ARTIGO

A.to.Z Novas práticas em Informação e Conhecimento

# Mineração de dados: busca de conhecimento sobre a evolução do canto da família *Thamnophilidae*

Letícia da Costa e Silva[1], Denise Fukumi Tsunoda[2], Viviane Deslandes[3]

**Resumo**

**Introdução**: Descreve a utilização de uma técnica de mineração de dados sobre o canto, a biologia e o micro-habitat da família *Thamnophilidae* (Aves) a fim de encontrar padrões que os relacionem. **Método**: Uma base de dados foi construída em planilha Excel® relacionando 82 espécies da família da ave *Thamnophilidae* com diversos atributos referentes às características do canto, da biologia e do micro-habitat em que são encontradas. Na análise utilizou-se o algoritmo Apriori no *software* WEKA 3.7.1. **Resultados**: Ao associar os diferentes atributos de 82 espécies diferentes considerando o suporte mínimo de 10% e a confiança mínima de 90% foram encontrados 172 padrões, dos quais 42 continham um dos atributos do canto: PC1 e PC2. Os padrões que relacionavam o atributo PC2 foram os mais significativos ao indicar a relação deste com o tamanho e gênero da família. Os resultados colaboraram para gerar a hipótese de que os atributos do canto possuem comportamentos não relacionados. **Conclusões**: O experimento demonstrou que o algoritmo pode ser melhor aproveitado em bases de dados maiores e/ou cuja padronização dos dados apresente menor número de categorias, o que pode ser uma limitação no campo da macroecologia. Mas, ao mesmo tempo, se mostrou um instrumento alternativo para o estudo exploratório de relações entre diversos atributos, cujos resultados podem servir de objetos de análises mais aprofundadas.

**Palavras-chave**

Mineração de dados; Bases de dados; Aves florestais; *Thamnophilidae* (ave); Cantos dos pássaros.

## Introdução

Em diversas áreas do conhecimento, dados estão sendo coletados e acumulados em grande escala nos meios digitais. Há urgência na descoberta de teorias computacionais e ferramentas que auxiliem os humanos na extração de conhecimentos úteis a partir deste crescente volume de bancos de dados. A Descoberta de Conhecimento em Base de Dados (DCBD), processo criado por Fayyad, Piatetsky-Shapiro, Smyth (1997) surge como uma alternativa viável de atender parte desta demanda, e procura, em um nível abstrato, desenvolver métodos e técnicas que ofereçam significado a dados armazenados digitalmente.

Tudo o que é coletado do nosso ambiente são as básicas evidências usadas para construir teorias e modelos do universo em que vivemos (BAÇÃO e PAINHA, 2003). O computador tem auxiliado os humanos nesta coleta ao armazenar dados em bancos que crescem continuamente de volume em dois sentidos: no número de registros ou objetos e no número de campos ou atributos relacionados ao objeto; além de permitir o relacionamento entre eles (FAYYAD, PIATETSKY-SHAPIRO E SMYTH, 1997). Dessa forma, o emprego do processo de DCBD, que inclui técnicas de mineração de dados, torna possível a extração de conhecimento a partir da navegação em variados ambientes informacionais.

Diversas ciências vêm utilizando a mineração de dados para o aprimoramento do saber em seus campos, entre elas: a administração, o marketing, a astronomia, a medicina, a física, a ciência da informação. No âmbito das ciências naturais, a astronomia destaca-se no uso de técnicas computacionais para adquirir conhecimento, além do recente interesse na utilização de bancos

de dados compartilhados sobre o genoma humano e de outros animais.

Na biologia e na geografia, a mineração de dados geralmente está associada ao sensoriamento remoto de imagens e no Sistema de Informações Geográficas (SIG). Ferramentas como o SIG, proporcionam aos usuários a visualização de fenômenos complexos que ocorrem na superfície terrestre por meio do armazenamento e geração de grandes quantidades de dados geo-referenciados (BAÇÃO e PAINHA, 2003, p. 6) que facilitam a sua classificação e sumarização. Estudos como o de Li, Di e Li (2000) aplicam algoritmos de mineração de dados sobre saídas de um sistema de informação gerencial e de ferramentas de sensoriamento remoto de dados para melhorar a classificação de imagens do uso de terra.

