se deseja testar a associação constrói-se uma Tabela de Contingência, como a apresentada na Figura 4:

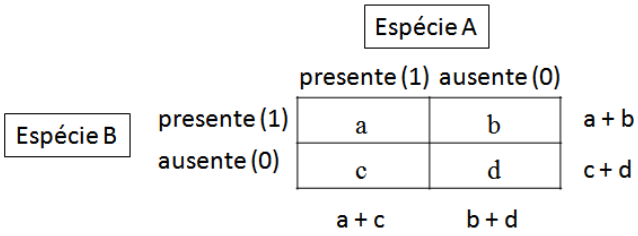


Figura 4. Matriz de contingências. Adaptado de VALENTIN (2000)

Convenções adotadas:

- **a** indica o número de amostras possuindo as duas espécies;
- **b** indica o número de amostras onde A está presente e B ausente;
- **c** indica o número de amostras onde A está ausente e B presente;
- **d** indica o número de amostras onde não ocorrem as espécies;
- A soma n = a + b + c + d

Dados quantitativos das espécies, em geral, exigem medidas de distância assimétricas. Nessa categoria são usados coeficientes como Bray-Curtis (LEGENDRE, LEGENDRE, 1998) e Chi-Quadrado. O teste de Chi-Quadrado pode ser usado para verificar se há dependência entre as distribuições de duas espécies. No caso de uma probabilidade de 5% e 1 grau de liberdade o Chi-Quadrado teórico da tabela é 3.84. Se o cálculo for maior que este valor, rejeita-se a hipótese nula informando então que há associação entre as espécies envolvidas.

Seu cálculo pode ser realizado com $n(ad - bc)^2 / (a + b)(c + d)(a + c)(b + d)$

Outro teste, com esse objetivo é o de Bray-Curtis, que é uma estatística utilizada para quantificar a dissimilaridade de composição entre dois locais diferentes, com base nas contagens em cada local.

Seu índice é definido por $BC_{ij} = 2C_{ij} / (S_i + S_j)$, onde, C_{ij}, é a soma do menor valor apenas para as espécies em comum entre dois *plots* (parcelas). Os parâmetros S_i e S_j são o número total de amostras contadas em ambos os sítios. O índice reduz-se a C = 2C / 2 onde abundâncias em cada local são expressos como uma percentagem. A dissimilaridade de Bray-Curtis varia entre 0 e 1, onde 0 significa que as duas parcelas têm a mesma composição, ou seja, compartilham todas as espécies, e 1 significa que as duas parcelas não compartilham qualquer espécie.

Outro índice bastante utilizado como medida de associação é o **índice de Jaccard** (JACCARD, 1901), que rejeita a hipótese de dupla ausência. Porém não trabalharemos com ele por não dispor de dados de ausência. Seu cálculo é dado pela fórmula: IJ = a / (a + b + c).

Mais recentemente, outra forma usada para a determinação de coocorrências entre espécies é a de **Modelos Nulos** (GOTELLI, 2000). Esse método baseia-se na definição de uma hipótese nula e observa-se que os padrões revelados não são atribuídos a uma determinada explicação causal indicada previamente. Por meio dos Modelos Nulos podem ser obtidos padrões a partir dos dados gerados nos estudos ecológicos. Simulações complexas podem ser realizadas com os meios computacionais, pois em uma determinada parte dos dados é realizada a randomização para a produção de padrões, que podem ser esperados na ausência de um determinado mecanismo ecológico. Entre as aplicações dos Modelos Nulos podemos utilizá-lo para argumentar se as diferenças obtidas nas análises em relação à diversidade são simplesmente o resultado da amostragem. Considerando ainda que os resultados obtidos dependem claramente de como a comunidade é estruturada (MAGURRAN, 2011). Embora os Modelos Nulos muitas vezes envolvam comparações de ocorrência das espécies entre vários locais geograficamente distintos, a abordagem geral de construção de modelos

nulos é a da randomização de um padrão observado, sendo também usada para a construção de modelos em comunidades simples (FULLER, ENQUIST, 2011). Modelos Nulos podem ser utilizados com os *softwares* Ecosim (ENTSMINGER, 2012) e Vegan no R (R DEVELOPMENT CORE TEAM, 2010) realizando simulações com modelos baseados em dados randômicos.

2.2.8 Análise de Padrões de Pontos Espaciais

Nos últimos anos a Análise de Padrões de Pontos Espaciais (PPA) (WANG *et al*., 2010, WIEGAND, MOLONEY, 2004, PERRY *et al.*, 2006, PROTÁZIO, 2007) vem obtendo grande destaque entre os métodos usados na avaliação da distribuição de pontos de ocorrência de espécies. Seu objetivo é o de verificar se existe uma tendência nos pontos observados conforme padrões que podem ser regulares, agregados ou aleatórios (SCHNITZER *et al*., 2012). Pela avaliação em diferentes distâncias, a função de correlação de pares permite uma precisa identificação da escala onde interações significantes interações ponto-a-ponto ocorrem (LUO *et al*., 2012, WIEGAND *et al.*, 2007). Ela pode ser aplicada da forma univariada ou bivariada pela função K de Ripley (RIPLEY, 1987), apresentando como principal vantagem o fato de identificar o padrão espacial em diversas escalas de distâncias de forma simultânea, avaliando a dependência espacial entre grupos de árvores. A função bivariada permite avaliar a relação espacial entre dois grupos de árvores, como por exemplo, duas espécies competidoras. Entre os programas que disponibilizam a função estão o Programita.

2.2.9 Análise de *Softwares* Existentes para o Estudo de Coocorrências entre Espécies Vegetais

Foram observados que poucos *softwares* foram desenvolvidos para o estudo de coocorrências em comunidades vegetais. Na Tabela 4, alguns dos principais *softwares*

foram analisados quanto a sua finalidade, recursos, formato de dados necessário para a carga e possíveis limitações de uso.

De uma forma geral, as opções existentes de *softwares* com o objetivo de analisar as coocorrências de espécies vegetais apresentam limitações relacionadas à integração com bancos de dados; fato esse verificado pelo formato de dados textual necessário a entrada de dados para o processamento, o mesmo ocorrendo na saída dos resultados. Desta forma, há uma necessidade maior de tempo por parte do pesquisador para a realização dos estudos na preparação dos dados.

Outro ponto observado foi que a maioria dos testes usando PPA é realizada com a pré-seleção das espécies a serem analisadas (WANG *et al*., 2010, SCHNITZER *et al*., 2012, LUO *et al*., 2012, DETTO, MULLER-LANDAU, 2013). Isto tem como justificativa o fato que a execução dos testes com todos os pares de espécies (WIEGAND *et al*., 2012) não é muito prática, criando uma dificuldade adicional quando a pesquisa analisa múltiplas espécies.

Tabela 4 – Comparativo de comparativo entre *softwares* utilizados na análise de coocorrências de espécies vegetais

Programa	Análises Realizadas	Formato de entrada	Limites
Twinspan (HILL, 1979)	***Tw****o-Way* ***In****dicator* ***Sp****ecies* ***An****alysis* é um programa projetado primariamente para dados coletados em parcelas. Tem por objetivo verificar as espécies e suas preferências, positivas e negativas, a partir de ocorrências	Escrito em Fortran. Arquivo no formato texto	Máximo de 25.000 parcelas e 10.000 espécies
CoOccurrence (ULRICH, 2006)	No programa estão implementadas as seguintes medidas*: c-score*, *checkboard score*, *species combinations score* (COMBO), correlação de Spearman, teste de variância de Schluter, índice de Morisita e teste de Mantel	O programa trabalha com matriz de presença com linhas representando espécies e os sites (parcelas) as colunas. O arquivo de dados tem de ser criado no formato ASCII	Máximo de 150 parcelas a serem analisadas
EcoSim (ENTSMINGER, 2012)	Programa executa análises baseadas em Modelos Nulos em bases de dados de comunidades vegetais, usando métodos como: ANOVA, chi-quadrado e *checkboard*	Usa o mesmo formato de dados do CoOccurrence, arquivos delimitados por espaços, textos separados por vírgulas e .csv	1400 espécies
Vegan (DIXON, 2003)	Package do R que possui, além de diversos índices ecológicos, como por exemplo, o índice *checkboard* e Modelos Nulos	Arquivo texto	Não observado
Programita (WIEGAND MOLONEY, 2004)	Permite a execução de análises univariadas e bivariadas para a identificação de padrões na distribuição de pontos com a aplicação das funções K e L de Ripley e da estatística de O-ring	Desenvolvido em Delphi 4 tendo como formatos de entrada arquivos .dat e .asc	Não observado
Turnover (ULRICH, 2012)	O programa analisa coocorrências negativas e positivas de espécies baseada no método *Nearest Neighbour*, índice de Morisita, *c-score* e outros	Escrito em Fortran, aceita arquivos delimitados por espaços, com os *plots* sendo informados nas colunas e as espécies nas linhas; igual ao Ecosim	Máximo de 40 *plots* e de 80 espécies

Capítulo 3 – Aplicação da Mineração de Dados em Bases de Dados da Biodiversidade

3.1 Mineração de Dados

O crescimento exponencial dos dados nos últimos anos vem provocando um aumento do número de discussões em torno da necessidade de pesquisas de novos métodos de acesso, análise e gerenciamento de dados biológicos (HOWE, 2008). Porém, a demanda para o desenvolvimento de pesquisas objetivando a descoberta de conhecimento foi percebida há muito tempo. Encontramos, por exemplo, em FAYYAD *et al*. (1996) a seguinte afirmação: *"há uma necessidade urgente por uma nova geração de teorias e ferramentas computacionais para auxiliar humanos na extração de informação útil (conhecimento) em contrapartida ao acelerado crescimento do volume de dados digital"*. O interesse pela descoberta de conhecimento em banco de dados teve início, historicamente registrado, em 1989 com o *Workshop on Knowledge Discovery in Databases* (PIATETSKY-SHAPIRO *et al*., 1991) e tem evoluído muito nas últimas décadas. Assim, como um ramo da inteligência artificial, a extração de dados e descoberta de conhecimento objetiva descobrir automaticamente regras e modelos estatísticos a partir dos dados (SILBERSCHATZ *et al*., 1999).

O processo de extração de conhecimento de bancos de dados (KDD - *Knowledge Discovery in Databases*) envolve diversas etapas para sua total aplicação. Na Figura 5 são apresentadas as fases realizadas para a extração do conhecimento, onde pode-se observar que somente depois da fase de pré-processamento é que as técnicas de mineração de dados são usadas para extrair as regras e padrões. Ela é uma das principais fases deste processo de geração de conhecimento e tem por objetivo realizar a

descoberta automática de informações úteis em grandes depósitos de dados (ELMASRI, NAVATHE, 2011).

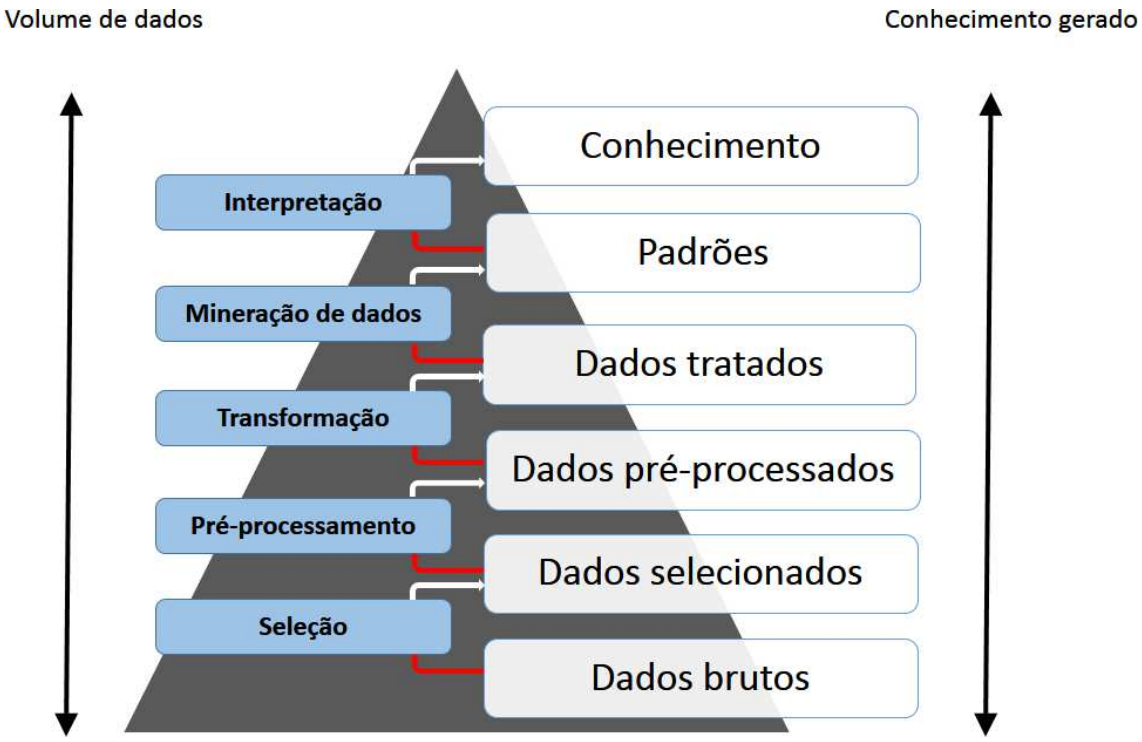


Figura 5. Processo de geração de conhecimento – exibindo a proporção do conhecimento gerado em relação aos dados originais. Adaptada de FAYYAD *et al.* (1996)

O processo de KDD vêm sendo utilizado, com sucesso, em diversas áreas tradicionais como: o marketing, medicina, economia, engenharia e administração de empresas (CARVALHO, 2005) e mais recentemente nas redes sociais (NGAI *et al.*, 2009, LIAO *et al.*, 2012); geografia (SU *et al*., 2004, MILLS *et al*., 2011), genética (KOONCE, TSAI, 2000). Nas Ciências da Terra as possibilidades de uso da mineração de dados já eram indicadas em (HAN *et al.*, 2002), consistindo de dois componentes principais: a modelagem de dados ecológicos e o desenvolvimento de algoritmos eficientes para a descoberta de padrões espaço-temporal. Isso ocorrendo em razão da necessidade dos centros de pesquisas por esses tipos de métodos para agilizar o processo de reconhecimento de padrões (RAYMOND *et al.*, 2005). Na Ecologia

também foram observadas algumas pesquisas com o uso das potencialidades das técnicas de mineração de dados, como podemos comprovar em trabalhos com o uso de algoritmos de classificação (HOCHACHKA *et al.*, 2007, PINO-MEJÍAS *et al*., 2010, LORENA *et al*., 2011, WISDOM, 2011, CUTLER *et al*., 2007); análise de agrupamentos (BRANDAO *et al.*, 2009, KENT, CARMEL, 2011, KUMAR *et al*., 2011); mas muito pouco se comparado com outras áreas (INMAN-NARAHARI *et al*., 2010). A seguir são analisadas as duas categorias de algoritmos da mineração de dados selecionadas para uso na metodologia proposta.

3.2 Análise de Associação

3.2.1 Conceituação

Uma dos métodos não-supervisionados mais populares da mineração de dados é o método destinado a encontrar conjuntos de itens frequentes a partir de transações registradas em banco de dados, com a extração de regras de associação entre os itens presentes nas transações, sem levar em consideração as implicações de causalidade (AGRAWAL *et al.*, 1993, AGRAWAL, 1994, TAN *et al*., 2009, HAN, 2011, WU *et al*., 2007). A análise de associação visa apresentar regras que muitas vezes não são claras, como por exemplo, o famoso caso de compras de fraldas e cervejas revelando o padrão de pais que compravam fraldas e também cervejas nas sextas-feiras à noite para assistir a jogos pela TV. Assim, uma *transação* é o conjunto de itens encontrados em operações tais como os produtos comprados por um determinado cliente, tal como a análise de cestas de compras (SILVERSTEIN *et al.*, 1997).

Na análise de associação, *itens frequentes* são conjuntos de itens com frequência maior ou igual a um valor mínimo informado ou o valor de *suporte*. A manipulação desses itens pode ser uma tarefa não muito simples, em razão da complexidade

ocasionada pela explosão combinatória que pode ser originada com base na quantidade de itens envolvidos na análise. De uma forma geral, o conjunto de itens frequentes pode ser encontrado por $2^k - 1$, excluindo o elemento nulo (TAN *et al.*, 2009). Depois da obtenção dos itens frequentes, o próximo passo na análise é a aplicação do algoritmo para a identificação das regras de associação. O valor de confiança juntamente com o valor de suporte serão usados para a seleção do conjunto de regras interessantes obtidas do conjunto de transações. O algoritmo então deve encontrar todas as regras que satisfaçam a condição na qual o valor do suporte deve ser maior ou igual ao valor mínimo de suporte e a confiança deve ser maior ou igual ao valor mínimo de confiança. O número total de regras possíveis que podem ser extraídas para um conjunto de itens contendo d itens é dado por: $3^d - 2^{d+1} + 1$.

Nesta pesquisa, as regras de associação foram formalizadas da seguinte forma:

a) Um conjunto de itens (*itemset*) é um subconjunto de espécimes do conjunto total de espécies, sendo $\boldsymbol{c\_sp} = \{sp\_1, sp\_2, sp\_3, \ldots, sp\_n\}$;

b) Uma transação t_{treeid}, onde *treeid* é o identificador do espécime, contém *itemsets* que ocorrem próximos de um espécime qualquer da espécie *x* e, cada transação, representada como $\boldsymbol{t_{treeid}\_sp_x} = \{sp\_1_{treeid}, sp\_2_{treeid}, sp\_3_{treeid}, \ldots, sp\_N_{treeid}\}$, contém outros espécimes próximos do espécime analisado em uma distância, em metros, previamente especificada. O conjunto de transações de uma espécie é representado como $\boldsymbol{t\_sp} = \{t1, t2, t3, \ldots, t\_n\}$;

c) Uma regra de associação é uma afirmação lógica entre o antecedente (LHS - *left hand side*) e o consequente (RHS – *right hand side*) e, pode ser exemplificada como implicação de $sp\_x \rightarrow sp\_y$, podendo ser compreendida como um padrão onde *sp_x* e

sp_y aparecem juntas nas ocorrências; desde que *sp_x* e *sp_y* estejam contidos em *c_sp* e; *sp_x* e *sp_y* sejam conjuntos disjuntos, ou seja, *sp_x* ∩ *sp_y* = ø.

3.2.2 Principais Algoritmos da Análise de Associação

Muitos algoritmos foram desenvolvidos para a análise de itens frequentes e geração de regras de associação sobre os itens. Em termos de classificação, podem ser divididos em três categorias:

- **Algoritmos Baseados no Princípio Apriori:**

O algoritmo Apriori (AGRAWAL, 1994) explora a propriedade anti-monotônica da medida do suporte, na qual o suporte para um conjunto de itens nunca excede o suporte de seus subconjuntos. Essa estratégia de diminuir o espaço de pesquisa exponencial baseado na medida de suporte é conhecida como poda baseada em suporte (TAN *et al*., 2009). O algoritmo segue o princípio "se um conjunto de itens é frequente, então todos os seus subconjuntos também devem ser frequentes". A propriedade é formalizada assim:

$$\forall X, Y : (X \subseteq Y) \Rightarrow \sup(X) \geq \sup(Y)$$

Variações envolvendo o uso de árvore hash e redução do tamanho da transação podem ser utilizadas para tornar o processo ainda mais eficiente. Nas variações envolvendo o particionamento dos dados, ocorre a mineração de cada partição e depois os resultados são combinados. O objetivo dessas variações é o de reduzir o número de varreduras necessárias no banco de dados na identificação dos itens frequentes como realizado no Apriori.

- **Algoritmos Baseados no Padrão de Crescimento Frequente:**

Finding Frequent Itemsets Without Candidate Generation - FP-Growth (HAN *et al*., 2000) é um método da mineração de conjuntos de itens frequentes, realizado sem a geração de itens candidatos. O algoritmo constrói uma estrutura de dados altamente compacta (FP-tree) na base de dados transacional original. Ao invés de empregar a estratégia de gerar e testar, dos métodos semelhantes ao Apriori, para a geração de itens candidatos; o algoritmo objetiva identificar o padrão frequente, evitando a custosa geração de candidatos e, resultando assim em uma maior eficiência no desempenho.

- **Algoritmos que usam o Formato de Dados Verticais:**

Outro método de mineração de conjuntos de itens frequentes é realizado utilizando o formato de dados verticais. O *Equivalence CLASS Transformation* - ECLAT (ZAKI, 1998) é um método que transforma um determinado conjunto de dados de transações de dados, do formato horizontal para um conjunto de dados verticais. O algoritmo analisa os dados transformados com base na propriedade Apriori e técnicas de optimização adicionais.

Após a avaliação dos algoritmos de associação, foi realizada a escolha pelo algoritmo Apriori. Esta escolha é justificada pelo fato do foco do trabalho estar orientado na avaliação de uma nova metodologia, sem considerar a questão de desempenho como um requisito não-funcional da proposta.

3.3 Análise de Agrupamentos

Os algoritmos usados na análise de agrupamentos, ou simplesmente clusterização, tem como finalidade particionar um conjunto de registros em grupos tais que os registros dentro de um grupo sejam similares entre si e, os registros pertencentes

a dois grupos diferentes apresentem características diferentes (RAMAKRISHMAN, GEHRKE, 2008). Existe uma variedade de algoritmos de agrupamentos que podem ser classificados em grupos (ESTER *et al*., 2001, BERKHIN, 2002). Na mineração de dados espaciais, a análise de agrupamentos é usada de forma acentuada objetivando agrupar objetos espaciais similares em classes, e é um importante componente da mineração de dados espaciais (HAN *et al*., 2005, ESTER *et al.*, 2001). Como uma função da mineração de dados, a mineração espacial pode ser usada como: a) uma ferramenta isolada para obter indícios de uma distribuição de dados; b) um mecanismo para observar as características de cada agrupamento e; c) uma estratégia para focar em um particular conjunto de agrupamentos para futuras análises. Também pode servir como uma etapa no pré ou pós-processamento de outros algoritmos, tal como os relacionados com a classificação e que serão utilizados na detecção de agrupamentos.

Diversos são os algoritmos utilizados na mineração de dados espaciais. Esses podem ser classificados em métodos de: particionamento, hierárquicos, baseados em densidade e baseados em grade. A seguir são listados algoritmos de agrupamentos utilizados em aplicações que envolvem dados espaciais e que foram estudados para a avaliação de seu uso na presente pesquisa. Foram encontradas diversas classificações para os algoritmos da análise de agrupamentos, dentre os quais optamos por seguir a proposta apresentada por HAN (2005):

3.3.1 Métodos de Particionamento

Dado um grupo de *n* objetos em uma dimensão espacial *d* e um parâmetro de entrada *k*, os algoritmos de particionamento organizam os objetos em *k clusters* de modo que o total de desvio de cada objeto a partir do centro do *cluster* ou a partir de a

distribuição do *cluster* ser minimizada. A seguir são listados os principais algoritmos do método de particionamento e algumas de suas características:

a) **K-means** (MACQUEEN, 1967) – O algoritmo tende a descobrir *clusters* de formato esférico e de tamanho similar. Só pode ser aplicado quando é possível definir a média dos grupos, o que não ocorre com variáveis categóricas. Sua aplicação é muito sensível a presença de ruídos ou valores extremos nos dados, sendo que estes influenciam fortemente o cálculo da média dos centros dos agrupamentos. Funciona bem quando os grupos são compactos e separados uns dos outros. O método, por ser escalonável, é eficiente e rápido no processamento mesmo em grandes quantidades de dados. É menos sensível a valores extremos. Segundo LEGENDRE *et al*. (2011), o algoritmo pode ser usado na Ecologia para encontrar regiões de ocorrências de espécies que apresentam um alta densidade;

b) ***Expectation Maximization* - EM** (LAURITZEN, 1995) – O algoritmo de maximização de expectativas utiliza a distribuição probabilística para representar cada *cluster*. Usa a probabilidade Gaussiana que se baseia na teoria de estimativa de densidade. O algoritmo faz uma previsão inicial para os parâmetros e, depois melhora-os interativamente. Gera *clusters* com tamanho similar e forma esférica, facilmente percebidos pelos olhos humanos;

c) ***Partitioning Around Medoids* - PAM** (KAUFMAN, ROUSSEEUW, 1990) – PAM Foi o primeiro algoritmo a utilizar o método *k-medoid* de particionamento (HAN *et al*., 2011). Porém, como usa o objeto mais central (*medoid*) ao invés da média, é menos sensível a erros. Procura pelo melhor *k-medoid* entre os dados selecionados. Trabalha bem para pequenos grupos de dados. Tende a descobrir agrupamentos de formato esférico e de tamanho similar. Proibitivo para grandes bases de dados;

d) ***Clustering Large Applications*** **- Clara** (KAUFMAN, ROUSSEEUW, 1990) – Definido para uso com grande quantidade de dados. Usa amostras do conjunto de dados para procurar os melhores *k-medoids*. Pode não encontrar o melhor agrupamento para todo o conjunto de dados. Um bom agrupamento baseado em amostragem não representa necessariamente um bom agrupamento para todo o conjunto de dados;

e) ***Clustering Large Applications based upon RANdomized - Search Clarans*** (NG, HAN, 1994) – Para melhorar a qualidade e a escalabilidade o algoritmo escolhe as amostras do conjunto de dados dinamicamente durante a execução. Se a melhor solução não é encontrada depois de um certo número de tentativas, o local é assumido como ótimo;

3.3.2 Métodos Hierárquicos

Esta categoria de algoritmos cria uma decomposição hierárquica de um conjunto de dados formando um dendograma, gerando uma árvore que divide o banco de dados recursivamente em pequenos subconjuntos. Seguem duas abordagens: a *bottom-up* (aglomerativa) ou a *top-down* (divisiva). Os principais algoritmos são listados abaixo:

a***) AGglomerative NESting*** **- Agnes** (KAUFMAN, ROUSSEEUW, 1990) e ***DIvisive ANAlysis*** **- Diana** (KAUFMAN, ROUSSEEUW, 1990) – No Agnes os *clusters* são aglomerados pelo algoritmo. No Diana todos os objetos são usados para formar um *cluster* inicial. O *cluster* é dividido seguindo algum princípio, tal como a distância Euclidiana entre os objetos vizinhos. São algoritmos antigos e tendem a produzir *clusters* errados, e com isso, conduzir a uma propagação desses erros. A razão para a propagação está no formato, pois depois de formados os agrupamentos, não permitem a troca dos objetos entre os grupos;

b) ***Balanced Iterative Reducing and Clustering Using Hierarchies* – Birch** (ZHANG *et al*., 1996) **–** Não trabalha bem com *clusters* que não são esféricos, pois trabalha com a noção de raio ou diâmetro. O algoritmo faz a compressão dos objetos em muitos agrupamentos pequenos e depois executa o agrupamento desses conjuntos. Esta abordagem permite uma execução mais rápida;

c) **CobWeb** (FISHER, 1987) **-** É um algoritmo da análise de agrupamentos espacial para dados categóricos. O CobWeb tem duas importantes qualidades. A primeira utiliza o aprendizado incremental, ao invés das abordagens aglomerativas e divisivas. Constrói um dendograma dinamicamente, processando um ponto de dados por vez. O CobWeb pertence ao modelo conceitual baseado em aprendizado. Isto significa que cada *cluster* é considerado como um modelo que pode ser intrinsecamente descrito, melhor do que uma coleção de pontos;

d) **Cure** (GUHA *et al*., 1998) – Se ajusta bem para *clusters* com formato não-esférico. É mais robusto em relação a *outliers*. Trabalha bem com grandes bases de dados sem sacrificar a qualidade dos *clusters*;

e) ***A Hierarchical Clustering Algorithm Using Dynamic Modeling* - Chamaleon** (KARYPIS *et al*., 1999) – Mais eficiente do que o Cure para a descoberta de *clusters* de formatos arbitrários. Exige um alto custo de processamento com muitas dimensões.

3.3.3 Métodos Baseados em Densidade

Os métodos de particionamento baseiam-se na distância entre os objetos. Tais métodos tendem a encontrar agrupamentos com formato esférico e encontram dificuldades para encontrar agrupamentos com outros formatos. Assim, os métodos baseados em densidade retornam agrupamentos com regiões densas separadas de outras

com baixa densidade. Os algoritmos desta categoria podem ser usados para a descoberta de *outliers*. Por utilizarem índices, os métodos baseados em densidade podem perder eficiência se o número de variáveis for excessivo. São eles:

a) ***Density-Based Spatial Clustering of Applications with Noise*** **- DBScan** (ESTER *et al*., 1996) – O algoritmo descobre regiões com alta densidade em *clusters* e descobre regiões de formatos arbitrários. Durante o processo, um determinado objeto que pertence a um agrupamento pode ser novamente encontrado e, desta forma, ocorrerá uma nova junção destes agrupamentos. O processo termina quando nenhum ponto pode ser adicionado em outro *cluster*. Necessita de dois parâmetros – raio e número mínimo de pontos para sua execução;

b) ***Ordering Points to Identify the Clustering Structure*** **- Optics** (ANKERST *et al.*, 1999) – Pode descobrir *clusters* de formato arbitrário baseado em densidade uniforme, cores, etc. Tamanhos e formas arbitrárias. Trabalha com argumentos para ordenar os *clusters* segundo uma análise interativa e automática;

c) ***DENsity-based CLUstEring*** **- Denclue** (BREESE *et al*., 1998) – Modela a densidade global de um ponto analiticamente como a soma das funções de influência dos pontos de dados ao redor de um determinado ponto. Para computar a soma das influências eficientemente, uma estrutura de grade é utilizada.

3.3.4 Métodos Baseados em Grid

Utilizam uma estrutura baseada em grade, quantificando o espaço em um finito número de células onde as operações podem ser realizadas. A maior vantagem está no desempenho que é tipicamente independente do número de objetos, dependendo apenas do número de células em cada dimensão. São eles:

a) ***STatistical INformation Grid* - Sting** (WANG *et al*., 1997) – Explora informações estatísticas armazenadas nas células. É uma técnica de clusterização na qual a área especial é dividida em células retangulares;

b) ***CLustering InQUEst* - Clique** (AGRAWAL *et al.*, 1998) – Utiliza representação em grade e utiliza a abordagem baseada em densidade para agrupar em um espaço de dados com muitas dimensões;

c) **Wave*Cluster*** (SHEIKHOLESLAMI *et al*., 1998) – Utiliza um método baseado em ondas. No seu processamento principal, o algoritmo sumariza os dados colocando-os em uma estrutura de grade multidimensional.

3.3.5 Métodos Baseados em Redes Neurais

Foi avaliado também o uso do algoritmo do tipo Mapa Auto Organizável – SOM (KOHONEN, 1990). O SOM consiste em um método eficiente na análise exploratória dos dados espaciais que permite levantar hipóteses e centralizar o foco em um conjunto reduzido de variáveis. É um método computacional de aprendizagem não-supervisionada que permite identificar regiões homogêneas a partir de parâmetros associados ao ambiente, tais como variáveis climáticas e geomorfológicas (XIMENES, 2008).

A Tabela 5 apresenta uma comparação das principais características dos algoritmos de agrupamentos que foram observados para a escolha dos que poderiam ser aplicados na trabalho. Na tabela utiliza-se a seguinte legenda: *k* é o número de *clusters*, *n* é o número de elementos.

Tabela 5 – Comparativo de algoritmos usados na análise de agrupamentos

	Algoritmo	Complexidade	Indicações de uso	Desvantagem / tipos de clusters
Particionamento	EM	Se k e d (dimensão) são fixas, pode ser resolvido em O (n^{dk+1} log n)	Dada uma estimativa para os parâmetros, calcula probabilidades para avaliar os novos parâmetros. Não exige a definição inicial do número de *clusters* ou da distância entre os pontos que formam os *clusters*	Encontra *clusters* que são de formato esférico e similar em tamanho. Pode ser lento, não é prático para modelos com grande quantidade de componentes
	K-means	O (nkt) t – nr iterações	Em aplicações onde se deseja que a distância para o centro do *cluster* seja curta. Funciona bem quando os grupos são compactos e separados. É menos sensível a valores extremos. É eficiente em grandes bases	Não pode ser aplicado em variáveis categóricas. Precisa estimar o parâmetro K. Sensível a ruídos ou valores extremos, com forte influência destes nos cálculos. *Clusters* esféricos e similares em tamanho
	Clara	O ($ks^2 + k(n – k)$)	Grande quantidade de dados. Utiliza amostras do conjunto de dados para procurar os melhores k medóides dentre essas amostras	Pode não encontrar o melhor agrupamento para todo o conjunto de dados. Assim, um bom agrupamento baseado em amostragem não significa a certeza de um bom agrupamento para todos o conjunto de dados
	PAM	O ($k(n – k)^2$)	Procura o melhor k medóide entre os dados usados. Em aplicações onde se deseja que a distância para o centro do *cluster* seja curta. Apresenta menor sensibilidade a erros e *outliers*	Proibitivo para grandes bases de dados. Muito tempo de processamento. *Clusters* de formato esférico e similar em tamanho
	Clarans	O (n^2)	Escolhe as amostras do conjunto de dados dinamicamente durante a execução	A qualidade dos resultados não pode ser garantida quando N é tem um alto valor. Assume que todos os objetos estão na memória principal
	Agnes	O (n^2 log n)	Usa a abordagem aglomerativa	Tende a produzir *clusters* errôneos
Hierárquicos	Diana	O (n^2)	Usa a abordagem divisiva	Tende a produzir *clusters* errôneos
	Cure	O (n^2)	Mais robusto a *outliers*. Trabalha bem com grandes bases de dados sem perder em qualidade dos *clusters*	Ajusta-se bem para *clusters* com formato não esféricos

	Chameleon	O (n^2)	Mais eficiente do que o Cure para a descoberta de *cluster* de formatos arbitrários	Custo alto de processamento quando usa muitas dimensões
	Birch	O (n)	Usa a noção de raio ou diâmetro	Não trabalha bem se os *clusters* não são esféricos
	Optics	O (log n)	Pode descobrir *clusters* de formato arbitrário	Não é muito sensível aos parâmetros. Baixo desempenho. Tamanhos e formas arbitrárias
Densidade	Denclue	O (n log n)	Para computar a soma das influências eficientemente, uma estrutura de grade é utilizada	Requer uma escolha cuidadosa dos parâmetros de agrupamento
	Dbscan	O (n log n), pior caso, O (n^2)	Descobre regiões com alta densidade e descobre regiões de formatos arbitrários. Não é muito afetado por ruídos. Identifica grupos de diferentes tamanhos	É sensitivo aos parâmetros de entrada. Baixo desempenho. Tem problemas quando os grupos têm densidade variada e muitas dimensões
	Sting	O (k) K é nr de grids	Trabalha bem com dados espaciais	Dependente dos parâmetros de entrada e pode perder grupos naturais. Como as células são retangulares, não captura com precisão a densidade das áreas circulares. O aumento de dimensionalidade provoca o aumento exponencial do número de potenciais células da grade
Grid	WaveCluster	O (n)	Remove automaticamente *outliers*. Pode manipular dados com mais de 20 variáveis	Descobre *clusters* de formato arbitrário
	Clique	O (C^k + nk) k – mais alta dimensão	Trabalha com dados numéricos	É de difícil aplicação por que o tamanho do *grid* e o valor limite de densidade de corte, são usados para todas as combinações das dimensões no *data set*

3.4 Análise dos Algoritmos para Uso na Metodologia Proposta

Dadas as características do problema apresentado, e objetivando analisar as coocorrências de espécies e a necessidade de entendimento dos mecanismos de coexistência, competição e distribuição de espécies (WIEGAND *et al.*, 2012), foram selecionados os seguinte algoritmos de mineração de dados para uso na metodologia proposta:

3.4.1 Seleção do Algoritmo para a Análise de Associação

Considerando a natureza da aplicação e a possibilidade de vários espécimes ocorrerem próximos uns de outros, a transação, que é uma unidade de aplicação da análise de associação, pode conter um elevado número de itens (espécimes em nosso estudo), resultando na dificuldade de determinar os itens frequentes e, consequentemente, a geração de regras. Portanto, selecionamos o algoritmo Apriori (AGRAWAL, 1994), que considera que, para um conjunto de itens frequentes, todos os seus subconjuntos também devem ser frequentes, com a mesma abordagem a ser utilizada para itens infrequentes. Esta abordagem permite que o algoritmo diminua os números de itens candidatos na geração de regras, acelerando o processamento. Além disso, o Apriori foi selecionado pois os testes realizados na tese demonstraram sua eficiência tanto em termos de resultados quanto em termos de desempenho para a base de dados usada na pesquisa.

Em relação aos outros algoritmos aplicados no levantamento de regras de associação, esses apresentaram como vantagem principal em relação ao Apriori, somente o desempenho. Porém como o foco da pesquisa foi o desenvolvimento de uma metodologia da mineração de dados, a questão de desempenho recebeu um peso menor para a seleção do algoritmo.

3.4.2 Seleção do Algoritmo para a Análise de Agrupamentos

No processo de seleção do algoritmo de agrupamento para uma determinada aplicação, muitos fatores têm de ser considerados. Nesses fatores são normalmente incluídos: o objetivo da aplicação, a análise da qualidade dos dados, os requisitos de desempenho em relação ao tempo de execução do algoritmo, as características dos dados quanto aos seus tipos e o número de dimensões. Foi observado que os registros (espécimes) coletados são, normalmente, encontrados em comunidades onde a distância para o centro do *cluster* é mínima; caracterizando a ocorrência agregada.

Assim, após o estudo dos algoritmos de agrupamentos espaciais e considerando a natureza dos dados da aplicação, identificamos que o algoritmo EM (*Expectation Maximization*) é o mais adequado para a implementação do componente de agrupamentos na metodologia proposta pois, além da qualidade dos resultados gerados, o algoritmo apresentou ainda a facilidade de não exigir a inicialização do número de *clusters* esperados nos resultados ou da distância entre os pontos, como no Kmeans, por exemplo. Ainda como justificativa, podemos relatar os resultados obtidos na pesquisa realizada com a utilização do EM no desenvolvimento de um artigo (BRANDAO *et al*., 2009). Nesse trabalho foi realizada a análise de uma base de dados de bromélias, identificando padrões altitudinais em diferentes escalas espaciais. A partir dos resultados encontrados foi sugerido o uso para fins de conservação de espécies ameaçadas.

Capítulo 4 – Metodologia de Aplicação da Mineração de Dados na Análise de Agrupamentos de Espécies Vegetais

Considerando que o objetivo da metodologia é a extração de padrões de coocorrências de espécies, no desenvolvimento do método foram definidos os componentes e a sequência correta de aplicação desses componentes para alcançar esse objetivo. Foi identificada a necessidade de aplicação da análise de associação e da análise de agrupamentos espaciais baseada em densidade para uso, assim como a sequência sendo realizada inicialmente pela aplicação da análise de associação e subsequentemente pela análise de agrupamentos.

4.1 Fundamentação Teórica para a Seleção dos Componentes da Metodologia

Os componentes e a sequência para uso na metodologia foram definidos após o estudo da fundamentação teórica geral apresentada no Capítulo 2 sobre comunidades de espécies e ainda pelo relacionamento possível entre esses conceitos e os tipos de análises realizadas pela categorias de algoritmos selecionados da mineração de dados. A seguir são justificadas essas escolhas:

4.1.1 Análise de Associação

A estrutura fitossociológica das comunidades vegetais, seu padrão de ocorrência ou coocorrências de espécies e a sua variação no espaço é um campo de pesquisa importante na Biologia. Segundo Capelo *"no que diz respeito a cada um destes aspectos, em termos de vegetação, os botânicos têm tentado induzir proposições empíricas ao nível da comunidade por análise dos dados de campo e, a posterior, de forma teórica, deduzindo estes padrões a partir dos processos e relações ao nível das populações das espécies"* (CAPELO, 2003). Observou-se ainda em relação às

comunidades que existe certa probabilidade de que dadas espécies ocorram juntas, tendo em vista que as comunidades possuem uma unidade funcional precisa, com estruturas tróficas e padrões de corrente de energia característicos.

O trabalho de LEGENDRE, FORTIN (1989) também corrobora esta ideia, afirmando que as comunidades são conjuntos de ocorrências de plantas em que padrões podem ser observados: "*na natureza, os organismos frequentemente apresentam uma distribuição agregada, formando manchas, gradientes ou outros tipos de padrões espaciais*". Assim, as regras de associação obtidas pela aplicação do Apriori permitirão identificar as coocorrências entre as espécies.

4.1.2 Análise de Agrupamentos

O componente de análise de agrupamentos espacial tem por finalidade agrupar espacialmente os espécimes presentes na amostra. Nesta análise, a densidade de ocorrências das espécies é usada como um indicador da existência de uma comunidade; os agrupamentos obtidos serão considerados as regiões preferenciais onde as espécies coexistem. Observou-se o uso da análise de agrupamentos em níveis, tendo em vista o "método de controle dos fatores" (BAILEY, 1983) no qual a hierarquia espacial é construída pela sucessão de subdivisões em larga escala do ecossistema. No método, as ecorregiões foram divididas em três níveis (unidades ecológicas): *domínio, divisão* e *província*. Na escala global, primeiro nível, o clima foi considerado o principal fator controlador. No segundo nível, os *domínios* foram subdivididos com base no sistema de classificação climático regional dando origem às *divisões*, que foram novamente subdivididas gerando as *províncias* baseadas nas características da vegetação potencial dominante, que refletem melhor o clima. No presente estudo a clusterização pode

auxiliar na identificação das províncias em relação à hierarquia espacial citada anteriormente.

A Tabela 6 abaixo apresenta o relacionamento entre os conceitos obtidos da teoria estudada no Capítulo 2 com as categorias

Tabela 6 – Principais conceitos sobre comunidades e suas associações com as categorias da mineração de dados usadas na metodologia

Conceito	Análise
Agrupamentos	Associação, Agrupamentos
Caracterização	Agrupamentos
Competição	Associação, Agrupamentos
Confiança	Associação, Agrupamentos
Coocorrência	Associação, Agrupamentos
Grupos Funcionais	Associação, Agrupamentos
Limites	Associação, Agrupamentos
Localização	Agrupamentos

4.2 Fundamentação Teórica para a Determinação da Sequência de Aplicação das Análises

Foi considerado como fator determinante para a definição da sequência de aplicação a premissa apresentada em trabalhos como WITTMAN *et al.* (2010), na qual as comunidades compartilham das mesmas condições ambientais em termos de habitat. Assim, pelo fato dos pontos de ocorrência estarem localizados em uma mesma região (*plot*), com características ambientais semelhantes, foi considerada a aplicação da análise de agrupamentos após a identificação das coocorrências entre espécies. Neste caso, em uma escala local, a aplicação da clusterização tem por objetivo facilitar a identificação da existência de comunidades ou micro-habitats.

4.3 Visão Geral da Metodologia

Na definição da metodologia foram desenvolvidos três componentes principais para a obtenção de conhecimento: 1) Componente de levantamento dos espécimes frequentes na distância especificada em metros; 2) Componente de análise de associação das coocorrências existentes entre pares e agrupamentos de espécies e; 3) Componente de análise de agrupamentos baseados em densidade.

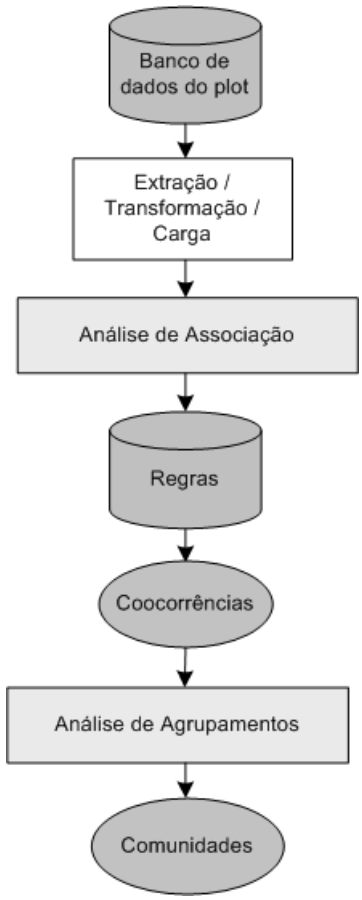


Figura 6. Visão macro da metodologia e os resultados gerados por cada componente

Na Figura 6 acima é apresentada a visão macro da metodologia com os principais componentes usados, exibindo a sequência definida para a geração do conhecimento. Após a aplicação do método os resultados possíveis estarão disponíveis para que estudos mais aprofundados possam ser realizados por parte dos especialistas na

investigação das razões ou hipóteses responsáveis pelas coocorrências e as regiões onde elas estão localizadas no *plot*.

São as seguintes as etapas a serem realizadas:

I. Extração e transformação dos dados de ocorrências de espécies de levantamentos florísticos em um formato de transações que permita o uso da análise de associação;
II. Execução da análise de associação para a identificação das coocorrências: positivas, negativas e de agrupamentos de espécies;
III. Avaliação das regras obtidas com o uso de métricas para permitir a seleção das melhores regras para cada um dos tipos identificados na etapa anterior;
IV. Análise de agrupamentos espaciais com o objetivo de verificar a existência de comunidades ou micro-habitats a partir da identificação das regiões ambientais propícias para tal existência.

4.4 Método Proposto para o Levantamento de Coocorrências

O método, apresentado na Figura 7, é dividido em duas etapas principais. A primeira delas objetiva a preparação dos dados para a aplicação do algoritmo da análise de associação, onde registros contendo espécimes que satisfazem a distância especificada para o raio em metros são convertidos em um formato transacional, necessário para a aplicação do Apriori; enquanto a segunda etapa é realizada pela aplicação do algoritmo e avaliação das regras geradas para reduzir o total de regras obtidas pela seleção das regras de maior qualidade. O cálculo da distância entre os pontos de ocorrência é realizado na primeira etapa do método, e isto é possível porque cada ponto de ocorrência tem suas localizações determinadas pelas coordenadas Cartesianas *Gx* (horizontal) e *Gy* (vertical), segundo o método de mapeamento descrito

em CONDIT (1998). Após a segunda etapa, as regras são incluídas no banco de dados para a extração do conhecimento.