Recentemente, a macroecologia aplica técnicas de mineração utilizando metadados organizados por diversas pessoas e instituições a fim de descobrir padrões gerais em ecologia. Para Blackburn (2004, p. 1) entender padrões e processos na macroecologia é um desafio metodológico, uma vez que as escalas espaciais e temporais para identificar a distribuição e abundância de espécies são amplas e demandam muitos dados. Conforme Gotelli (2008, p. 4) o uso da mineração de dados na macroecologia (e na biogeografia) pode proporcionar correlações que revelem mecanismos em modelos estatísticos compreensíveis que incluam tantos dados possíveis de serem reunidos. Bekker *et al.* (2008) discorrem sobre o uso da ecoinformática em grandes bancos de dados que reúnem informações fitossociológicas das espécies buscando padrões de sua relação com o ambiente e de sua distribuição em escalas locais e regionais.

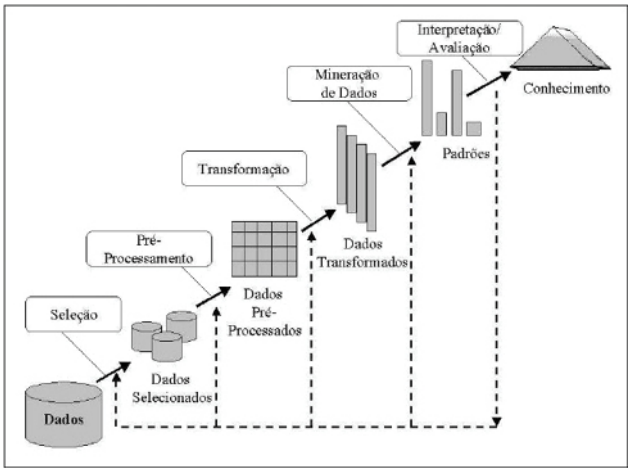
Apesar da aplicação da mineração de dados em diversos estudos na área de macroecologia ainda há um promissor cenário na sua exploração de forma a propiciar relevância, utilidade e validade na geração de novos conhecimentos. Este estudo tem o propósito de extrair conhecimento sobre a relação entre parâmetros do canto de aves *Thamnophilidae* com características de sua biologia e seu habitat. Para isso foi utilizada uma base de dados que reúne informações sobre a biologia, o canto e características do micro-habitat dessas aves.

O artigo divide-se em quatro partes. A primeira apresenta uma breve síntese sobre os conceitos de DCBD e mineração de dados, baseados em publicações de Fayyad, Piatetsky-Shapiro e Smyth, e indica os materiais e métodos utilizados a fim de explicitar como este trabalho foi empiricamente sistematizado. A segunda detalha como foi realizado o pré-processamento dos dados antes de submetê-lo ao algoritmo de mineração de dados. A terceira apresenta os resultados gerados pelo algoritmo. E por fim, a quarta, discute os conhecimentos obtidos após a mineração dos dados.

## DCBD e mineração de dados

Fayyad, Piatetsky-Shapiro e Smyth (1997, p.6) apresentam o processo de DCBD em cinco etapas:

**Figura 1 – Uma visão geral das fases inclusas no processo de DCBD**




Fonte: Neves, 2003, p. 28.

seleção da base de dados, pré-processamento (seleção e tratamento dos dados), transformação, mineração dos dados, interpretação e avaliação, conforme demonstrado na Figura 1.

A DCBD transforma dados de bancos tipicamente volumosos em produtos que facilitam o entendimento das relações entre eles. Vários são os produtos que podem se extrair deste processo como relatórios, modelagem de processos,

ou ainda modelos de previsão. Dessa forma, dependendo do que se pretende extrair da base de dados selecionada é escolhido um algoritmo que minerará os dados em busca da resposta procurada.

A mineração de dados envolve enquadrar modelos a (ou encontrar padrões em) dados observados. A maioria dos métodos de mineração é baseada em técnicas testadas de aprendizado de máquina e reconhecimento de padrões e estatística, como classificação, associação, regressão, *clustering*, sumarização, modelagem de dependência, entre outros (FAYYAD, PIATETSKY-SHAPIRO, SMYTH, 1997, p. 8-10).