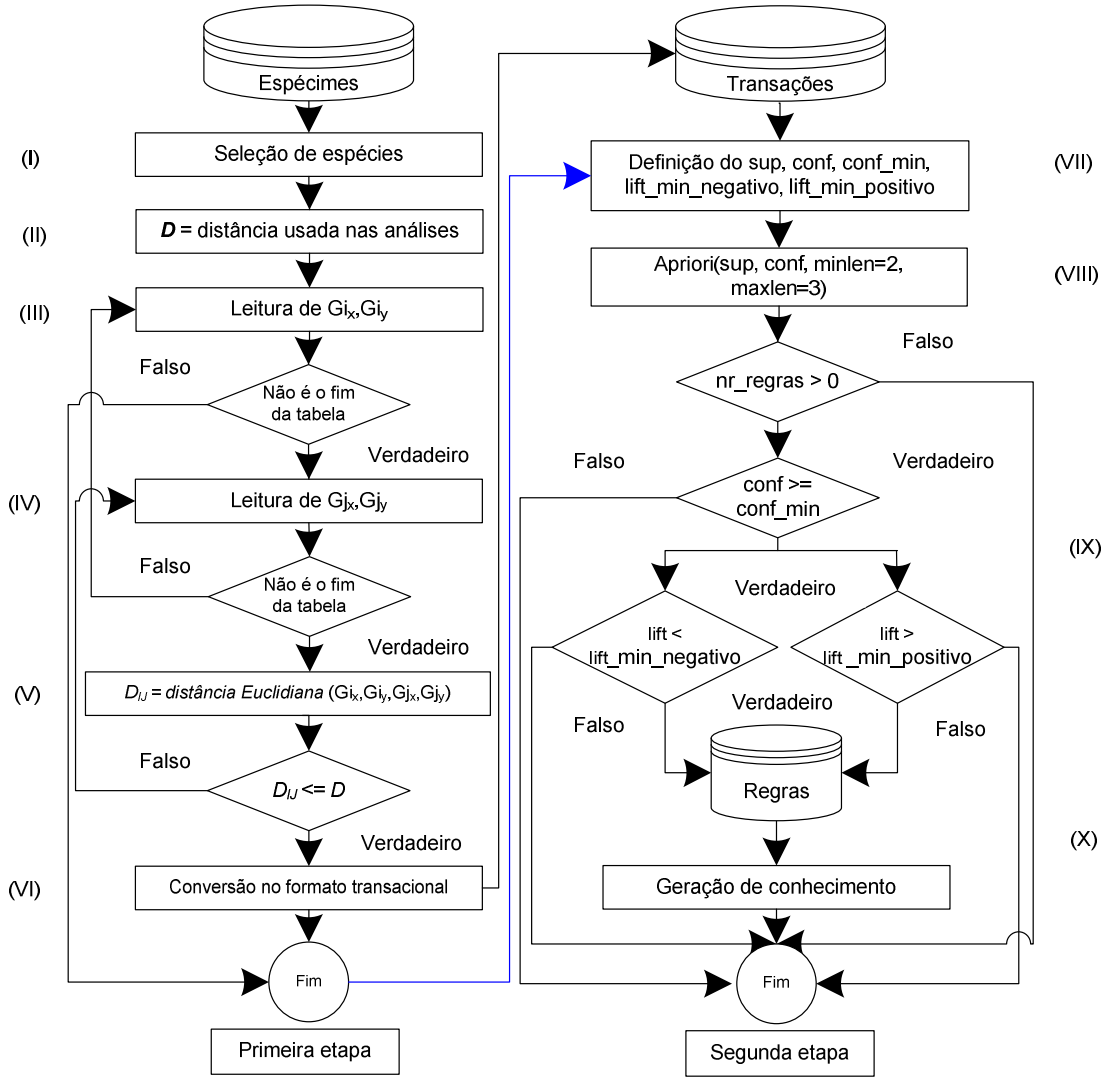


Figura 7. Estrutura do método proposto para o levantamento de padrões de coocorrências – as seguintes abreviaturas são usadas na figura: *Dij* = distância euclidiana entre os pontos *i* e *j*; *sup* = suporte mínimo; *conf* = confiança; *conf_min* = confiança mínima; *minlen* e *maxlen* definição do valor mínimo e máximo do número de elementos envolvidos nas regras; *nr_regras* = número de regras; *lift_min_negativo* = valor mínimo negativo do *lift*; *lift_min_positivo* = valor mínimo positivo do *lift*. A seta em azul indica a passagem entre as etapas

As etapas apresentadas na Fig. 7 são descritas a seguir:

I. Seleção das espécies e definição de outros critérios, como filtros. Na pesquisa, por exemplo, foi usado o valor do DAP > 100 mm. O diâmetro da altura do peito foi

definido como filtro na tese por ser uma medida de qualidade usada amplamente em pesquisas ecológicas caracterizando os árvores maduras ou “recrutadas”;

II. Definição do valor da distância máxima *D* (em metros) entre os pontos (espécimes) a ser utilizada na análise;

III. É realizada a leitura de cada ocorrência da espécie *i*, determinada pelo uso das coordenadas Gi_x e Gi_y, até o fim do banco de dados;

IV. É realizada a leitura de cada ocorrência da espécie *j*, determinada pelo uso das coordenadas Gj_x e Gj_y, até o fim do banco de dados;

V. É realizado o cálculo da distância euclidiana *Dij* para todos os pares de espécimes da base (representados por *j*) no banco de dados por:

$$D_{ij} = ((Gi_x - Gj_x)^2 + (Gi_y - Gj_y)^2)^{1/2};$$

VI. Se *Dij* for menor ou igual a distância *D* especificada, é realizada a agregação dos pontos próximos ao espécime base. Permitindo a transformação no formato transacional requerido para aplicação do Apriori e inclusão na tabela de transações;

VII. Verificação da espécie de frequência mais baixa para a definição do valor do suporte mínimo (*sup*), da confiança mínima (*conf*), do valor mínimo negativo do *lift* (*lift_min_negativo*) e do valor mínimo positivo do *lift* (*lift_min_positivo*);

VIII. Aplicação do algoritmo Apriori com os parâmetros: suporte mínimo *sup*, confiança mínima *conf_min* de verificação das regras, número mínimo (*minlen*) e máximo de elementos **(***maxlen*). Esses dois últimos parâmetros especificam se a análise será para avaliar pares ou grupos de espécies;

IX. Avaliação das regras armazenadas segundo as métricas de qualidade aplicadas na base de dados;

X. Geração do conhecimento com as regras de associação e os *clusters* obtidos.

4.5 Descrição das Etapas do Método

A seguir são acrescentados detalhes sobre os principais componentes do método segundo as etapas apresentadas na Figura 7.

4.5.1 Primeira Etapa – Extração, Transformação e Carga dos Dados

A pesquisa foi inicialmente desenvolvida com base na observação de que existe uma grande variação na distância entre dois espécimes em uma parcela. Outra razão é que pelo valor da distância, diferentes tipos de estudos podem ser realizados. Por exemplo, um estudo para identificar as preferências de habitats das espécies pode ser empreendido com um valor de até 20 metros e outro para analisar a competição entre espécies pode ser realizado com a distância de até 5 metros ou menor. Tudo depende da escala em que esses processos ocorrem. Dadas estas considerações, e na busca por resultados precisos, o método proposto exige a definição da distância em metros a ser usada na pesquisa e, a partir desta definição, é realizado o cálculo da distância euclidiana entre todos os pares de amostras no banco de dados. Como podem ocorrer várias possibilidades de distribuição das espécies no *plot*, a avaliação da distância entre os pontos de ocorrência pode auxiliar no estudo de alguns casos, sendo mais precisa do que apenas na identificação da parcela. Assim, a Figura 8 apresenta alguns exemplos de distribuição de pontos com apenas dois *plots*, nos quais as letras A a G indicam tipos de espécies e, os números são usados para indicar as amostras. A determinação da distância permite a inclusão nas análises de casos tais quais os que ocorrem quando dois espécimes estão localizados a poucos metros de distância, mas em diferentes parcelas, como por exemplo as ocorrências D_2 em *P1* e E_1 em *P2*. Neste caso, análises que consideram as parcelas como unidades de amostragem não poderiam identificar esse tipo de associação. Outro caso possível é aquele em que duas espécies ocorrem em

maior abundância quando estão próximas, como se mostra pelos pontos de espécies A e B em *P1*, indicando que as espécies são indicadores de habitat. O oposto é apresentado entre F e G, sugerindo correlação negativa.

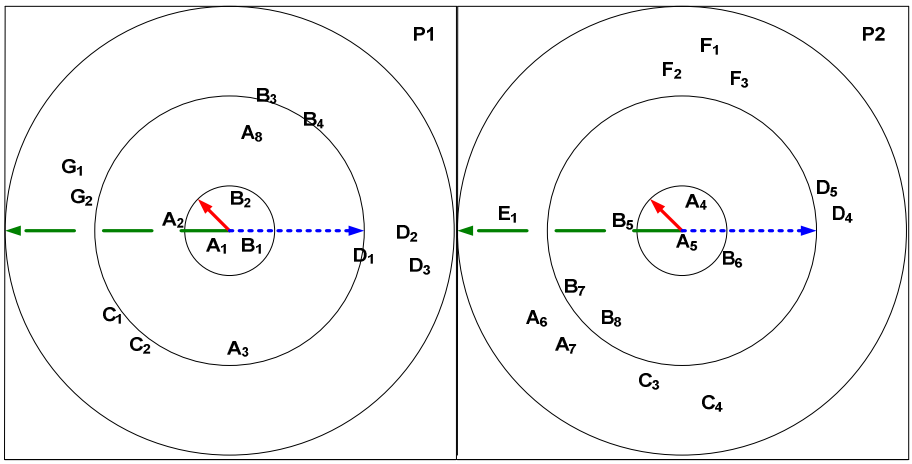


Figura 8. Visão geral da ocorrência das espécies em duas parcelas - as setas vermelhas indicam um raio de até 2.5 m em torno de duas ocorrências centrados nas diferentes parcelas P1 e P2; em azul pontilhado o raio de até 5 metros e de até 10 metros em verde

A Figura 9 facilita a visualização das unidades de aplicação obtidas e contém os espécimes próximos dentro do raio especificado. Cada conjunto de espécimes será transformado em uma transação para a aplicação da análise de associação.

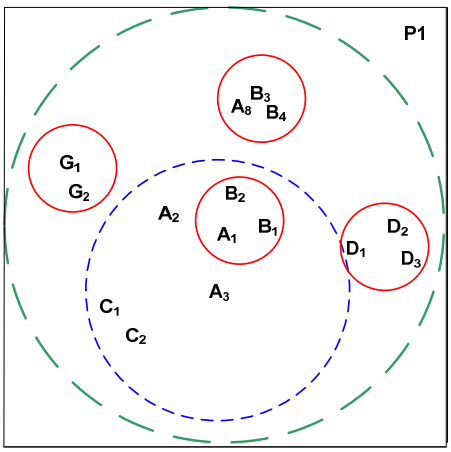


Figura 9. Unidades de aplicação (transações) obtidas com o método

4.5.2 Segunda Etapa – Aplicação das Análises de Associação e Agrupamento

Após a etapa de identificação e transformação dos dados, é realizada a aplicação dos algoritmos de análise de associação e de agrupamentos por meio dos componentes, descritos a seguir:

4.5.2.1 Componente de Associação

De forma similar a análise de associação realizada a partir do levantamento dos itens frequentes em transações de cestas de compras, a granularidade definida como unidade de estudo para a aplicação da análise de associação na tese foi o conjunto de espécimes presentes dentro do raio definido em metros, sendo esses itens obtidos após o cálculo da distância realizada no passo número V do método apresentado na Figura 7.

Na pesquisa bibliográfica realizada na tese não foram encontrados trabalhos relacionados com a aplicação de algoritmos de associação da mineração de dados no processo de análise de coocorrências de espécies vegetais com dados de levantamentos florísticos em parcelas. Essa abordagem se destaca ainda por permitir o levantamento de coocorrências além de pares de espécies, como ocorre com a Análise de Padrões de Pontos Espaciais. Apresentando ainda como outra vantagem, o fato de não exigir de uma pré-seleção de espécies para diminuir o custo de execução dos testes, como citado no fim do Capítulo 2.

4.5.2.2 Métricas Usadas na Avalição das Regras Geradas

Para o obtenção do conhecimento sobre as espécies há a necessidade da avaliação das regras. Em relação às métricas, existem diversas propostas na literatura para avaliação da qualidade das regras geradas pela análise de associação. Neste trabalho foram avaliadas várias métricas e para a escolha do conjunto final o critério foi

a seleção das métricas que permitiram a identificação de coocorrências entre os itens obtidos pela análise de associação. Segundo TAN (2005), se existir uma consistência nas medidas apresentadas, então pode-se utilizar escolher qualquer. O conjunto de métricas avaliadas é apresentado na Tabela 7.

Em relação as métricas básicas da análise de associação, o suporte *sup* é a porcentagem de transações onde um *itemset* está contido. O contador de suporte $\boldsymbol{\sigma}$ de um itemset $\boldsymbol{z}$ é declarado como: $\sigma(z) = |\{t\_sp_i | z \subseteq t\_spi, t\_sp_i \in t\_sp\}|$

Onde |.| expressa o número de elementos de um conjunto de espécimes de uma determinada espécie sp_i. O suporte é expresso como sup $(sp\_x \rightarrow sp\_y) = \sigma (sp\_x \cup sp\_y) / n$, onde $\boldsymbol{n}$ é o número total de transações *t_sp*.

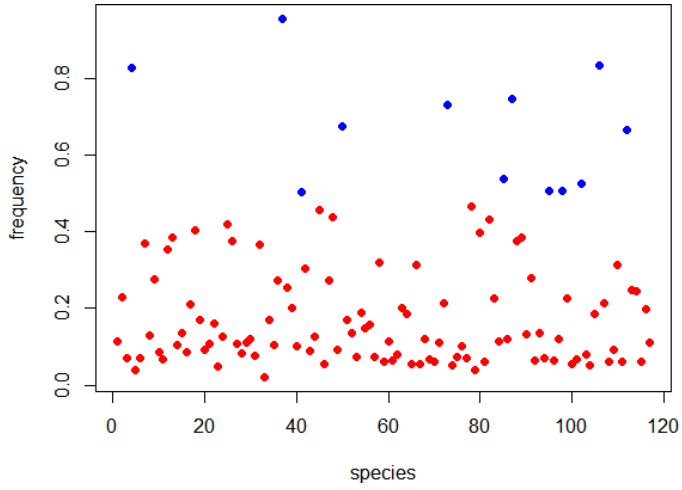


Figura 10. Distribuição da frequência na distância máxima de até 20 metros

Para a aplicação do algoritmo, existe a necessidade de determinação do valor do suporte. Porém, isso depende do número de registos, é importante notar a frequência das espécies analisadas. Altos valores de suporte podem excluir das análises as espécies com baixa frequência de ocorrência e limitar a avaliação apenas as espécies mais frequentes. Por exemplo, a Figura 10 apresenta em azul apenas 12 espécies com frequência relativa acima de 0.5 na distância de até 20 metros de um total de 117

espécies. A confiança ***conf***, é definida como a frequência na qual os itens da *sp_y* são encontrados nas transações que contém *sp_x*. A confiança é expressa como:

$$conf\ (sp\_x \rightarrow sp\_y) = \sigma\ (sp\_x \cup sp\_y) / \sigma\ (sp\_x)$$

Para que o algoritmo Apriori possa identificar os padrões de coocorrências positivas e negativas entre espécies (SANJEREHEI, 2011) há a necessidade do uso de métricas. O *lift* é uma medida objetiva de avaliação da qualidade de uma regra muito utilizada porque o uso somente da combinação suporte-confiança muitas vezes conduz a avaliações limitadas, tendo em vista que no cálculo da confiança o suporte do conjunto de itens do consequente da regra é ignorado (TAN *et al*., 2009). O *lift* identifica associações assumindo que os itens ocorrem inicialmente independentemente um dos outros (SILVERSTEIN, 1997) e é expresso como:

$$\sigma\ (sp\_x \cup sp\_y\ ) / (\sigma\ (sp\_x\ ) * \sigma\ (sp\_y))$$

O uso do *lift* será exemplificado no Capítulo 5, que traz os resultados dos testes realizados nos estudos de casos. O uso comum da métrica é aquele que permite a identificação das correlações positivas e negativas entre os itens envolvidos nas regras geradas. Seus valores podem ser interpretados da seguinte forma: um valor igual a um significa que há independência entre o antecedente (LHS – *left hand side*) e o consequente (RHS – *right hand side*) da regra; um valor maior que um significa que o relacionamento é positivo, ou seja, itens aparecendo com uma frequência maior do que a esperada e; um valor menor que um indica um relacionamento negativo, ou seja, a presença de um item implica na diminuição do outro. Além das 3 métricas citadas, foram utilizadas ainda na busca por qualidade nos resultados mais 9 métricas: *chi-square*¸ *p-value*, *coverage*, *conviction*, Gini, *HyperLift*, *HyperConfidence***,** *Leverage* e *OddsRatio*.

Tabela 7 – Principais métricas da análise de associação avaliadas para uso na metodologia

Nr	Métrica	Definição / Uso	Variação (min:max)
1	**Suporte** (*support*)	É a porcentagem de transações que contém X e Y. Um itemset com um valor de suporte maior que o valor mínimo é chamado de frequente. Seu uso apresenta desvantagem quanto aos itens raros, que são excluídos, embora poderiam produzir regras interessantes. Isto ocorre em transações com alta desigualdade no suporte dos itens	sup (X → Y) = σ (X ∪ Y) / n, onde *n* é o total de transações onde X e Y ocorrem e σ representa o contador de suporte do itemset. Varia de 0 a 1
2	**Confiança** (*confidence*)	A confiança indica a quantidade de transações contendo X e que também contém Y	conf (X → Y) = sup (X ∪ Y) / sup (X). Varia de 0 a 1
3	***Lift***	Utilizado para a verificação de correlação entre o LHS e o RHS da regra. É também conhecido como fator de interesse e, é usado por que regras de alta confiança podem, às vezes, gerar resultados enganosos, pois a medida de confiança ignora os itens que aparecem no consequente da regra	*lift* (X → Y) = conf (X → Y) / sup(Y) Valor igual a 1 indica independência entre LHS e RHS; valor < 1 indica correlação negativa e; > 1 indica correlação positiva
4	**Chi-Square** (HAHSLER *et al.*, 2013, LIU *et al.*, 1999)	É usado para testar a independência entre o LHS e o RHS da regra. Altos valores de qui-quadrado indicam que o LHS e o RHS não são independentes. Qui-quadrado = Σ (observado - esperado)^2 / esperado	Valor igual a 1 indica independência entre LHS e RHS; para um valor fora do limite entre 0,05 e 3.84, a independência rejeitada
5	Teste Exato de Fisher (***p-value***) (LIU *et al.*, 2011)	É um teste estatístico de significância usado na análise de tabelas de contingência. Retorna o *p-value* quando usado em regras de associação	Valores baixos indicam que a regra tem pouca chance de ocorrer com itens independentes
6	**Coverage** (HAHSLER *et al.*, 2013)	Calcula a cobertura da regra (suporte do LHS). Foi usada para verificar a diferença de suporte entre o LHS e o RHS da regra	Coverage (X → Y) = sup (X) = P(X). Varia de 0:1
7	**Conviction** (JORGE, AZEVEDO, 2005)	Calculada como: *conviction* (X → Y) = 1 – sup (Y) / 1 – conf (X → Y), mede a independência entre os itens envolvidos	Varia de 0:5 a 1 e de 1:infinito. O valor 1 indica independência

8	**Gini** (TAN *et al*., 2002)	O índice de Gini é uma métrica que não é simétrica, ou seja, o valor para a regra X → Y é diferente de Y → X. É usado em vários contextos e quantifica impurezas	Valor igual a zero indica independência entre itens; o valor entre 0 e 0.5 indica dependência
9	**HyperLift** (HAHSLER *et al*., 2013)	É uma adaptação do *lift* para sofrer uma influência menor em análises com poucos registros	Varia de 0:infinito
10	**HyperConfidence** (HAHSLER *et al*., 2013)	Calcula o nível de confiança para valores muito altos em relação a itens com baixa quantidade usando um modelo hipergeométrico	Varia de 0:1
11	**Leverage** (HAHSLER *et al*., 2013)	Definido como P(X → Y) – (P (X)P(Y)), mede a diferença entre os itens X e Y aparecendo juntos, em relação ao que seria esperado no caso de dependência estatística	Varia de -1:1
12	**OddsRatio** (HAHSLER *et al*., 2013)	Mede as chances de encontrar X em transações que contém Y divididas pelas chances de encontrar X em transações que não contenham Y	Varia de 0:1 e de 1:infinito O valor 1 indica que não tem associação
13	RLD (KENETT, SALINI, 2008)	O *Relative Linkage Disequilibrium* avalia o desvio do suporte de todos os itens de uma regra a partir do suporte esperado sob independência do antecedente e do consequente da regra	Varia de 0:1
14	Coseno (TAN *et al*., 2009)	A métrica coseno pode ser entendida como a métrica *lift* harmonizada. A diferença para o *lift* é que o coseno é influenciado somente pelo suporte de X, Y e X ∪ Y e não pelo número total de transações. Nos testes realizados, foi eficiente apenas quando aplicado em agrupamentos com correlação positiva previamente identificada, com o uso da métrica *lift*, por exemplo	Varia de 0:1. Valor > 0.5 indica correlação positiva; um valor = 0.5 independência e; Valor < 0.5 correlação negativa
15	Phi (TAN *et al*., 2009)	Não aplicado para variáveis binárias	Varia de -1:1

4.5.2.3 Componente de Análise de Agrupamentos

O objetivo do componente de análise de agrupamentos é o de mapear as regiões onde as espécies apresentaram coocorrências, identificadas na etapa anterior com a análise de associação. A aplicação da análise de agrupamentos pode ser útil para a promoção da investigação sobre os motivos ou hipóteses da formação das comunidades por habitat, presentes no *plot*. Como a clusterização pode ser afetada pela seleção de parâmetros, como o número de *clusters* e a distância entre os pontos, foi selecionado o algoritmo EM (*Expectation Maximization*) com o *package Mclust* (FRALEY, RAFTERY, 2006) que permite a aplicação com os parâmetros padrões, sendo esses adequados na maioria dos casos.

4.5.3 Requisitos Não-Funcionais

Foi considerada na implementação da proposta o uso de *software* livre. A implementação foi feita com o programa estatístico R (R DEVELOPMENT CORE TEAM, 2010), com o uso do pacote *Arules* para o algoritmo Apriori (HAHSLER *et al.*, 2005) e o do pacote *Mclust* (FRALEY, RAFTERY, 2006). Foi utilizado o sistema gerenciador de banco de dados *Postgresql*[5].

Como justificativas para o uso do R podem ser listadas:

- R é um dos *software*s mais utilizados na mineração de dados. Apresenta recursos para, praticamente, todas as fases do processo de extração do conhecimento e disponibiliza diversos algoritmos nas várias categorias de aplicações da mineração de dados (TAN *et al.*, 2005);
- O R é também amplamente usado na Biologia[6];
- O R é livre, de código aberto e tem uma linguagem de programação própria;

[5] www.postgresql.org
[6] http://r-forge.r-project.org/softwaremap/trove_list.php

- O R possui inúmeros projetos e pacotes (*packages*) ofertados em praticamente todos os sistemas operacionais;
- Permite a importação e exportação de dados em diferentes formatos e sua integração com bancos de dados, tais como o PostgreSQL.

4.5.4 Modelo de Dados do Repositório de Regras

O repositório de regras foi desenvolvido, entre outros fatores, para preservar e facilitar o acesso ao conhecimento obtido. Mantendo os dados em um SGBD, como o PostgreSQL, obtém-se como vantagem da conexão direta com o R a possibilidade de exploração dos dados de formas variadas, pois o uso da Linguagem de Consulta Estruturada (SQL) possibilita uma maior flexibilidade no desenvolvimento das consultas. As classes podem ser divididas em dois grupos:

- Classes comuns: são as classes necessárias para o controle das espécies (*specie*), dos espécimes (*specimen*) e dos censos (*censu* e *specimen_censu*) do *plot* e;
- Classes necessárias aos experimentos – essas classes são usadas para o armazenamento das transações (*transaction*) produzidas segundo as distâncias usadas nos experimentos, do controle dos detalhes de cada experimento (*experiment*) e das regras (*rules* e *specie_rule*) obtidas com a aplicação da análise de associação.

A Figura 11 apresenta o diagrama de classes da UML, mapeando as tabelas do banco de dados utilizadas como repositório das regras geradas pela análise de associação.

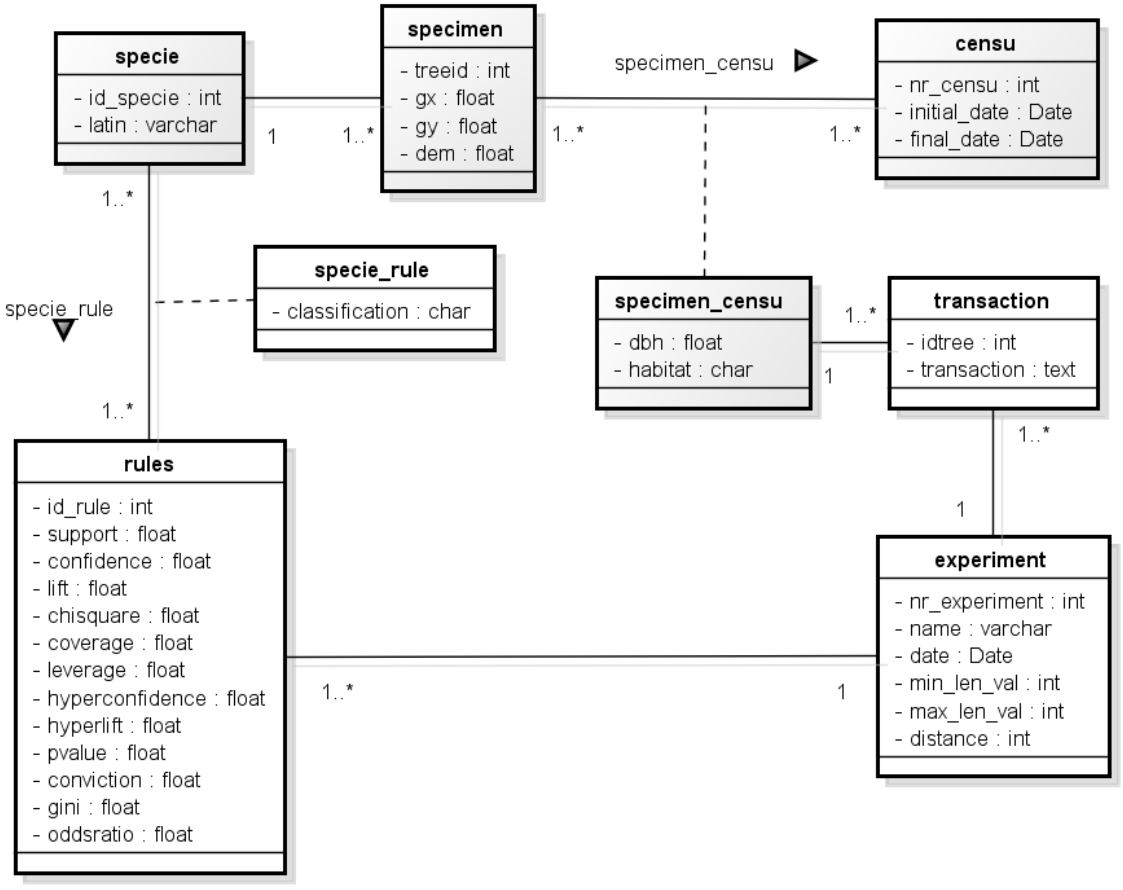


Figura 11. Estrutura do banco de dados

A classe ***specie*** contém a lista de espécies presentes na base de dados. Os nomes científicos são armazenados no atributo *latin*.

O conjunto de espécimes do *plot* fica registrado na classe ***specimen***, onde cada árvore recebe um identificador próprio *treeid* e, onde são armazenados também valores como as coordenadas desses espécimes nos atributos *gx* e *gy* e, a medida da elevação do ponto de ocorrência pelo atributo *dem*.

Os dados dos censos são mantidos na classe ***censu*** com atributos que permitem a identificação do período de realização do levantamento. O controle do período pode ser útil, por exemplo, para o desenvolvimento de pesquisas sobre mudanças climáticas.

A classe de associação ***specimen_censu*** controla o armazenamento do histórico das medidas de cada espécime por censo realizado, para possibilitar o acompanhamento da evolução dos espécimes. Nesta pesquisa foram usados os atributos: *dbh* (*diameters at breast height*), o mesmo que o DAP e; o tipo de *habitat*.

A classe ***transaction*** tem por finalidade armazenar o conjunto de transações geradas para cada experimento realizado com a aplicação da análise de associação.

A classe ***experiment*** contém os detalhes dos testes realizados, com destaque para o atributo *distance* usado como parâmetro do método de construção das transações para a especificação da distância máxima entre os espécimes analisados. Os atributos *min_len_val* e *max_len_val* são usados para controlar o número de itens envolvidos na análise de associação.

A classe ***rules*** foi criada para o controle das regras geradas e das métricas selecionadas para o estudo: *support, confidence, lift, chisquar, coverage, leverage, hyperconfidence, hyperlift, pvalue, conviction, gini* (índice Gini) e *oddsratio*.

A classe ***specie_rule*** serve para identificar as espécies envolvidas nas regras e o tipo de envolvimento de cada uma. Os tipos tem suas classificações armazenadas no atributo *classification* e são identificados pelas siglas *lhs* (*left hand side*) para o antecedente da regra e *rhs* (*right hand side*) para o consequente da regra.

Capítulo 5 – Estudos de Casos e Avaliação da Proposta

5.1 Base de Dados – *Barro Colorado Island*

Os dados usados na pesquisa pertencem ao Projeto Floresta Dinâmica (FDP) da Ilha de Barro Colorado no Panamá (Barro Colorado Island - BCI) (HUBBELL *et al.*, 1999, HUBBELL *et al.*, 2005). Os dados do projeto são atualizados por meio de censos realizados a cada 5 anos para cada espécime com a medida do diâmetro da altura do peito (DAP) maior ou igual a 10 mm. O banco de dados analisado é referente ao censo realizado em 2010 e é composto de 277.351 registros, contendo o total de 312 espécies. Desse total, 221 espécies possuem espécimes com DAP > 100 mm. Foi definida ainda a quantidade mínima de 20 pontos de ocorrência por espécie para uma maior robustez nos resultados. Com esses filtros, o número de espécies foi reduzido para 117 ao final. Este limite de corte seleciona espécies com a frequência mínima igual a 0.001 para uso no estudo de caso com distância máxima de até 5 metros (escala pequena) e 0.02 para a distância de até 20 metros (escala larga). Estas distâncias são similares aos estudos realizados em (LAN *et al*. 2012).

5.2 Análise Exploratória dos Dados

Para facilitar a análise dos resultados que serão apresentados nos estudos de casos, alguns gráficos serão apresentados nesta seção. A Figura 12 mostra a distribuição dos espécimes pelos 7 tipos de habitats presentes no *plot* que é organizado em um total de 1250 parcelas medindo 20 metros^2 em uma área da ilha de 50 hectares.

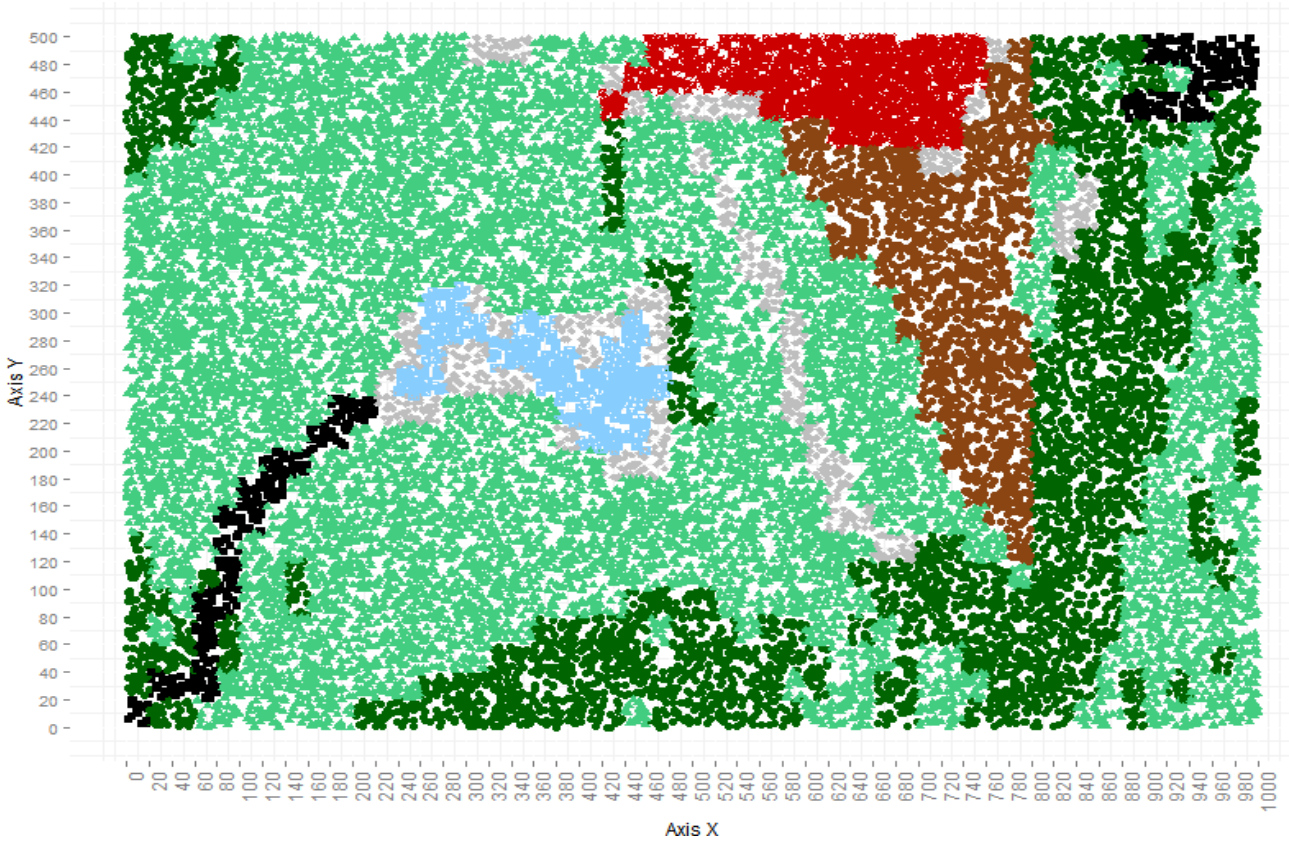


Figura 12. Distribuição de espécimes por habitat no *plot* – distribuição dos espécimes com DAP maior que 100 mm. Os habitats são identificados por: W – pântano (azul), L – baixo platô (verde claro), H – alto platô (marrom), S – encosta (verde escuro), T – beira de córrego (preto), Y – floresta nova (vermelho), e M – habitat misto (cinza) – Adaptada de HARMS *et al*. (2001)

Na Figura 13 é apresentada a distribuição da quantidade de registros por espécies (eixo das coordenadas), destacando a presença da maioria das espécies com baixa quantidade de registros e poucas espécies com alta quantidade.

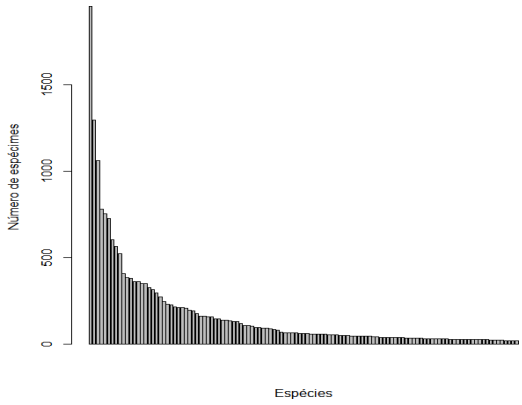


Figura 13. Distribuição do total de registros por espécies.

Na Figura 14 é exibida a distribuição da quantidade de registros pelas 1250 parcelas (coordenadas) do plot de *Barro Colorado Island*, caracterizando uma distribuição similar em grande parte das parcelas e, uma floresta com alta densidade. Isto é importante para aumentar a confiança nos resultados, pois a existência de parcelas com poucas ocorrências poderia influenciar nos resultados.

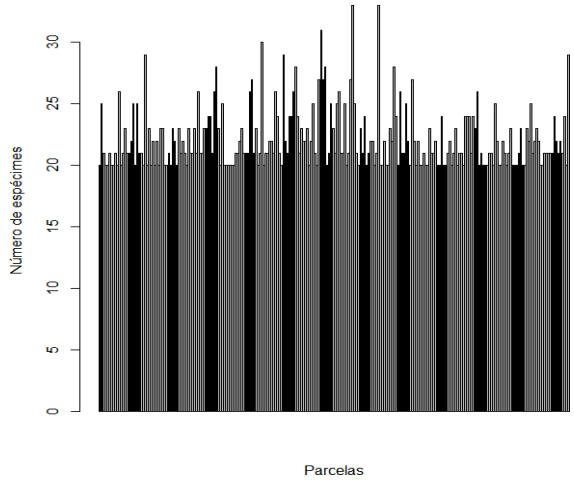


Figura 14. Distribuição da quantidade de registros por parcelas

A Figura 15 apresenta a distribuição das frequências relativas das espécies para as distâncias de 5 (esquerda) e 10 metros (direita), respectivamente. Em vermelho são exibidas as espécies com frequência abaixo de 50% e, em azul, acima ou igual.

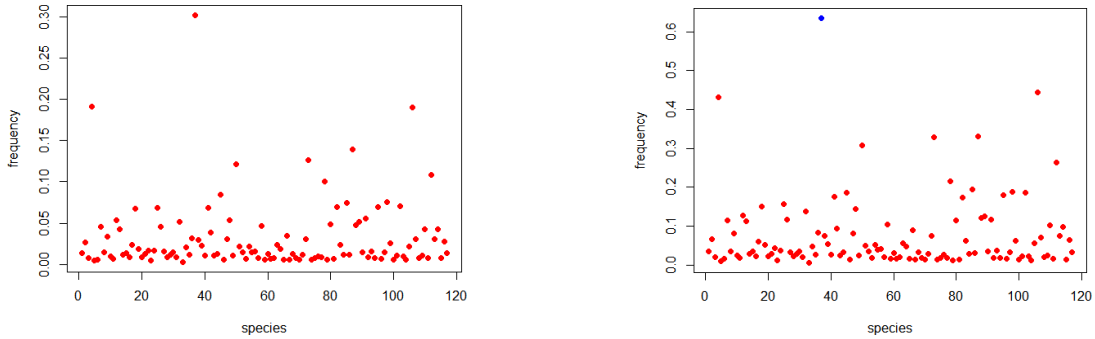


Figura 15. Distribuição das frequências das espécies

A Tabela 8 apresenta as 117 espécies analisadas e as suas respectivas quantidades. O objetivo da tabela é o de permitir avaliar os resultados dos estudos de casos em termos da abundância das espécies envolvidas nos resultados.

Tabela 8 – Lista de espécies analisadas e suas respectivas quantidades

Espécie	Total	Espécie	Total	Espécie	Total
Adelia triloba	64	*Guarea bullata*	42	*Pourouma bicolor*	20
Alchornea costaricensis	134	*Guarea guidonia*	352	*Pouteria reticulata*	214
Allophylus psilospermus	32	*Guatteria dumetorum*	164	*Pouteria stipitata*	28
Alseis blackiana	1059	*Guazuma ulmifolia*	42	*Prioria copaifera*	362
Anacardium excelsum	21	*Guettarda foliacea*	57	*Protium costaricense*	111
Andira inermis	31	*Gustavia superba*	604	*Protium panamense*	48
Apeiba membranacea	197	*Hampea appendiculata*	29	*Protium tenuifolium*	410
Aspidosperma spruceanum	60	*Hasseltia floribunda*	149	*Pterocarpus rohrii*	52
Astrocaryum standleyanum	147	*Heisteria concinna*	296	*Quararibea asterolepis*	728
Astronium graveolens	38	*Hieronyma alchorneoides*	41	*Randia armata*	232
Attalea butyracea	32	*Hirtella triandra*	782	*Simarouba amara*	230
Beilschmiedia pendula	249	*Hura crepitans*	90	*Sloanea terniflora*	64
Brosimum alicastrum	195	*Inga acuminata*	69	*Socratea exorrhiza*	273
Calophyllum longifolium	49	*Inga goldmanii*	32	*Spondias mombin*	37
Casearia arborea	59	*Inga marginata*	99	*Spondias radlkoferi*	62
Casearia sylvestris	38	*Inga nobilis*	68	*Sterculia apetala*	31
Cassipourea elliptica	110	*Inga sapindoides*	69	*Swartzia simplex*	352
Cecropia insignis	363	*Inga thibaudiana*	38	*Tabebuia guayacan*	27
Cecropia obtusifolia	121	*Jacaranda copaia*	208	*Tabebuia rosea*	57
Ceiba pentandra	40	*Lacistema aggregatum*	25	*Tabernaemontana arborea*	383
Celtis schippii	48	*Lacmellea panamensis*	55	*Tachigali versicolor*	106
Chrysophyllum argenteum	71	*Laetia thamnia*	28	*Terminalia amazonia*	26
Chrysophyllum cainito	21	*Lindackeria laurina*	40	*Terminalia oblonga*	44
Cordia alliodora	60	*Lonchocarpus heptaphyllus*	95	*Tetragastris panamensis*	386
Cordia bicolor	327	*Luehea seemannii*	86	*Trattinnickia aspera*	41
Cordia lasiocalyx	212	*Macrocnemum roseum*	25	*Trema integerrima*	29
Coussarea curvigemmia	66	*Maquira guianensis*	159	*Trichilia pallida*	99
Croton billbergianus	49	*Maytenus schippii*	24	*Trichilia tuberculata*	1294
Cupania seemannii	50	*Miconia argentea*	58	*Triplaris cumingiana*	131
Dendropanax arboreus	61	*Nectandra cissiflora*	29	*Trophis caucana*	38
Dipteryx oleifera	32	*Ocotea cernua*	28	*Turpinia occidentalis*	41
Drypetes standleyi	316	*Ocotea oblonga*	47	*Unonopsis pittieri*	178
Elaeis oleifera	20	*Ocotea whitei*	162	*Virola multiflora*	28
Eugenia coloradoensis	83	*Oenocarpus mapora*	754	*Virola sebifera*	524
Eugenia nesiotica	48	*Perebea xanthochyma*	27	*Virola surinamensis*	141
Eugenia oerstediana	159	*Picramnia latifolia*	31	*Xylopia macrantha*	215
Faramea occidentalis	1950	*Platymiscium pinnatum*	43	*Zanthoxylum acuminatum*	27
Garcinia intermedia	137	*Platypodium elegans*	30	*Zanthoxylum ekmanii*	139
Guapira standleyana	94	*Poulsenia armata*	567	*Zanthoxylum panamense*	52

5.3 Estudos de Casos

Três estudos de casos foram realizados para verificar a eficiência da proposta na identificação positiva e negativa de coocorrências entre pares ou grupos de espécies, associando essas ocorrências com os habitats mostrados na Figura 12. Nos testes foram utilizadas como parâmetros as distâncias máximas entre os espécimes de 5 até 20 metros. Os estudos de coocorrências de espécies seguem uma tendência atual para analisar pares de espécies (VEECH 2013). Porém, a abordagem proposta também se mostrou eficiente no levantamento das coocorrências em grupos de espécies.

A aplicação da primeira etapa do método resultou na criação de transações contendo 18.686 e 19.647 itens para as distâncias informadas anteriormente, destacando o potencial de extração de conhecimento em grandes volumes de dados. É possível melhorar ainda mais o conhecimento sobre os resultados obtidos utilizando medidas subjetivas de interesse, tais como a *visualização* (TAN *et al*., 2009). Portanto, as Figuras numeradas de 16 a 22 foram criadas para exibir a distribuição espacial das espécies em gráficos de dispersão (*scatter plots*) com os maiores valores nas métricas utilizadas.

Nos gráficos de dispersão, foram usadas elipses para realçar a concentração da distribuição dos dados. Em outros casos, a regressão linear suave mostra a distribuição de dados geográficos de uma forma melhor. O total de espécies presentes no estudo de caso com a distância máxima de até 20 metros foi de 29 espécies e essas encontram-se listadas na Tabela 9, juntamente com os valores das frequências e a quantidade de espécimes por habitat. A legenda para os habitats segue a que foi apresentada na Figura 12.

Tabela 9 – Frequência relativa e distribuição das espécies por habitat

Espécie	Frequência	S	W	L	H	T	Y	M	Total
Adelia triloba	0.1146231	6	0	**29**	16	0	11	2	64
Attalea butyracea	0.06570978	5	7	**10**	2	4	0	4	32
Beilschmiedia pendula	0.3538963	83	0	**129**	20	4	1	12	249
Casearia sylvestris	0.0867817	7	1	**16**	4	1	6	3	38
Drypetes standleyi	0.3650939	129	2	**148**	22	5	2	8	316
Guarea guidonia	0.5030285	100	5	**185**	32	12	8	10	352
Guatteria dumetorum	0.3044231	64	0	**82**	5	4	3	6	164
Gustavia superba	0.4568636	56	2	146	87	21	**266**	26	604
Inga goldmanii	0.07303914	8	4	**17**	0	1	0	2	32
Inga thibaudiana	0.07456609	2	0	**35**	0	1	0	0	38
Lacmellea panamensis	0.1127908	7	4	**25**	8	3	3	5	55
Ocotea whitei	0.2137222	**111**	0	43	1	4	1	2	162
Perebea xanthochyma	0.05018578	3	0	**24**	0	0	0	0	27
Platypodium elegans	0.07146129	4	0	**19**	2	1	1	3	30
Poulsenia armata	0.4652619	**313**	0	212	8	24	3	7	567
Pourouma bicolor	0.04051509	6	0	**14**	0	0	0	0	20
Pterocarpus rohrii	0.1184405	7	5	**32**	0	1	2	5	52
Quararibea asterolepis	0.746119	180	4	**374**	77	23	31	39	728
Socratea exorrhiza	0.2775487	91	0	**171**	4	3	1	3	273
Spondias mombin	0.0651499	0	7	**17**	0	5	1	7	37
Tabernaemontana arborea	0.5055225	49	10	**267**	11	10	12	24	383
Terminalia oblonga	0.06601517	10	0	**31**	0	0	2	1	44
Tetragastris panamensis	0.5253219	81	4	**256**	12	7	5	21	386
Triplaris cumingiana	0.2128569	22	20	**53**	18	3	0	15	131
Trophis caucana	0.0618924	**30**	0	6	0	2	0	0	38
Unonopsis pittieri	0.3118033	56	0	**105**	5	4	2	6	178
Virola multiflora	0.05955108	**16**	0	11	1	0	0	0	28
Virola sebifera	0.6642235	146	1	**296**	34	8	22	17	524
Xylopia macrantha	0.2453301	**105**	2	98	6	1	1	2	215
Zanthoxylum ekmanii	0.1973838	34	0	**97**	5	2	0	1	139
Total		**1731**	**78**	**2948**	**380**	**154**	**384**	**231**	**5906**

5.3.1 Análise de Associação Par-a-Par com Correlação Positiva

Como resultado, não foram observadas regras caracterizando padrões de coocorrências na distância de até 5 metros, tanto positiva quanto negativamente, mas em uma distância de 20 metros a aplicação do método resultou em 8 regras envolvendo 11 espécies no total. Com o objetivo de selecionar as melhores regras em termos de qualidade, foi aplicado o limite de corte para métrica com o valor *lift* >= 2.0. Dois dos pares de espécies obtidos pela análise são mostrados nas Figuras 16 e 17. Na Figura 16,

a aplicação do método mostra que o centro de distribuição de espécies *Virola multiflora* e *Xylopia macrantha* é alto no habitat encosta, com 44.8% de ocorrências.

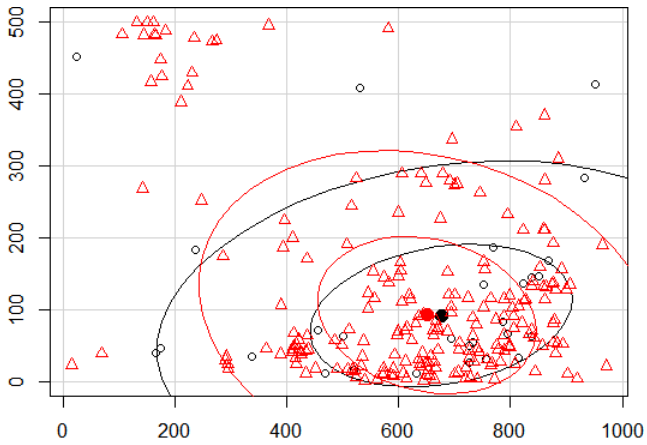


Figura 16. Distribuição espacial com coocorrência positiva entre as espécies *Virola multiflora* (preta) e *Xylopia macrantha* (vermelha)

Como conhecimento gerado, as métricas desta regra indicaram que em 4% (suporte) do conjunto de dados foram encontradas ambas espécies e em 60% (confiança) dos casos, quando encontrada *Virola multiflora* também foi encontrada *Xylopia macrantha* a uma distância máxima de 20 metros. Na Figura 17, para 50% dos casos com ocorrências da espécie *Terminalia oblonga* também foram observadas ocorrências de *Ocotea whitei*, totalizando 3% da base de dados. Para facilitar a visualização do resultado, em vez de utilizar os centros das distribuições, a regressão linear foi utilizada no gráfico de dispersão que mostra valores semelhantes ocorrendo nos habitats do tipo baixo platô e encosta. O *p-value* para todas as regras foi 0. O Apêndice B mostra os *scatter plots* das espécies com correlação positiva.

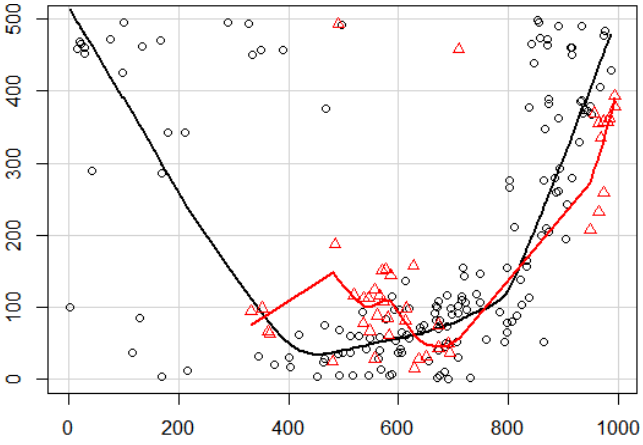


Figura 17. Distribuição espacial com coocorrência positiva entre as espécies *Ocotea whitei* (preta) e *Terminalia oblonga* (vermelha)

5.3.2 Análise de Associação Par-a-Par com Correlação Negativa

Para um valor de *lift* <= 0.7, foram criadas 10 regras em um total de 15 espécies, sendo 3 delas exibidas nas Figuras 18 e 19, apresentando uma correlação negativa em alguns habitats. O *p-value* para todas as regras foi 1.

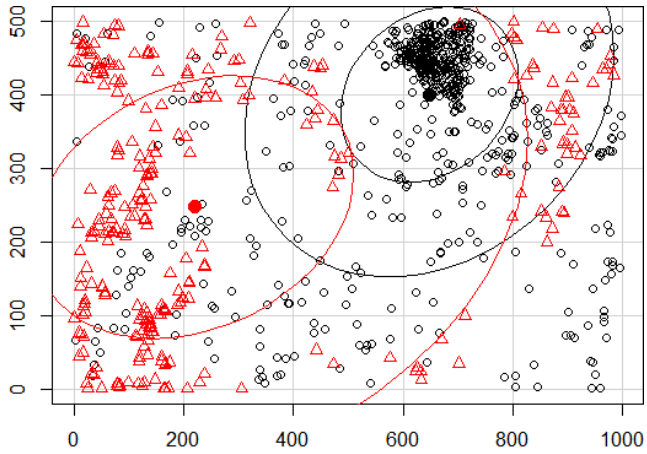


Figura 18. Distribuição espacial com coocorrência negativa entre as espécies *Gustavia superba* (preta) e *Socratea exorrhiza* (vermelha)

A Figura 18 exibe a espécie *Gustavia superba* com uma alta incidência no habitat do tipo floresta nova e a espécie *Socratea exorrhiza* no baixo platô. Ambas as espécies aparecem juntas em apenas 13% do total de ocorrências. Foi verificado que

apenas um exemplar da *Socratea exorrhiza* foi encontrado na floresta nova. A Figura 19 a distribuição ampla das espécies *Tetragastris panamensis* e *Triplaris cumingiana*, mas com preferências contrárias de habitat. A porcentagem de coocorrênca das duas espécies foi de 13.8% em 20 metros.

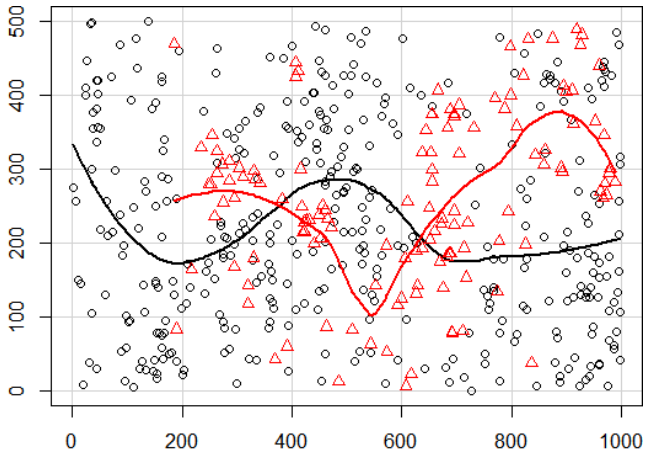


Figura 19. Distribuição espacial com coocorrência negativa entre as espécies *Tetragastris panamensis* (preta) e *Triplaris cumingiana* (vermelha)

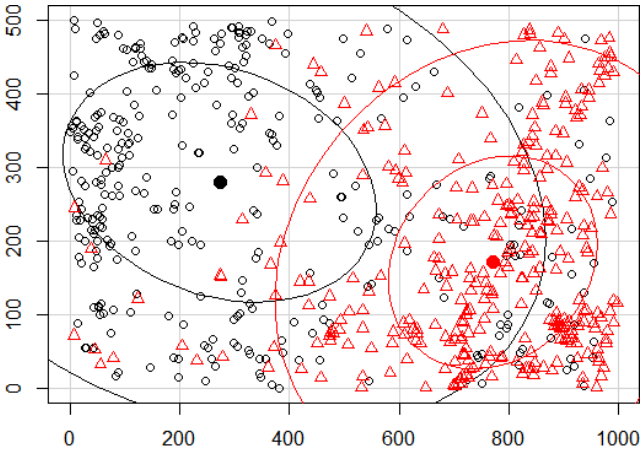


Figura 20. Distribuição espacial com coocorrência negativa entre as espécies *Cordia bicolor* (preta) e *Drypetes standleyi* (vermelha)

O método também foi eficiente para detectar correlações negativas mesmo com espécies que apresentam pequenas diferenças nas frequências. Um exemplo é apresentado na Figura 20 entre as espécies *Cordia bicolor* e *Drypetes standleyi* com

uma diferença de somente 0.05. O Apêndice C mostra *scatter plots* para todas as espécies com correlação negativa.

5.3.3 Análise de Associação para Agrupamentos de Espécies

A Figura 21 mostra a ocorrência das espécies *Ocotea whitei*, *Virola multiflora*, e *Xylopia macrantha* destacando o centro da distribuição. Com o conhecimento gerado, identificamos que no habitat encosta, quando foram encontradas as duas primeiras espécies havia uma chance de 70% de ser encontrada a terceira espécie, em uma distância de até 20 metros em 2% das transações.

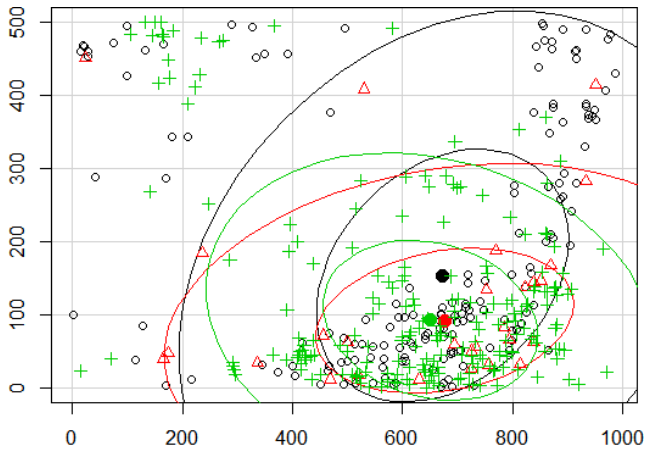


Figura 21. Distribuição espacial com coocorrência positiva entre *Ocotea whitei* (preta), *Virola multiflora* (vermelha) e *Xylopia macrantha* (verde)

A Figura 22 mostra a distribuição da espécie *Inga goldmanii*, *Socratea exorrhiza* e *Beilschmiedia pendula* ao longo do *plot*. A métrica *confiança* revela que quando a primeira das duas espécies foi encontrada, em 83% a terceira espécie também foi encontrada, com suas distribuições sendo destacadas com curvas suaves no gráfico de dispersão. O Apêndice D mostra os *scatter plots* para as correlações em grupos.

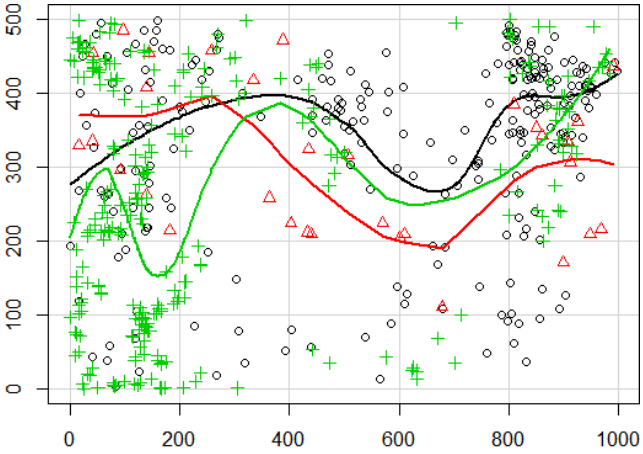


Figura 22. Distribuição espacial com coocorrência positiva entre *Inga goldmanii* (vermelha), *Socratea exorrhiza* (verde) e *Beilschmiedia pendula* (preta)

Na Tabela 10, todos os resultados são mostrados. Nela, o *lhs* (*left hand side*) é o antecedente da regras e o consequente o *rhs* (*right hand side*). Nos experimentos foram usados os seguintes parâmetros necessários para a aplicação ao do Apriori: *minlen* = 2, *maxlen* = 3, para verificar grupos de três espécies e suporte baseado no valor da espécie com a menor frequência para não eliminar espécie alguma da análise; e confiança mínima = 0.3.

A identificação de grupos pode permitir uma melhor caracterização do tipo de vegetação em uma determinada região. Os grupos também podem servir como indicadores da presença de outras espécies.

Tabela 10 – Regras e métricas obtidas com a distância máxima de até 20 metros

Legenda: *suporte* (sup), *coverage* (cove), *leverage* (leve), *confiança* (conf), *hyperconfidence* (hc), *hyperlift* (hlift), chi-quadrado (χ^2), *p-value* (p), *conviction* (conv), índice *Gini* (Gini) e *oddsratio* (odds). A tabela encontra-se dividida nos blocos I (correlação positiva em pares), II (correlação negativa em pares) e, III (correlação positiva em grupos de espécies)

	Fig	Antecedente (LHS) → Consequente (RHS)	sup	cove	leve	conf	hc	*lift*	hlift	χ^2	p	conv	Gini	odds
I	16	{*Virola multiflora*} → {*Xylopia macrantha*}	0.04	0.06	0.02	0.60	1	2.44	2.18	833.0	0	1.87	0.0157	5.17
	17	{*Terminalia oblonga*} → {*Ocotea whitei*}	0.03	0.07	0.02	0.50	1	2.32	2.07	657.3	0	1.56	0.0112	4.09
	23	{*Pourouma bicolor*} → {*Socratea exorrhiza*}	0.03	0.04	0.02	0.79	1	2.83	2.50	1066.1	0	3.36	0.0218	10.62
	24	{*Inga thibaudiana*} → {*Pterocarpus rohrii*}	0.02	0.07	0.01	0.30	1	2.55	2.19	509.2	0	1.26	0.0054	3.74
	25	{*Trophis caucana*} → {*Ocotea whitei*}	0.03	0.06	0.02	0.54	1	2.51	2.24	806.1	0	1.70	0.0138	4.87
	26	{*Virola multiflora*} → {*Ocotea whitei*}	0.03	0.06	0.02	0.52	1	2.44	2.17	704.5	0	1.65	0.0121	4.54
	27	{*Perebea xanthochyma*} → {*Socratea exorrhiza*}	0.03	0.05	0.02	0.67	1	2.43	2.17	811.3	0	2.21	0.0166	5.97
	28	{*Inga thibaudiana*} → {*Zanthoxylum ekmanii*}	0.03	0.07	0.02	0.45	1	2.27	2.02	626.5	0	1.45	0.0101	3.76
II	18	{*Socratea exorrhiza*} → {*Gustavia superba*}	0.08	0.28	-0.04	0.30	0	0.67	0.65	705.1	1	0.78	0.0178	0.41
	19	{*Triplaris cumingiana*} → {*Tetragastris panamensis*}	0.08	0.21	-0.04	0.36	0	0.68	0.66	595.0	1	0.74	0.0151	0.42
	29	{*Spondias mombin*} → {*Virola sebifera*}	0.02	0.07	-0.02	0.35	0	0.53	0.51	603.1	1	0.52	0.0137	0.25
	30	{*Adelia triloba*} → {*Tetragastris panamensis*}	0.04	0.11	-0.02	0.33	0	0.63	0.60	387.6	1	0.71	0.0098	0.40
	31	{*Attalea butyracea*} → {*Virola sebifera*}	0.03	0.07	-0.02	0.43	0	0.64	0.61	351.5	1	0.59	0.0080	0.35
	32	{*Unonopsis pittieri*} → {*Gustavia superba*}	0.09	0.31	-0.05	0.30	0	0.66	0.64	858.6	1	0.78	0.0217	0.39
	33	{*Platypodium elegans*} → {*Guarea guidonia*}	0.02	0.07	-0.01	0.33	0	0.66	0.62	177.1	1	0.74	0.0045	0.47
	34	{*Casearia sylvestris*} → {*Poulsenia armata*}	0.03	0.09	-0.01	0.31	0	0.67	0.64	174.8	1	0.78	0.0044	0.49
	35	{*Zanthoxylum ekmanii*} → {*Gustavia superba*}	0.06	0.20	-0.03	0.31	0	0.68	0.66	407.0	1	0.79	0.0103	0.47
	36	{*Lacmellea panamensis*} → {*Poulsenia armata*}	0.04	0.11	-0.02	0.33	0	0.70	0.67	192.7	1	0.79	0.0049	0.52

	Fig	Antecedente (LHS) → Consequente (RHS)	sup	cove	leve	conf	hc	*lift*	hlift	χ^2	p	conv	Gini	odds
III	21	{*Ocotea whitei, Virola multiflora*} → {*Xylopia macrantha*}	0.02	0.02	0.01	0.70	1	2.87	2.46	715.8	0	2.55	0.0135	7.93
	22	{*Inga goldmanii,Socratea exorrhiza*}→{*Beilschmiedia pendula*}	0.02	0.03	0.01	0.83	1	2.34	2.07	577.2	0	3.76	0.0134	9.36
	37	{*Unonopsis pittieri, Virola multiflora*} → {*Xylopia macrantha*}	0.02	0.03	0.02	0.85	1	3.46	2.94	1104.7	0	4.97	0.0208	18.92
	38	{*Terminalia oblonga, Xylopia macrantha*} → {*Ocotea whitei*}	0.02	0.03	0.02	0.72	1	3.36	2.86	1023.0	0	2.79	0.0175	10.43
	39	{*Poulsenia armata, Pourouma bicolor*} → {*Socratea exorrhiza*}	0.02	0.02	0.01	0.87	1	3.13	2.69	880.4	0	5.54	0.0180	18.76
	40	{*Drypetes standleyi, Virola multiflora*} → {*Xylopia macrantha*}	0.03	0.04	0.02	0.74	1	3.00	2.62	1025.8	0	2.86	0.0193	9.58
	41	{*Guatteria dumetorum, Virola multiflora*} → {*Xylopia macrantha*}	0.02	0.03	0.01	0.72	1	2.93	2.51	762.0	0	2.69	0.0144	8.60
	42	{*Pourouma bicolor, Quararibea asterolepis*} → {*Socratea exorrhiza*}	0.02	0.03	0.02	0.81	1	2.93	2.55	886.4	0	3.89	0.0181	12.44
	43	{*Pourouma bicolor, Virola sebifera*} → {*Socratea exorrhiza*}	0.03	0.03	0.02	0.80	1	2.90	2.52	877.9	0	3.70	0.0179	11.68
	44	{*Perebea xanthochyma, Poulsenia armata*} → {*Socratea exorrhiza*}	0.03	0.03	0.02	0.80	1	2.87	2.51	947.4	0	3.53	0.0193	11.12

5.3.4 Análise de Agrupamentos de Espécies

Após o levantamento das espécies que apresentaram coocorrências positivas, foi definida na metodologia proposta a aplicação da análise de agrupamentos para a identificação espacial das comunidades ou micro-habitats onde as coocorrências estão presentes no *plot*. A clusterização foi aplicada com o uso do algoritmo *Expectation Maximization* (EM), usando a estimativa de densidade baseada no modelo de mistura, como indicado na Seção 3.4.2.

A aplicação da análise de agrupamentos revelou que 4 *clusters* foram identificados envolvendo as espécies *Virola multiflora* e *Xylopia macranta*. A Figura 23 apresenta o gráfico de dispersão onde pode-se visualizar os *clusters* identificados pelo algoritmo.

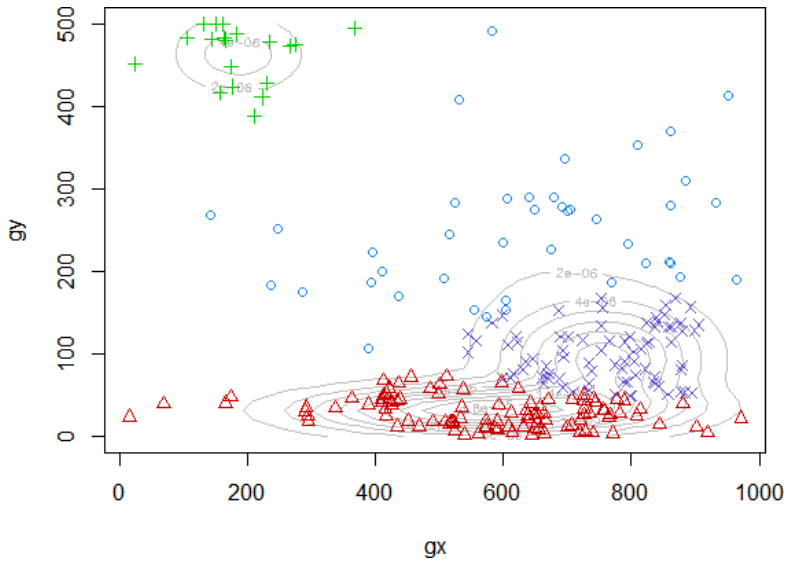


Figura 23. Coocorrência positiva entre *Virola multiflora* e *Xylopia macranta*

Na Figura 24 são destacadas as regiões onde as densidades mais altas são ressaltadas, permitindo a verificação dos habitats com as maiores distribuições para as duas espécies.

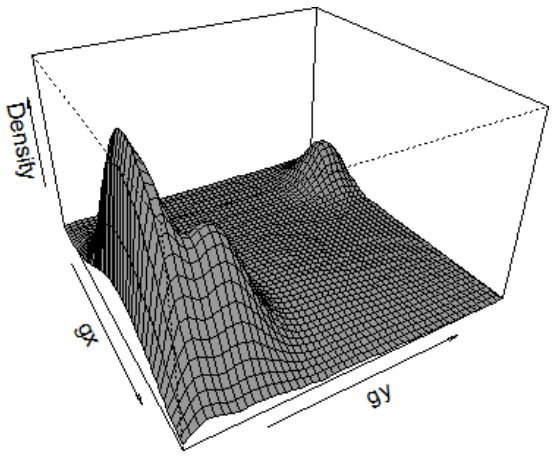


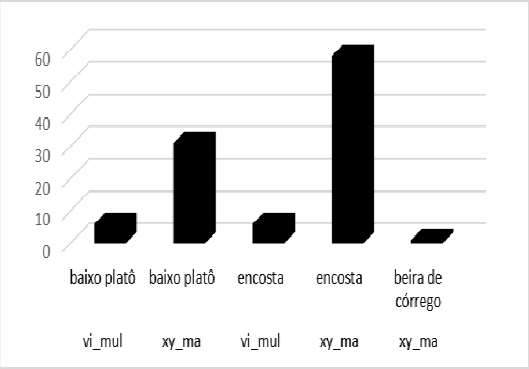
Figura 24. Plot 3D para coocorrência positiva entre *Virola multiflora* e *Xylopia macranta*

Cada *cluster* contém respectivamente: *cluster*_1 (azul) com 42 ocorrências, *cluster*_2 (vermelho) com 102, *cluster*_3 (verde) com 19 e o *cluster*_4 (roxo) com 80 ocorrências. Na Tabela 11 são apresentadas as quantidades de cada espécie nos *clusters*. As maiores quantidades foram observadas nos *clusters* 2 e 4.

Tabela 11 – Distribuição das espécies pelos *clusters* para *Virola multiflora* e *Xylopia macranta*

Espécie	***Cluster***	**Total**	**total_*cluster***
Virola multiflora	1	5	42
Xylopia macranta		37	
Virola multiflora	2	12	102
Xylopia macranta		90	
Virola multiflora	3	1	19
Xylopia macranta		18	
Virola multiflora	4	10	80
Xylopia macranta		70	

Nas Figuras 25 e 26 podem ser verificadas as distribuições das espécies pelos habitats, e assim observar que as espécies têm preferências pelos habitats do tipo encosta e baixo platô.



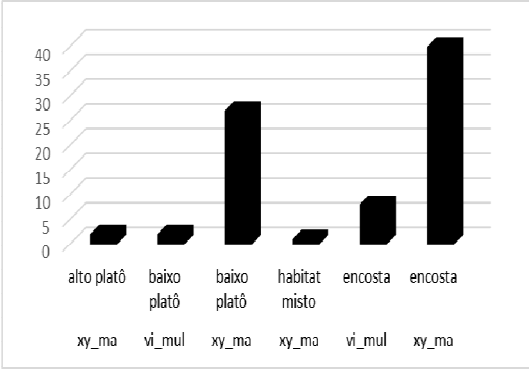


Fig. 25 **Fig. 26**

Figura 25. Distribuição de *Virola multiflora* e *Xylopia macranta* por habitats no *cluster*_2

Figura 26. Distribuição de *Virola multiflora* e *Xylopia macranta* por habitat no *cluster*_4

A aplicação da clusterização também foi realizada no estudo de caso com agrupamentos de espécies, como o exemplo apresentado na Figura 21 envolvendo as espécies *Ocotea whitei*, *Virola multiflora*, e *Xylopia macranta*. O resultado da análise de agrupamentos é apresentado nas Figuras 27 e 28, revelando o total de 6 *clusters*.

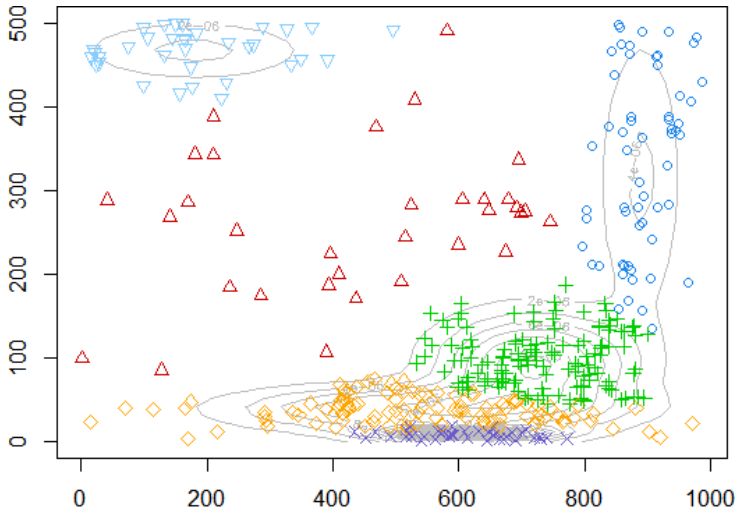


Figura 27. Agrupamentos para a coocorrência positiva entre *Ocotea whitei, Virola multiflora* e *Xylopia macranta*

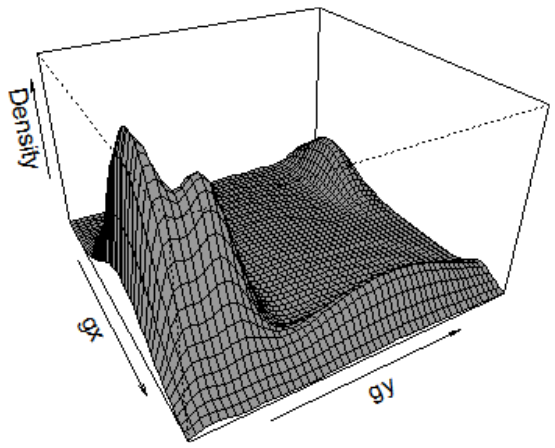


Figura 28. Gráfico 3D indicando as regiões com coocorrência positiva entre *Ocotea whitei, Virola multiflora* e *Xylopia macrantha*

Na Tabela 12 abaixo são apresentados os valores da distribuição da quantidade de espécimes em cada *cluster*. Os dois *clusters* que apresentam as maiores quantidades são os de número 3 e 5, sendo que os habitats observados no *plot* são os de tipo baixo platô e encosta.

Tabela 12 – Distribuição de *Ocotea whitei, Virola multiflora* e *Xylopia macranta* nos *clusters*

Espécie	*Cluster*	Total	total_*cluster*
Ocotea whitei	1	43	59
Virola multiflora		3	
Xylopia macrantha		13	
Ocotea whitei	2	7	33
Virola multiflora		2	
Xylopia macrantha		24	
Ocotea whitei	3	50	**135**
Virola multiflora		11	
Xylopia macrantha		74	
Ocotea whitei	4	12	38
Virola multiflora		2	
Xylopia macrantha		24	
Ocotea whitei	5	33	**105**
Virola multiflora		9	
Xylopia macrantha		63	
Ocotea whitei	6	17	35
Virola multiflora		1	
Xylopia macrantha		17	

Capítulo 6 – Conclusão

6.1 Contribuições

O volume de dados de coleções científicas tem crescido e o número de pesquisadores necessários para a realização de estudos nessas bases de dados não tem acompanhado tal crescimento. Essa demanda pela análise de grandes bancos de dados tem provocado uma busca por novos métodos, como tem ocorrido em outras áreas. A botânica pode então ser incorporada nesse contexto, apresentando a demanda por ferramentas para auxiliar na extração de conhecimento de forma automática, apoiando o trabalho dos pesquisadores. Na busca por soluções, uma abordagem multidisciplinar pode ser o caminho para a solução do problema, o que naturalmente provoca a formação de equipes com abrangência de conhecimento em diferentes áreas, sendo essas equipes compostas por profissionais com perfis específicos para o gerenciamento de bancos de dados, análises estatísticas e a interpretação dos padrões biológicos obtidos.

Neste trabalho foi proposta uma metodologia para a análise de coocorrências de espécies, inovando quanto ao uso dos algoritmos da análise de associação e de agrupamentos da mineração de dados como componentes da proposta. Considerou-se a aplicação da mineração de dados na análise de dados ecológicos, tendo em vista que se mostrou eficiente em diversos outros tipos de aplicações.

A metodologia foi utilizada para a obtenção de conhecimento relativo à indicação da coocorrências de espécies, classificando segundo o tipo de habitat. Os resultados alcançados por meio dos estudos de casos comprovaram que tal conhecimento pode ser extraído, observando-se a correlação positiva e negativa entre espécies. Para que esses resultados fossem alcançados, uma metodologia foi proposta e testada em uma base contendo dados mapeados em parcelas.

Um método foi especificado com uma abordagem baseada na análise de regras de associação examinando a distância (em metros) entre os espécimes, como fator determinador para a criação das transações a serem avaliadas. O conjunto de regras obtido foi avaliado com a aplicação de diversas métricas sobre um banco de dados de um projeto de monitoramento de uma complexa floresta tropical, amplamente conhecido, como é o *plot* de *Barro Colorado Island*, facilitando portanto a avaliação dos resultados.

Como principais contribuições do trabalho realizado, destacamos:

- A comprovação de que é possível obter conhecimento por meio da aplicação da análise de associação tradicional para os dados de levantamento florísticos, o que foi provado pelos estudos de casos e com os resultados satisfazendo as métricas utilizadas;

- Que a aplicação da análise de agrupamentos facilita a identificação das regiões onde as coocorrências existem;

- Que uma abordagem permite a análise simultânea de ocorrências de espécies entre pares e grupos de espécies, tanto positiva quanto negativamente correlacionadas;

- Que é possível realizar a análise de coocorrências sem a necessidade de uma pré-seleção inicial de espécies para que os testes possam ser mais práticos. O que foi observado em diversos trabalhos, mesmo os mais recentes, acarretando em uma grande dificuldade no desenvolvimento das pesquisas envolvendo uma quantidade maior de espécies e, o que normalmente ocorre em regiões tropicais;

- Outra contribuição importante do trabalho foi a avaliação do conjunto de métricas da análise de associação para a seleção daquelas que permitiram obter as melhores

regras, no sentido de alcançar os objetivos pré-estabelecidos para a pesquisa quanto à identificação de coocorrências;

- Alguns dos resultados obtidos, direta ou indiretamente, ao longo do desenvolvimento desta tese foram registrados na forma de trabalhos apresentados:

    - BRANDAO, S. N., SILVA, W. N., SILVA, L. A. E., FAGUNDES, V., MELLO, C. E. M., ZIMBRAO, G.; SOUZA, J. M. Analysis and Visualization of the Geographical Distribution of Atlantic Forest Bromeliads Species. **In: IEEE Symposium on Computational Intelligence and Data Mining**, pp. 375 - 380, Nashville, TN, USA, mar. 2009;

    - SILVA, L. A. E, BARROS, R. O., DALCIN, E. C., ZIMBRÃO, G. S., SOUZA, J. M. Abordagem Colaborativa para a Melhoria da Qualidade de Dados em Bases de Dados Botânicas, **In: II Workshop de Computação Aplicada à Gestão do Meio Ambiente e Recursos Naturais, Belo Horizonte. XXX Congresso da Sociedade Brasileira de Computação - Computação Verde: Desafios Científicos e Tecnológicos**. Belo Horizonte: 2010;

    - SILVA, L. A. E., DALCIN, E. C., ZIMBRÃO, G., SOUZA, J. M., SAMPAIO, J. O. Gestão do Conhecimento Taxonômico Aplicado na Conservação da Flora Brasileira, **In: IV e-Science Workshop, XXX Congresso da Sociedade Brasileira de Computação - Computação Verde: Desafios Científicos e Tecnológicos**. Belo Horizonte: 2010;

    - SILVA, L. A. E., ZAN, K. M., VAZ, M., S, ZIMBRÃO, G., CARMO, F. B., LIMA, D. A. Aplicação da Abordagem Colaborativa em Herbários Virtuais,

In: VIII Simpósio Brasileiro de Sistemas Colaborativos. VIII Simpósio Brasileiro de Sistemas Colaborativos 2011. Paraty: 2011;

- DALCIN, E. C., SILVA, L. A. E., ZIMBRÃO, G., SOUZA, J. M., CABANILLAS, C. C., MONTEIRO, V. F., LOURES, M. G. S. M. Data Quality Assessment at the Rio de Janeiro Botanical Garden Herbarium Database and Considerations for Data Quality Improvement, **In: 8th International Conference on Ecological Informatics**, Brasília: 2012;

Como ponto de discussão nesta pesquisa, apontamos a sugestão de inclusão da mineração de dados no processo de geração de conhecimento na Ecologia, ressaltando que diferentes áreas da pesquisa que começaram a fazer uso dessas possibilidades obtiveram resultados interessantes, como citado no Capítulo 3. Sugerimos a inclusão da mineração de dados na etapa anterior a aplicação da Análise de Padrões de Pontos (WIEGAND, MOLONEY, 2004), que revelam *insights* sobre as espécies que têm algum tipo de coocorrência, evitando com isso a necessidade do testes com todos os pares de espécies.

Outro importante ponto de discussão está relacionado com a necessidade de um estudo mais aprofundado da definição ou seleção de outras métricas para a avaliação das regras obtidas com a análise de associação, tendo em vista especialmente o propósito deste tipo de aplicação.

6.2 Trabalhos Futuros

Ao longo do desenvolvimento da pesquisa diversas possíveis aplicações da metodologia proposta foram surgindo. Entre elas consideramos que a abordagem pode ser especialmente útil para espécies ainda pouco mapeadas, tendo em vista as

dificuldades de coleta de espécies em algumas regiões. A afirmativa considera que os resultados obtidos com agrupamentos de espécies, onde as constam as espécies com poucas coletas, pode se utilizado como reforço de que determinadas espécies seriam indicadoras de habitats, ou mesmo de micro-habitats, dada a escala em que essa abordagem atua.

Outro trabalho possível é o desenvolvimento de um *workflow* científico para agilizar e permitir o compartilhamento das pesquisas. O *workflow* pode possibilitar também o monitoramento constante das mudanças ocorridas tanto com os espécimes quanto, quanto nos habitats na floresta.

Outras métricas também podem ser estudadas na avaliação dos resultados sobre as regras geradas, observando-se as características muito específicas das pesquisas envolvendo dados ecológicos.

Os resultados promissores obtidos sobre pares e grupos de espécies motivam outras investigações relacionadas com a inclusão de outras variáveis ambientais na análise de agrupamentos, como por exemplo, dados com tipos de solo e nível de insolação.

Independentemente das formas de aplicação, estudos são necessários para determinar o uso do método no suporte à prova da existência de hipóteses envolvendo comunidades de espécies, tal como a competição. Neste caso, é necessário o uso de variáveis bióticas. Tudo isso sendo possível pela flexibilidade do método, que pode ser usado em análises com diferentes escalas.

Referências Bibliográficas

AGRAWAL, R. Fast Algorithms for Mining Association Rules. In: **Proceedings VLDB Endowment**, pp. 487-499, Santiago, Chile, Set. 1994.

AGRAWAL, R.; GEHRKE, J.; GUNOPULOS, D.; RAGHAVAN, P. Automatic subspace clustering of high dimensional data for data mining applications. In: **Proceedings ACM SIGMOD International Conference on Management of Data**, pp. 94-105, Seattle, Washington, Jun. 1998.

AGRAWAL, R., IMIELINSKI, T., SWAMI, A. Mining association rules between sets of items in large databases. In: **Proceedings ACM SIGMOD International Conference on Management of Data**, pp. 207-216, Washington, D. C., Mai. 1993.

ANKERST, M., BREUNIG, M., KRIEGEL, H.-P., SANDER., J. OPTICS: Ordering points to identify the clustering structure. In: **Proceedings ACM SIGMOD International Conference on Management of Data**, pp. 49–60, Philadelphia, Pennsylvania, Jun. 1999.

ARAÚJO, M. B., BASELGA, A. "Do community-level models describe community variation effectively?", **Biogeography**, v. 37, n. 10, pp. 1842–1850, Jun. 2010.

AUSTIN, M. "Species distribution models and ecological theory: A critical assessment and some possible new approaches", **Ecological Modelling**, v. 200, n. 1, pp. 1–19, Jan. 2007.

BAILEY, R. G. "Identifying ecoregion boundaries", **Environmental Management**, v. 7, n. 4, pp. 365–373, Jan. 1983.

BARROS, A. J. S., LEHFELD, N. A. S. **Fundamentos de Metodologia Científica**. 3. ed. São Paulo: Editora Pearson Prentice Hall, 2007.

BERKHIN, P. Survey of Clustering Data Mining Techniques. In: KOGAN, J., NICHOLAS, C., TEBOULLE, M. (Ed.). **Grouping Multidimensional Data**. Berlin: Springer Berlin Heidelberg, 2006. pp. 25–71.

BOCARD, D., GILLET, F., LEGENDRE, P. **Numerical Ecology with R**. New York: Springer, 2011.

BORGES, D. F. DE M. *Padrões de variação na riqueza de espécies em gradientes altitudinais : uma revisão multi-taxonômica.* Brasília: Dissertação M.Sc., Universidade de Brasília, 2011.

BRANDAO, S. N., SILVA, W. N., SILVA, L. A. E., FAGUNDES, V., MELLO, C. E. M., ZIMBRAO, G., SOUZA, J. M. Analysis and Visualization of the Geographical Distribution of Atlantic Forest Bromeliads Species. In: **IEEE Symposium on Computational Intelligence and Data Mining**, pp. 375 - 380, Nashville, TN, USA, Mar. 2009.

BREESE, J., HECKERMAN, D., KADIE, C. Empirical analysis of predictive algorithms for collaborative filtering. In: **Proceedinds Conference Uncertainty in Artificial Intelligence**, pp. 43–52, Madison, WI, Jul. 1998.

BROOKS, T. M. *et al.* "Habitat Loss and Extinction in the Hotspots of Biodiversity", **Conservation Biology**, v. 16, n. 4, p. 909–923, Ago. 2002.

CÁCERES, M. D., LEGENDRE, P., MORETTI, M. "Improving indicator species analysis by combining groups of sites", **Oikos**, v. 119, pp. 1674–1684, Out. 2010.

CAPELO, J. **Conceitos e métodos da fitossociologia - Formulação contemporânea e métodos numéricos de análise da vegetação**. Lisboa: Estação Florestal Nacional, Sociedade Portuguesa de Ciências Florestais, 2003.

CARVALHO, L. A. V. **Datamining – A Mineração de Dados no Marketing, Medicina, Economia, Engenharia e Administração**. Rio de Janeiro: Editora Ciência Moderna, 2005.

CIANCIARUSO, M. V., SILVA, I. A., BATALHA, M. A. "Diversidades filogenética e funcional novas abordagens para a Ecologia de comunidades", **Biota Neotropica**, v. 9, n. 3, pp. 93–103, Jul. 2009.

CIL, I. “Consumption universes based supermarket layout through association rule mining and multidimensional scaling”, **Expert Systems with Applications**, v. 39, n. 10, pp. 8611–8625, Ago. 2012.

CLARIDGE, M. F. “Introducing systematics Agenda 2000”, **Biodiversity & Conservation**, v. 4, n. 5, pp. 451–454, Jul. 1995.

CONDIT, R. **Tropical Forest Census Plots**. Berlin, Germany, and Georgetown, Texas: Springer-Verlag and R. G. Landes Company, 1998.

CUTLER, D. R. *et al.* “Random forests for classification in ecology”, **Ecology**, v. 192, n. 11, pp. 2783–2792, Nov. 2007.

DATE, C. J. **Introdução a Sistemas de Bancos de Dados**. 7. ed. Rio de Janeiro: Editora Campus, 2000.

DETTO, M., MULLER-LANDAU, H. C. “Fitting ecological process models to spatial patterns using scalewise variances and moment equations”, **The American naturalist**, v. 181, n. 4, pp. E68–82, Abr. 2013.

DINIZ, J. A. F. *et al.* “Null models and spatial patterns of species richness in South American birds of prey”, **Ecology Letters**, v. 5, n. 1, pp. 47–55, Jan. 2002.

DIXON, P. “VEGAN, a package of R functions for community ecology”, **Journal of Vegetation Science**, v. 14, n. 6, pp. 927–930, Dez. 2003.

DREW, L. W. “Are We Losing the Science of Taxonomy?: As need grows, numbers and training are failing to keep up”, **BioScience**, v. 61, n. 12, pp. 942–946, Dez. 2011.

ELITH, J. *et al.* “Novel methods improve prediction of species’ distributions from occurrence data”, **Ecography**, v. 29, n. 2, pp. 129–151, mar. 2006.

ELITH, J., LEATHWICK, J. R. “Species Distribution Models: Ecological Explanation and Prediction Across Space and Time”, **Annual Review of Ecology, Evolution, and Systematics**, v. 40, n. 1, pp. 677–697, Dez. 2009.

ELMASRI, R., NAVATHE, S. B. **Sistemas de banco de dados**. 6. ed. São Paulo: Editora Pearson Addison Wesley, 2011.

ENTSMINGER, G. **EcoSim Professional: Null modeling software for ecologists**MontroseAcquired Intelligence Inc., Kesey-Bear, & Pinyon Publishing, 2012. Disponível em: <http://www.garyentsminger.com/ecosim/index.htm>. Acesso em: 4 nov. 2013.

ESTER, M., KRIEGEL, H., SANDER, J. Algorithms and Applications for Spatial Data Mining. In: MILLER, H. V., HAN, J. (Ed.). **Geographic Data Mining and Knowledge Discovery**. 1. ed. London: Taylor and Francis, 2001. p. 160–187.

ESTER, M., KRIEGEL, H. -P., SANDER, J., XU, X. A density-based algorithm for discovering clusters in large spatial databases. In: **Second International Conference on Knowledge Discovery and Data Mining**, pp. 226–231, Portland, OR, Ago. 1996.

FAYYAD, U., PIATETSKY-SHAPIRO, G., SMYTH, P. "Knowledge Discovery and Data Mining: Towards a Unifying Framework", In: **Proceedings of the Second International Conference on Knowledge Discovery and Data Mining**, pp. 82-88, Portland, OR, Ago. 1996.

FERRIER, S., GUISAN, A. "Spatial modelling of biodiversity at the community level", **Applied Ecology**, v. 43, n. 3, pp. 393–404, Jun. 2006.

FISHER, D. Improving inference through conceptual clustering. In: **Proceedings of the sixth National conference on Artificial intelligence**, pp. 461-465, Seattle, WA, Jul. 1987.

FORZZA, R.C., LEITMAN, P.M., COSTA, A., CARVALHO JUNIOR, A.A., PEIXOTO, A.L., WALTER, B.M.T., BICUDO, C., MOURA, C.W.N., ZAPPI, D., COSTA, D.P., LLERAS, E., MARTINELLI, G., LIMA, H.C., PRADO, J., STEHMANN, J.R., BAUMGRATZ, J.F.A., PIRANI, J.R., SYLVESTRE, L.S., MAIA, L.C., LOHMANN, L.G., QUEIROZ, L.P., SILVEIRA, M., COELHO, M.N., MAMEDE, M.M.H., BASTOS, M.N.C., MORIM, M.P., BARBOSA, M.R., MENEZES, M., HOPKINS, M., SECCO, R., CAVALCANTI, T. &

SOUZA, V.C. 2010b. Lista de Espécies da Flora do Brasil. Jardim Botânico do Rio de Janeiro, Rio de Janeiro. Disponível em: <http://floradobrasil.jbrj.gov.br>

FRALEY, C., RAFTERY, A. E. MCLUST Version 3 for R: Normal Mixture Modeling and Model-Based ClusteringIn: Report No. 504. Department of Statistics, University of Washington, Washington.

FULLER, M. M., ENQUIST, B. J. "Accounting for spatial autocorrelation in null models of tree species association", **Ecography**, v. 35, n. 6, pp. 510–518, Jun. 2011.

GERING, J. C., CRIST, T. O. "The alpha-beta-regional relationship: providing new insights into local-regional patterns of species richness and scale dependence of diversity components", **Ecology Letters**, v. 5, n. 3, pp. 433–444, Mai. 2002.

GIOVANELLI, J. G. R. *et al.* "Modeling a spatially restricted distribution in the Neotropics: How the size of calibration area affects the performance of five presence-only methods", **Ecological Modelling**, v. 221, n. 2, pp. 215–224, Jan. 2010.

GOMES, A. S. **Análise de Dados Ecológicos**. Rio de Janeiro: Instituto de Biologia, Universidade Federal Fluminense, 2004.

GOTELLI, N. J. "Null Model Analysis of Species Co-Occurrence Patterns", **Ecology**, v. 81, n. 9, pp. 2606, Set. 2000.

GOTELLI, N. J. *et al.* "Patterns and causes of species richness: a general simulation model for macroecology", **Ecology letters**, v. 12, n. 9, pp. 873–86, Set. 2009.

GOTELLI, N. J., MCCABE, D. J., AUG, N. "Species Co-Occurrence : A Meta-Analysis of J . M . Diamond's Assembly Rules Model", **Ecology**, v. 83, n. 8, pp. 2091–2096, Ago. 2007.

GRAHAM, C. H. *et al.* "The influence of spatial errors in species occurrence data used in distribution models", **Journal of Applied Ecology**, v. 45, n. 1, pp. 239–247, 3 Ago. 2007.

GUHA, S., RASTOGI, R., SHIM, K. Cure: An efficient clustering algorithm for large databases. In: **Proceedings ACM SIGMOD International Conference on Management of Data**, pp. 73–84, Seattle, WA, Jun. 1998.

HAHSLER, M., BUCHTA, C., GRÜN, B., HORNIK, K. **arules: Mining Association Rules and Frequent Itemsets**. Disponível em: <http://cran.r-project.org/package=arules>. Acesso em: 2 maio. 2013.

HAHSLER, M., GRÜN, B., HORNIK, K. "arules – A Computational Environment for Mining Association Rules and Frequent Item Sets", **Journal of Statistical Software**, v. 14, n. 15, Set. 2005.

HAN, J. *et al.* "Emerging Scientific Applications in Data Mining", **Communications of ACM**, v. 45, n. 8, pp. 54–58, Ago. 2002.

HAN, J., KAMBER, M., PEI, J. **Data Mining: Concepts and Techniques**. 3. ed. San Francisco: Morgan Kaufmann, 2011.

HAN, J., KAMBER, M. & TUNG, K. H. Spatial Clustering Methods in Data Mining: A Survey. *In*: **Geographic Data Mining and Knowledge Discovery**. 1. ed. London: Taylor and Francis, 2001. p. 1–29.

HAN, J., PEI, P., YIN, Y. Mining frequent patterns without candidate generation. In: **Proceedings ACM SIGMOD International Conference on Management of Data**, pp. 1-12, Seattle, WA, Mai. 2000.

HARMS, K. E., CONDIT, R., HUBBELL, S. P., FOSTER, R. B. "Habitat associations of trees and shrubs in a 50-ha neotropical forest plot", **Journal of Ecology**, v. 89, n. 6, pp. 947–959, Dez. 2001.

HIJMANS, R. J., ELITH, J. **Species distribution modeling with R**. Disponível em: <http://cran.r-project.org/web/packages/dismo/vignettes/sdm.pdf>. Acesso em: 4 nov. 2013.

HILL, M. **TWINSPAN – a FORTRAN program for arranging multivariate data in an ordered two-way table by classification of the individuals and attributes**, New York, NY, USA, Cornel University, 1979.

HOCHACHKA, W. M. *et al.* “Data-Mining Discovery of Pattern and Process in Ecological Systems”, **The Journal of Wildlife Management**, v. 71, n. 7, pp. 2427, Set. 2007.

HOPKINS, G. W., FRECKLETON, R. P. “Declines in the numbers of amateur and professional taxonomists: implications for conservation”, **Animal Conservation**, v. 5, n. 03, pp. 245–49, Ago. 2002.

HOWE, D., COSTANZO, M., FEY, P., GOJOBORI, T., HANNICK, L., HIDE, W., HILL, DP., KANIA, R., SCHAEFFER, M., ST PIERRE, S., TWIGGER, S., WHITE, O., RHEE, S. “Big data: The future of biocuration”, **Nature**, v. 455, pp. 47–50, Set. 2008.

HUBBELL, S.P., CONDIT, R., FOSTER, R. B. **Barro Colorado Forest Census Plot Data**. Disponível em: <http://ctfs.si/edu/datasets/bci>. Acesso em: 4 nov. 2013.

HUBBELL, S.P., R.B. FOSTER, S.T. O’BRIEN, K.E. HARMS, R. CONDIT, B. WECHSLER, S. J. W., LAO, AND S. L. DE. “Light gap disturbances, recruitment limitation, and tree diversity in a neotropical forest”, **Science**, v. 283, pp. 554–557, Jan. 1999.

INMAN-NARAHARI, F., GIARDINA, C., OSTERTAG, R., CORDELL, S., SACK, L. “Digital data collection in forest dynamics plots”, **Methods in Ecology and Evolution**, v. 1, n. 3, pp. 274–279, set. 2010.

JACCARD, P. Distribution de la flore alpine dans le bassin des Dranses et dans quelques régions voisines. **Bulletin de la Société Vaudoise des Sciences Naturelles**, v. 37, pp. 241–272, Jan. 1901.

JORGE, A. M., AZEVEDO, P. J. An Experiment with Association Rules and Classification: Post-Bagging and Conviction. In: **Proceedings of the 8th International Conference on Discovery Science DS 2005**, pp. 137-149, Singapore, out. 2005.

KARYPIS, G., HAN, E., KUMAR, V. “CHAMELEON: A Hierarchical Clustering Algorithm Using Dynamic Modeling”, **IEEE Computer**, v. 32, n. 8, pp. 68–75, Ago. 1999.

KAUFMAN, L., ROUSSEEUW, P. J. **Finding Groups inData: An Introduction to Cluster Analysis**. New York: John Wiley & Sons, 1990.

KENETT, R., SALINI, S. Relative Linkage Disequilibrium: A New Measure for Association Rules. In: **8th industrial Conference on Advances in Data Mining: Medical Applications, E-Commerce, Marketing, and Theoretical Aspects**, pp. 189-199, Pisa, Italy, Dez. 2008.

KENT, R., CARMEL, Y. “Evaluation of five clustering algorithms for biodiversity surrogates”, **Ecological Indicators**, v. 11, n. 3, pp. 896–901, Mai. 2011.

KERSCHBERG, L., ROSS, J., PIATETSKY-SHAPIRO, G. Knowledge Discovery in Real Databases: A Report on the IJCAI-89 Workshop. **AI Magazine**, v. 11, n. 5, 1991.

KOHONEN, T. “The Self-Organizing Map”, **Proceedings of the IEEE**, v. 78, n. 9, pp. 1464–1480, Set. 1990.

KOONCE, D. A., TSAI, S. C. “Using data mining to find patterns in genetic algorithm solutions to a job shop schedule”, **Computers & Industrial Engineering**, v. 38, n. 3, pp. 361–374, Out. 2000.

KUMAR, J. *et al.* “Parallel k-Means Clustering for Quantitative Ecoregion Delineation Using Large Data Sets”, **Procedia Computer Science**, v. 4, pp. 1602–1611, Jan. 2011.

LAN, G., GETZIN, S., WIEGAND, T., HU, Y., XIE, G., ZHU, H., CAO, M. “Spatial Distribution and Interspecific Associations of Tree Species in a Tropical Seasonal Rain Forest of China”, **PLoS ONE**, v. 7, n. 9, Set. 2012.

LATTIN, J., CARROLL, J. D., GREEN, P. E. **Análise de Dados Multivariados**. São Paulo: Editora Cengage Lerning, 2011.

LAURITZEN, S. L. “The EM algorithm for graphical association models with missing data”, **Computational Statistics and Data Analysis**, v. 19, n. 2, pp. 191–201, Fev. 1995.

LAZCORRETA, E., BOTELLA, F., FERNÁNDEZ-CABALLERO, A. “Towards personalized recommendation by two-step modified Apriori data mining algorithm”, **Expert Systems with Applications**, v. 35, n. 3, pp. 1422–1429, Out. 2008.

LEGENDRE, P., FORTIN, M. Spatial pattern and ecological analysis. **Vegetatio**, v. 80, pp. 107–138, Jan. 1989.

LEGENDRE, P., LEGENDRE, L. **Numerical Ecology**. 2. ed. Amsterdam: Elsevier Science B. V., 1998.

LIAO, S., CHEN, Y., DENG, M. “Mining customer knowledge for tourism new product development and customer relationship management”, **Expert Systems with Applications**, v. 37, n. 6, pp. 4212–4223, Jun. 2010.

LIAO, S., CHU, P., HSIAO, P. “Data mining techniques and applications – A decade review from 2000 to 2011”, **Expert Systems with Applications**, v. 39, n. 12, pp. 11303–11311, Set. 2012.

LIN, Y., HUANG, Y., LEU, J. J. “A new logic correlation rule for HIV-1 protease mutation”, **Expert Systems with Applications**, v. 38, n. 5, pp. 5448–5455, Mai. 2011.

LIU, B., HSU, W., MA, Y. Pruning and Summarizing the Discovered Associations. In: **Proceedings of the fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**, pp. 125-134, San Diego, CA, 1999.

LIU, G., ZHANG, H., WONG, L. Controlling False Positives in Association Rule Mining. **In: Proceedings of the VLDB Endowment**, pp. 145-156, Seattle, WA, Ago. 2011.

LORENA, A. C. *et al.* “Comparing machine learning classifiers in potential distribution modelling”, **Expert Systems with Applications**, v. 38, n. 5, pp. 5268–5275, Mai. 2011.

LUO, Z. R. *et al.* “Spatial associations of tree species in a subtropical evergreen broad-leaved forest”, **Journal of Plant Ecology**, Jan. 2012.

MACHADO, R. B., M. B. RAMOS NETO, P. G. P. PEREIRA, E. F. CALDAS, D. A. GONÇALVES, N. S. SANTOS, K. TABOR, M. S. **Estimativas de perda da área do Cerrado brasileiro**. Disponível em: <http://www.conservation.org.br/ arquivos/ RelatDesmatamCerrado.pdf>. Acesso em: 4 nov. 2013.

MACQUEEN, J. Some methods for classification and analysis of multivariate observations. In: **Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability**, pp. 281-297, Berkeley, CA, Jul. 1967.

MAGURRAN, A. E. **Medindo a Diversidade Biológica**. Curitiba: Editora UFRR, 2011.

MILLS, R. T. *et al.* "Cluster Analysis-Based Approaches for Geospatiotemporal Data Mining of Massive Data Sets for Identification of Forest Threats”, **Procedia Computer Science**, v. 4, pp. 1612–1621, Jan. 2011.

NG, R., HAN, J. Efficient and effective clustering method for spatial data mining. In: **Proceedings International Conference Very Large Data Bases**, pp. 144-155, Santiago, Chile, Set. 1994.

NGAI, E. W. T., XIU, L., CHAU, D. C. K. “Application of data mining techniques in customer relationship management: A literature review and classification”, **Expert Systems with Applications**, v. 36, n. 2, pp. 2592–2602, Mar. 2009.

ODUM, E. P., BARRETT, G. W. **Fundamentos de Ecologia**. 1. ed. São Paulo: Editora Thomson Pioneira, 2007.

OKSANEN, J. **Multivariate Analysis of Ecological Communities in R : vegan tutorial**. Disponível em: <http://cc.oulu.fi/~jarioksa/opetus/metodi/ vegantutor.pdf>. Acesso em: 11 abr. 2013.

OLSON, D. M. *et al.* “Terrestrial Ecoregions of the World : A New Map of Life on Earth”, **BioScience**, v. 51, n. 11, pp. 933–938, Nov. 2001.

PEARSON, R. G. **Species’ Distribution Modeling for Conservation Educators and PractitionersNative Plants**American Museum of Natural History, , 2007. Disponível em: <http://biodiversityinformatics.amnh.org/files/SpeciesDistModelingSYN_1-16-08.pdf>. Acesso em: 4 nov. 2013

PEIXOTO, A. L., MORIM, M. P. “Botânicas : documentação da biodiversidade brasileira”, **Flora**, v. 55, n. 3, pp. 21–24, Jul. 2002.

PERRY, G. L. W., MILLER, B. P., ENRIGHT, N. J. “A comparison of methods for the statistical analysis of spatial point patterns in plant ecology”, **Plant Ecology**, v. 187, n. 1, pp. 59–82, Mar. 2006.

PINO-MEJÍAS, R., CUBILES-DE-LA-VEGA, M. D., ANAYA-ROMERO, M., PASCUAL-ACOSTA, A., JORDÁN-LÓPEZ, A., BELLINFANTE-CROCCI, N. “Predicting the potential habitat of oaks with data mining models and the R system”, **Environmental Modelling & Software**, v. 25, n. 7, pp. 826–836, Jul. 2010.

PROTAZIO, J. M. B. *Spatial Pattern Analysis Applied to Plant Ecology*. Bremen, Germany: Ph.D Thesis, Center of Tropical Marine Ecology, University of Bremen, 2007.

R DEVELOPMENT CORE TEAM. **R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing Vienna Austria**, 2011. Disponível em: <http://www.r-project.org>. Acesso em: 4 nov. 2013

RAMAKRISHMAN, R., GEHRKE, J. **Sistemas de Gerenciamento de Banco de Dados**. 3. ed. São Paulo: Editora McGraw-Hill, 2008.

RANGEL, T. F., DINIZ-FILHO, J. A. F., BINI, L. M. “SAM: a comprehensive application for Spatial Analysis in Macroecology”, **Ecography**, v. 33, n. 1, pp. 46–50, Fev. 2010.

RAYMOND, B. *et al.* “Data and Scientific Data Mining”, **Arctic, Antarctic, and Alpine Research**, v. 37, n. 3, pp. 348–357, Ago. 2005.

RIPLEY, B. D. Spatial point pattern analysis in ecology. *In*: LEGENDRE, P., LEGENDRE, L. (Eds.). **Developments in Numerical Ecology**. Berlin: Springer-Verlag, 1987. p. 407–429.

SANJEREHEI, M. M. “Determination of an appropriate quadrat size and shape for detecting association between plant species”, **Ecological Modelling**, v. 222, n. 10, pp. 1790–1792, Mai. 2011.

SCHNITZER, S. A *et al.* “Liana abundance, diversity, and distribution on Barro Colorado Island, Panama”, **PloS one**, v. 7, n. 12, pp. e52114, Dez. 2012.

SHEIKHOLESLAMI, G., CHATTERJEE, S., ZHANG, A. Wavecluster: A multi-resolution clustering approach for very large spatial databases. In: **Proceedings of the 24th International Conference Very Large Data Bases**, pp. 428-439, New York, USA, 1998.

SILBERSCHATZ, A., KORTH, H. F., SUDARSHAN, S. **Sistema de Banco de Dados**. 3. ed. São Paulo: Makron Books, 1999.

SILVERSTEIN, C., BRIN, S., MOTWANI, R. “Beyond Market Baskets: Generalizing Association Rules to Dependence Rules”, **Data Mining and Knowledge Discovery**, v. 2, n. 1, pp. 39–68, Jan. 1998.

SIQUEIRA, M. F. DE. *Uso de modelagem de nicho fundamental na avaliação do padrão de distribuição geográfica de espécies vegetais.* São Carlos, São Paulo: Tese D.Sc., Escola de Engenharia de São Carlos da Universidade de São Paulo, 2005.

SNEATH, P. H., SOKAL, R. R. **Numerical taxonomy: The principles and practice of numerical classification**. San Francisco: W.H. Freeman, 1973. p. 573

SOKAL, R. R., MICHENER, C. D. **A statistical method for evaluating systematic relationships**. Kansas: University of Kansas, 1958.

STEINMANN, K., LINDERB, H. P., ZIMMERMANN, N. E. “Modelling plant species richness using functional groups”, **Ecological Modelling**, v. 220, n. 7, pp. 962–967, Abr. 2009.

SU, F. *et al.* “A data-mining approach to determine the spatio-temporal relationship between environmental factors and fish distribution”, **Ecological Modelling**, v. 174, n. 4, pp. 421–431, Jun. 2004.

SWENSON, N. G. “The assembly of tropical tree communities – the advances and shortcomings of phylogenetic and functional trait analyses”, **Ecography**, v. 36, n. 3, pp. 264–276, Mar. 2013.

TAN, P.-N., KUMAR, V., SRIVASTAVA, J. Selecting the Right Interestingness Measure for Association Patterns. In: **Proceedings of the eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**, pp. 32–41, Alberta, Canada, Jul. 2002.

TAN, P. N., STEINBACH, M., KUMAR, V. **Introdução ao Data Mining**. 1st. ed. Rio de Janeiro: Editora Ciência Moderna, 2009.

TELENIUS, A. “Biodiversity information goes public: GBIF at your service”, **Nordic Journal of Botany**, v. 29, n. 3, pp. 378–381, Jun. 2011.

TOWNSEND, C. R., BEGON, M., HARPER, J. L. **Fundamentos em Ecologia**. 4. ed. Porto Alegre: Editora Artmed, 2007.

ULRICH, W. **CoOccurrence – a FORTRAN program for species co-occurrence analysis**, 2006. Disponível em: <www.uni.torun.pl/~ulrichw>. Acesso em: 4 nov. 2013

Turnover - a FORTRAN program for analysis of species associations, 2012. Disponível em: <www.umk.pl/~ulrichw>. Acesso em: 4 nov. 2013

URIARTE, M., CONDIT, R., CANHAM, C. D., HUBBELL, S. P. “A spatially explicit model of sapling growth in a tropical forest: does the identity of neighbours matter?”, **Ecology**, v. 92, n. 2, pp. 348–360, Abr. 2004.

VALENTIN, J. L. **Ecologia Numérica – Uma introdução à análise multivariada de dados ecológicos**. Rio de Janeiro: Editora Interciência, 2000.

VEECH, J. A. "A probabilistic model for analysing species co-occurrence", **Global Ecology and Biogeography**, v. 22, n. 2, pp. 252–260, Fev. 2013.

WALTHER, G. *et al.* "Ecological responses to recent climate change", **Nature**, v. 416, pp. 389–395, Mar. 2002.

WANG, X. *et al.* "Species associations in an old-growth temperate forest in north-eastern China", **Journal of Ecology**, v. 98, n. 3, pp. 674–686, Mai. 2010.

WANG, W., YANG, J. . M. R. STING: A statistical information grid approach to spatial data mining. In: **Proceedings International Conference Very Large Data Bases**, pp. 186-195, Athens, Greece, Ago. 1997.

WARD, J. H. J. 'Hierarchical Grouping to Optimize an Objective Function", **Journal of the American Statistical Association**, v. 58, n. 301, pp. 236–244, Abr. 1963.

WHITTAKER, R. H. "Vegetation of the Siskiyou Mountains, Oregon and California", **Ecological Monographs**, v. 30, n. 3, pp. 279–338, Jul. 1960.

WIEGAND, T. *et al.* Testing the independent species' arrangement assertion made by theories of stochastic geometry of biodiversity. **Proceedings of the Royal Society B Biological Sciences**, v. 279, n. 1741, pp. 3312–20, Mar. 2012.

WIEGAND, T., GUNATILLEKE, S., GUNATILLEKE, N. "Species associations in a heterogeneous Sri Lankan dipterocarp forest", **The American naturalist**, v. 170, n. 4, pp. E77–E95, Ago. 2007.

WIEGAND, T., MARTÍNEZ, I., HUTH, A. "Recruitment in tropical tree species: revealing complex spatial patterns", **The American naturalist**, v. 174, n. 4, pp. E106–E140, Out. 2009.

WIEGAND, T., MOLONEY, K. A. "Rings, circles, and null-models for point pattern analysis in ecology", **Oikos**, v. 104, n. 2, pp. 209–229, Fev. 2004.

WISDOM, M. D. “A data mining approach to predictive vegetation mapping using probabilistic graphical models”, **Ecological Informatics**, v. 6, n. 2, pp. 111–124, Mar. 2011.

WISZ, M. S. *et al.* “Effects of sample size on the performance of species distribution models”, **Diversity and Distributions**, v. 14, n. 5, pp. 763–773, Set. 2008.

WITTMAN, S. E. *et al.* “Species interactions and thermal constraints on ant community structure”, **Oikos**, v. 119, n. 3, pp. 551–559, Set. 2010.

WU, X., KUMAR, V., QUINLAN, J. R., GHOSH, J., YANG, Q., MOTODA, H., MCLACHLAN, G. F., NG, A., LIU, B., YU, P. S., ZHOU, Z., STEINBACH, M., HAND, D. J., S. D. “Top 10 algorithms in data mining”, **Knowledge and Information Systems**, v. 14, n. 1, pp. 1–37, Dez. 2007.

XIMENES, A. C. **Mapas Auto-Organizáveis para a Identificação de Ecorregiões no Interflúvio Madeira-Purus: Uma Abordagem da Biogeografia Ecológica**. [s.l.] Instituto Nacional de Pesquisas Espaciais, 2008.

ZAKI, M. J. “Scalable Algorithms for Association Mining”, **IEEE Transactions on Knowledge and Data Engineering**, v. 12, n. 3, pp. 372–390, Jun. 1998.

ZHANG, T., RAMAKRISHNAN, R., LIVNY, M. BIRCH: an efficient data clustering method for very large databases. In: **Proceedings ACM SIGMOD International Conference on Management of Data**, pp. 103-114, Montreal, Canada, Jun. 1996.

Apêndices

Apêndice A - Permissão de Uso da Base de Barro Colorado Island

Subject: BCI 50 ha Plot Data Download
Dear Luis Alexandre Estevao da Silva,
You have agreed to the following Terms and Conditions for the Use of the Barro Colorado Island 50 Hectare Plot Data:
You and your collaborators will not share the BCI 50-ha nor small plot data with other parties not included on the Request for Data Access Proposal that you filled online.
Publications using BCI data will include the following citations:

Hubbell, S.P., Condit, R., and Foster, R.B. 2005. Barro Colorado Forest Census Plot Data.

URL http://ctfs.arnarb.harvard.edu/webatlas/datasets/bci.

Condit, R. 1998. Tropical Forest Census Plots. Springer-Verlag and R. G. Landes Company, Berlin, Germany, and Georgetown, Texas.

Hubbell, S.P., R.B. Foster, S.T. O'Brien, K.E. Harms, R. Condit, B. Wechsler, S.J. Wright, and S. Loo de Lao. 1999. Light gap disturbances, recruitment limitation, and tree diversity in a Neotropical forest. Science 283: 554-557.

Publications will include an acknowledgement of the support of the Center for Tropical Forest Science of the Smithsonian Tropical Research Institute and the primary granting agencies that have supported the BCI plot:

The BCI forest dynamics research project was made possible by National Science Foundation grants to Stephen P. Hubbell:

DEB-0640386, DEB-0425651,DEB-0346488, DEB-0129874, DEB-00753102, DEB-9909347, DEB-9615226, DEB-9615226, DEB-9405933, DEB-9221033, DEB-9100058, DEB-8906869, DEB-8605042,DEB-8206992, DEB-7922197, support from the Center for Tropical Forest Science, the Smithsonian Tropical Research Institute, the John D. and Catherine T. MacArthur Foundation, the Mellon Foundation, the Small World Institute Fund, and numerous private individuals, and through the hard work of over 100 people from 10 countries over the past two decades. The plot project is part the Center for Tropical Forest Science, a global network of large-scale demographic tree plots.

The PIs will be sending you an email shortly concerning collaboration and/or authorship. If you do not receive a response within the next month, it is assumed that they do not wish to be involved as collaborators or authors.
Either way, copies of articles should be sent to the BCI PIs prior to submission.
Once published, any manuscript making use of the BCI data should be sent to the PIs.
Please go to the following site using Mozilla Firefox to select the data you want to download (IE will not work): https://ctfs.arnarb.harvard.edu/CTFSReports
username: XXX / password: XXX

Apêndice B – *Scatter plots* com Coocorrências Positivas

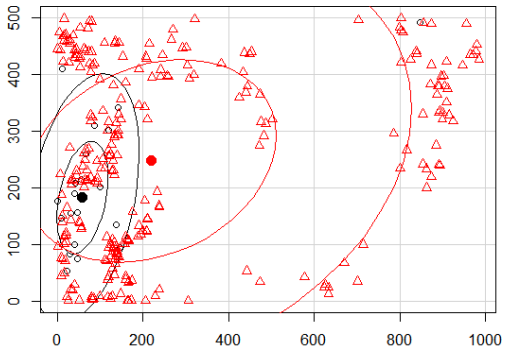


Fig. 29

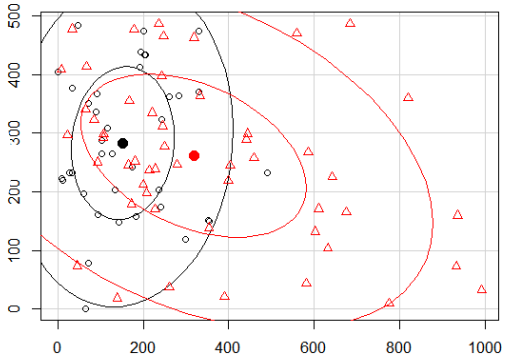


Fig. 30

Figura 29. ***Pourouma bicolor*** **(preta) e** ***Socratea exorrhiza*** **(vermelha)**

Figura 30. ***Inga thibaudiana*** **(preta) e** ***Pterocarpus rohrii*** **(vermelha)**

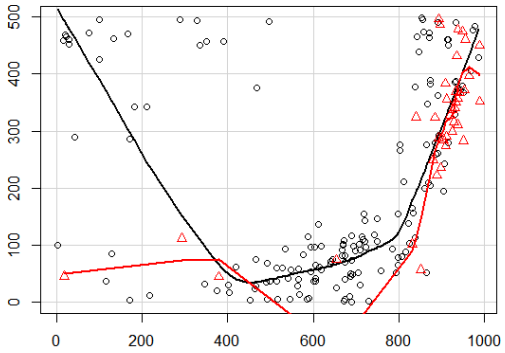


Fig. 31

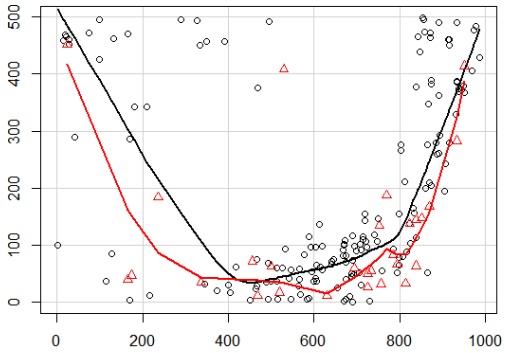


Fig. 32

Figura 31. ***Trophis caucana*** **(vermelha) e** ***Ocotea whitei*** **(preta)**

Figura 32. ***Ocotea whitei*** **(preta) e** ***Virola multiflora*** **(vermelha)**

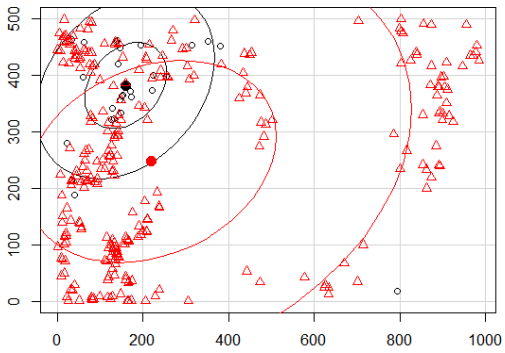


Fig. 33

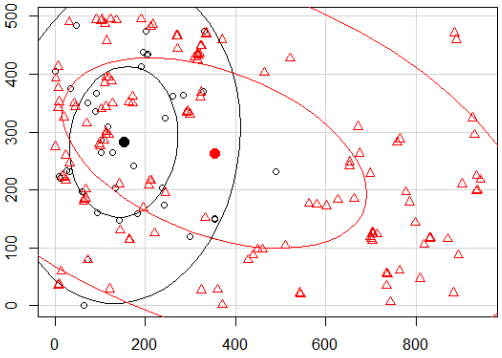


Fig. 34

Figura 33. ***Socratea exorrhiza*** **(vermelha) e** ***Perebea xanthochyma*** **(preta)**

Figura 34. ***Inga thibaudiana*** **(preta) e** ***Zanthoxylum ekmanii*** **(vermelha)**

Apêndice C – *Scatter plots* com Coocorrências Negativas

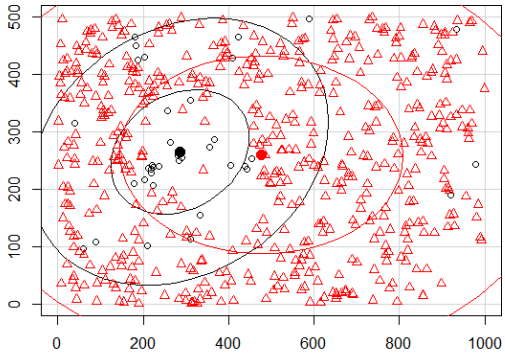


Fig. 35

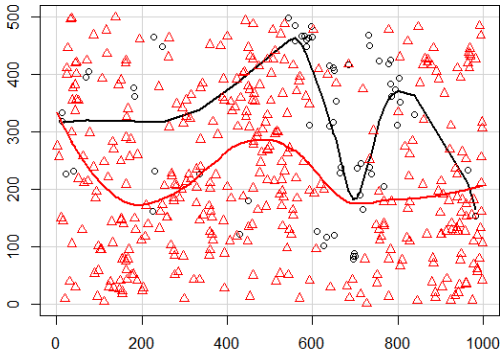


Fig. 36

Figura 35. ***Virola sebifera*** **(vermelha) e** ***Spondias mombin*** **(preta)**

Figura 36. ***Adelia triloba*** **(preta) e** ***Tetragastris panamensis*** **(vermelha)**

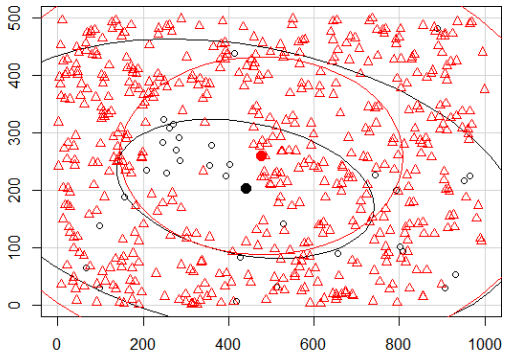


Fig. 37

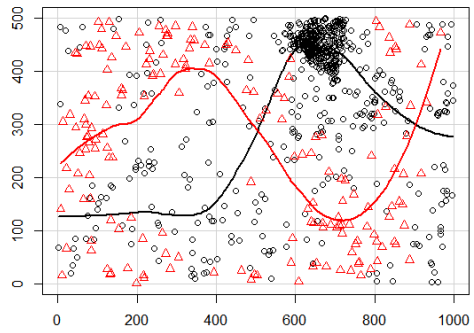


Fig. 38

Figura 37. ***Virola sebifera*** **(vermelha) e** ***Attalea butyracea*** **(preta)**

Figura 38. ***Gustavia superba*** **(preta) e** ***Unonopsis pittieri*** **(vermelha)**

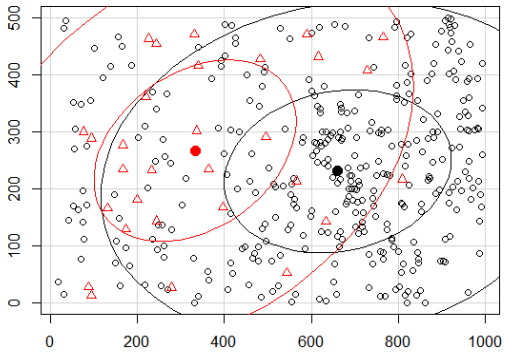


Fig. 39

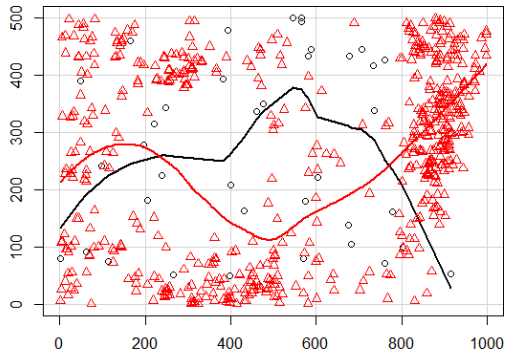
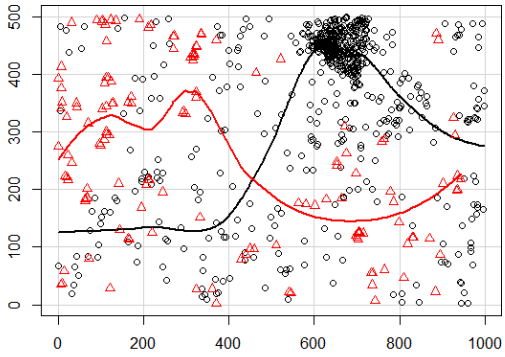


Fig. 40

Figura 39. ***Platypodium elegans*** **(vermelha) e** ***Guarea guidonia*** **(preta)**

Figura 40. ***Casearia sylvestris*** **(preta) e** ***Poulsenia armata*** **(vermelha)**



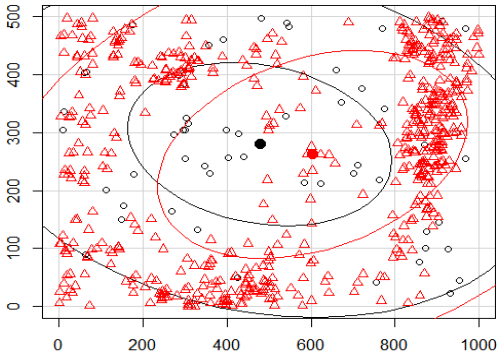


Fig. 41 **Fig. 42**

Figura 41. ***Zanthoxylum ekmanii*** **(vermelha) e** ***Gustavia superba*** **(preta)**

Figura 42. ***Lacmellea panamensis*** **(preta) e** ***Poulsenia armata*** **(vermelha)**

Apêndice D – *Scatter plots* com Coocorrências Positivas em Grupos

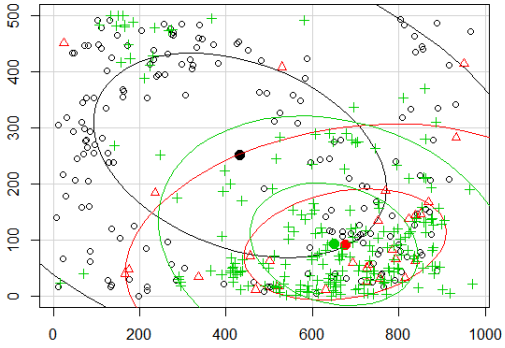


Fig. 43

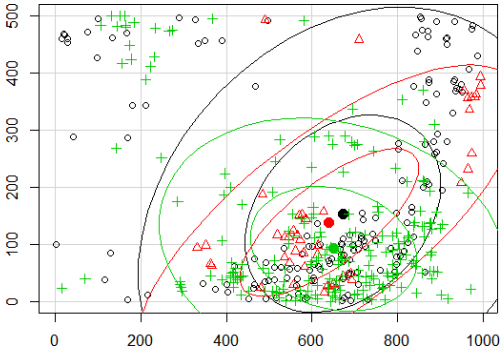


Fig. 44

Figura 43. ***Xylopia macrantha*** **(verde),** ***Virola multiflora*** **(vermelha) e** ***Unonopsis pittieri*** **(preta)**

Figura 44. ***Ocotea whitei*** **(preta),** ***Xylopia macrantha*** **(verde) e** ***Terminalia oblonga*** **(vermelha)**

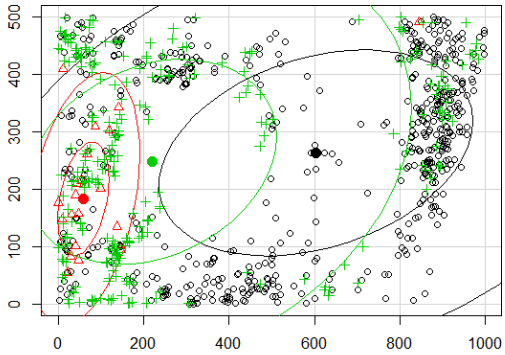


Fig. 45

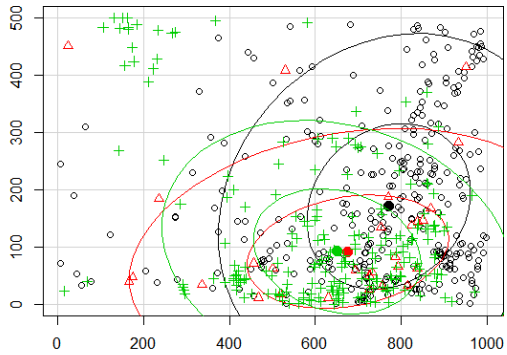


Fig. 46

Figura 45. ***Poulsenia armata*** **(preta),** ***Socratea exorrhiza*** **(verde) e** ***Pourouma bicolor*** **(vermelha)**

Figura 46. ***Drypetes standleyi*** **(preta),** ***Xylopia macrantha*** **(verde) e** ***Virola multiflora*** **(vermelha)**

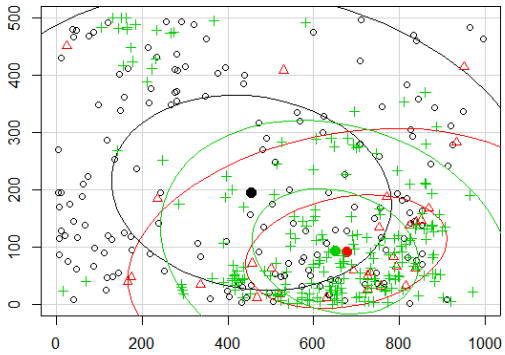


Fig. 47

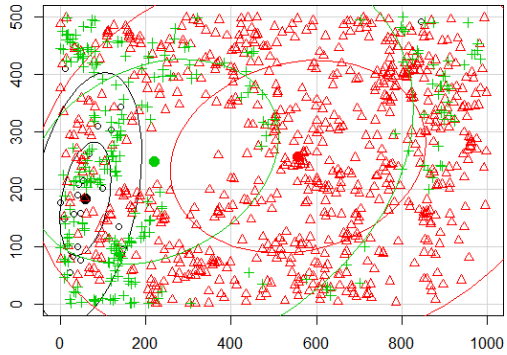


Fig. 48

Figura 47. ***Guatteria dumetorum*** **(preta),** ***Xylopia macrantha*** **(verde) e** ***Virola multiflora*** **(vermelha)**

Figura 48. ***Pourouma bicolor*** **(preta),** ***Socratea exorrhiza*** **(verde) e** ***Quararibea asterolepis*** **(vermelha)**

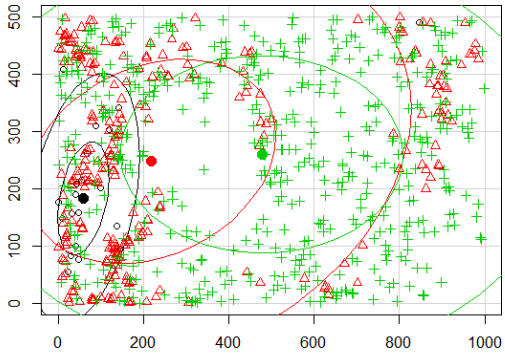


Fig. 49

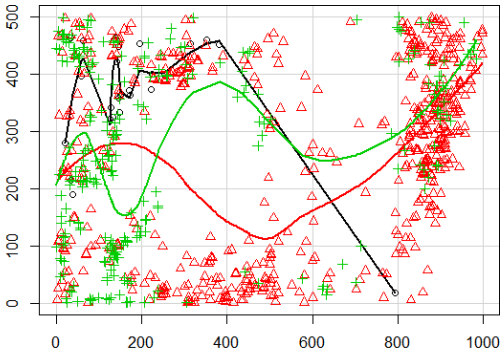


Fig. 50

Figura 49. ***Pourouma bicolor*** **(preta),** ***Virola sebifera*** **(verde) e** ***Socratea exorrhiza*** **(vermelha)**

Figura 50. ***Perebea xanthochyma*** **(preta),** ***Socratea exorrhiza*** **(verde) e** ***Poulsenia armata*** **(vermelha)**

Apêndice E – Parte do código fonte utilizado na implementação

```
#MINERPLOT
#PARTE 1: CARGA DAS BIBLIOTECAS
library(RODBC);     #No Postgresql - biblioteca ODBC (data source BCI)
library(scales);    #necessária para o plot
library(arules);    #Apriori
library(ggplot2);   #plots avançados
library(mclust);    #package MCLUST - EM
library(car)        #usado nos scatter plots

#PARTE 2: CARGA DOS DADOS - conexão do banco
canal <- odbcConnect("bci",case="postgresql",believeNRows=FALSE);

#PARTE 3: ETL
#ROTINA 3.1 - preparo do dataset por distância
execute_insert<-sqlQuery(canal,"INSERT INTO dataset_radius_20 (treeid,transacao)
(SELECT treeid, array_agg(DISTINCT latin)
FROM
(SELECT latin,treeid
    FROM
    (SELECT b.treeid,a.latin
     FROM (SELECT latin,treeid,gx,gy
        FROM censu7
        WHERE dbh >100
        AND latin IN (SELECT latin
                FROM censu7
                WHERE dbh>100
                GROUP BY latin
                HAVING count(*)>=20
                ORDER BY latin)) a,
                        (SELECT latin,treeid,gx,gy
                        FROM censu7
                        WHERE dbh >100
                        AND latin IN (SELECT latin
                                        FROM censu7
                                        WHERE dbh>100
                                        GROUP BY latin
                                        HAVING count(*)>=20
                                        ORDER BY latin)) b
WHERE (SQRT((a.gx-b.gx)*(a.gx-b.gx) + (a.gy-b.gy)*(a.gy-b.gy) ) <= 20) ) h) vv
GROUP BY treeid)");

#PARTE 4: ANÁLISE DE ASSOCIAÇÃO

#ROTINA 4.1: carga do dataset para o apriori - alternativa via .csv
txn<-read.transactions("recrutadas_20m.csv",format="basket",sep=",")
#ROTINA 4.2: análise exploratória do dataset
# verificação do total de espécies de cada dataset gerado
dim(txn)
#plot pela frequência
plot(itemFrequency(txn),xlab="species",ylab="frequency",col=ifelse(itemFrequency(txn)<0.5,'red','blue'),
pch=19)
#ROTINA 4.3 - determina a sp com a menor frequência
#variável para armazenar o menor valor
menor=1;
#total de espécies - 117
tot=dim(txn)[2];
#transformação em matriz
txn_mat<-as.matrix(txn_mat);
```

```
#verifica a espécie com a menor frequência
for (i in 1:tot)
{
  if (txn_mat[i]<menor)
    menor<-txn_mat[i];
}
```

#ROTINA 5 - APLICAÇÃO DO Apriori

#ROTINA 5.1 - preparação para a aplicacao do Apriori

```
tot_geral=0;
#Total de Regras com qualidade segundo o lift especificado
tot_com_qualidade=0;
#total de regras com correlação positiva
total_positivo=0;
#total de regras com correlação negativa
total_negativo=0;
#suporte iniciando no valor obtido para a variável menor
suporte_maximo=0.9;
#confiança iniciando em 0.3
confianca_minima=0.3;
#nr mínimo de elementos
minimo_len=2;
#nr máximo de elementos
maximo_len=3;
```

#ROTINA 5.2 - Aplicação do Apriori com os parêmetros

```
rules<-apriori(txn, parameter = list(sup=menor, conf=confianca_minima, minlen=minimo_len,
maxlen=maximo_len, target="rules"));
#limpa a tela
cat(rep("\n",64));
#variável usada para verificar se foi criada alguma regra tam=length(rules);
#definição da variável que vai acumular o total de regras
tot_geral = tot_geral + tam;
        #teste para verificar se foi gerada alguma regra
        if (tam > 0)
        {
                #verifica as métricas para as regras geradas
                m<-interestMeasure (rules, c("support", "confidence", "lift", "chiSquare", "leverage",
"hyperLift", "cosine","fishersExactTest", "gini", "conviction", "doc", "hyperconfidence", "oddsratio",
"rld", "coverage"), txn);
                #contador de regras geradas
                for (z in 1:tam)
                {
                        #exibição dos resultados parciais
                        cat("\n----------------------------------------------------");
                        cat("\n Total de regras analisadas:",tot_geral);
                        cat("\n Total de regras inseridas :",tot_com_qualidade);
                        #verifica a espécie antecedente
                        lhs = capture.output(print(inspect(rules[z]@lhs)));
                        #verifica a espécie consequente
                        rhs = capture.output(print(inspect(rules[z]@rhs)));
                        #cria a string para a inserir no banco de regras
                        insere<-paste("sqlQuery(canal,",rawToChar(as.raw(34)),"insert into
regras_20m (nr_regra, max_len_val, lhs1, lhs2, lhs3, rhs, suporte, confianca, lift, chisquare, leverage,
                        hyperlift, cosine, fisher, gini, conviction, doc, hyperconfidence, oddsratio, rld,
coverage)  values('",z, "', '", maximo_len, "', '", lhs[2][1], "', '", lhs[3][1], "', '", lhs[4][1], "', '", rhs[2], "', '",
m[z,1], "', '", m[z,2], "', '", m[z,3], "', '", m[z,4], "', '", m[z,5],  "', '", m[z,6], "', '", m[z,7], "', '", m[z,8], "',
'", m[z,9], "', '", m[z,10], "', '", m[z,11], "', '", m[z,12], "', '", m[z,13], "', '", m[z,14], "', '", m[z,15], "')",
rawToChar(as.raw(34)),")");
                        #inserção no banco das regras geradas
```

```
                        insere<- eval(parse(text=insere));
                        #conta regras de boa qualidade
                        tot_com_qualidade = tot_com_qualidade + 1;
                        #verifica se a correlação é positiva
                        if (m[z,3]>1){
                          total_positivo=total_positivo + 1;
                        }else {
                          if (m[z,3]<1){
                          #verifica se a correlação é negativa
                          total_negativo=total_negativo + 1;
                          };
                        }
                        #exibe os totais positivos e negativos parciais
                        cat("\n Total de regras com correlação positiva :",total_positivo);
                        cat("\n Total de regras com correlação negativa :",total_negativo);
                    }
   }
#listagem final dos resultados
cat(rep("\n",64));
cat("\notal de regras com qualidade:",tot_com_qualidade);
#rotina para ajuste final dos resultados no banco
if (tot_geral>0){
  sqlQuery(canal,"UPDATE regras_20m set lhs1=trim(substring(lhs1,('[A-Z][a-z]*'||' '||'[a-z][a-z]*')))");
  sqlQuery(canal,"UPDATE regras_20m set lhs2=trim(substring(lhs2,('[A-Z][a-z]*'||' '||'[a-z][a-z]*')))");
  sqlQuery(canal,"UPDATE regras_20m set lhs3=trim(substring(lhs3,('[A-Z][a-z]*'||' '||'[a-z][a-z]*')))");
  sqlQuery(canal,"UPDATE regras_20m set rhs=trim(substring(rhs,('[A-Z][a-z]*'||' '||'[a-z][a-z]*')))");
}else {
 print("Nenhuma regra gerada!");}
sqlQuery(canal,"UPDATE regras_20m set max_len_val=2 where lhs2 is null");
Sys.time(); #fim
```

#ROTINA 6 - AVALIAÇÃO DOS RESULTADOS POR MEIO DE SCATTER PLOTS

#ROTINA 6.1 - SMOOTH DA CORRELAÇÃO POSITIVA

```
axis<-sqlQuery(canal,"SELECT latin,habitat,gx,gy FROM censu7 WHERE DBH>100 AND latin IN
('Drypetes standleyi','Virola multiflora','Xylopia macrantha')");
```

#ROTINA 6.2 - SMOOTH COM O CENTRO DA DISTRIBUIÇÃO

```
scatterplot(axis$gy ~ axis$gx | axis$latin,data=axis,legend.plot=FALSE,
        ellipse=TRUE,reg.line=FALSE,smooth=FALSE,robust=TRUE)
```

#ROTINA 7 – CLUSTERIZAÇÃO - ROTINA 7.1

```
axis<-sqlQuery(canal,"SELECT treeid,habitat,gx,gy,latin FROM censu7 WHERE latin IN ('Ocotea
whitei','Virola multiflora','Xylopia macrantha') and  dbh>100");
myData <- data.frame(axis);
x = as.matrix(myData[,c(3,4)]);
dens = densityMclust(x);
summary(dens);
#plots coloridos com as regiões mais densas por estimativa
plot(dens,x,col="black",points.col=mclust.options()$classPlotColors[dens$classification],
    pch = dens$classification)
#plot por densidade
plot(dens,type="image",col=topo.colors(70))
#plot 3D
plot(dens,type="persp")
#acréscimo como campos desses resultados
myData$cluster <- dens$classification
myData$density <- dens$density
names(myData)
head(myData)
write.table(myData[2:7],file="myData2.csv",sep="\t",row.names=T)
```