A escolha apropriada da técnica de mineração a ser utilizada é essencial para que sejam extraídos os conhecimentos pretendidos. Entender como será aplicada a mineração facilita a compreensão do usuário sobre sua contribuição e aplicabilidade na geração de conhecimento a partir de uma grande quantidade de dados.

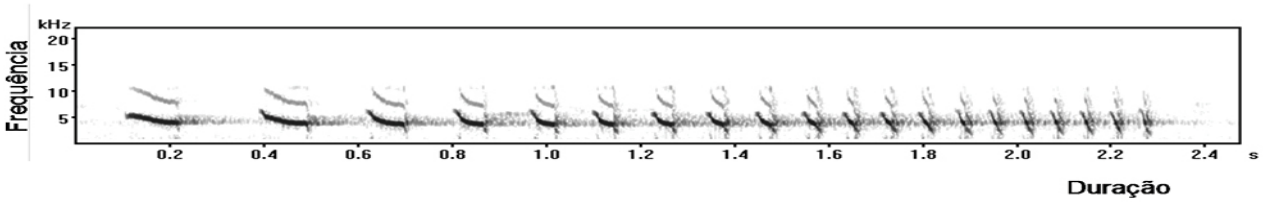
## Materiais e métodos

A base de dados utilizada no estudo relaciona 82 espécies da família da ave *Thamnophilidae* com diversos atributos referentes às características do canto, da biologia e do micro-habitat em que são encontradas. Esta base de dados foi construída em planilha Excel® do pacote Microsoft Office®.

Os cantos das 82 espécies de *Thamnophilidae* foram obtidos da base de dados do XENO-CANTO (2010). Respostas a *playback*[1] , apelos[2] e cantos com taxa de amostragem e resolução menores do que 22.050 Hz e 16 bits foram excluídos da amostra. A fim de evitar pseudo-repetição, foi medido somente um canto por indivíduo e localidade, com um total de 3 a 17 cantos por espécie, totalizando 676 cantos analisados. A partir das gravações foram gerados sonogramas, que são representações gráficas dos cantos, onde a frequência do canto é representada no eixo Y (em Hz) e a duração do canto é representada no eixo x (em segundos), conforme demonstrado na Figura 2.

Todos os sonogramas foram gerados no software RAVEN PRO 1.4 (CHARIF et al., 2010). As seguintes medidas acústicas foram obtidas: frequência maior (Hz), frequência menor (Hz), variação na frequência (Hz), frequência de pico (Hz), duração do canto (s), entropia agregada (s) e número de notas (unidades). Análise de Componentes Principais (PCA) foi aplicada aos parâmetros do canto, visando reduzir sua dimensionalidade. Dessa forma, os valores dos dois primeiros componentes principais (PC1 e PC2) do canto foram usados para representar o canto de cada uma das espécies do estudo, e os dados concernentes a biologia e aspectos do micro-habitat das espécies foram coletados dos trabalhos de Dunning (2008) e de Del Hoyio, Elliot e Smyth (2003).

**Figura 2 – Exemplo da representação de um sonograma (ou audioespectrograma) do canto de *Drymophila malura*, um *thamnophilideo* de sub-bosque**




Fonte: os autores.
Obs.: A frequência (kHz) é representada no eixo y do gráfico e a duração do canto (segundos), no eixo x. Os traços mais escuros dispostos em sequencia na figura representam as notas do canto, nesse caso, 20 notas.

[1] É a resposta da ave à gravação de voz de sua própria espécie.
[2] São as vocalizações menos complexas e não territoriais.

O intuito desse trabalho é a detecção de padrões que relacionem a biologia e micro-habitat das espécies aos atributos do canto. Devido à utilização de dois parâmetros do canto para a mineração de dados, optou-se por não utilizar atributo meta na busca de padrões. Dessa forma, a mineração dos dados foi realizada no *software* WEKA 3.7.1, por meio do algoritmo APRIORI (AGRAWAL *et al*, 1993), instalado em um *notebook* Toshiba Satellite U305-S5127, processador Intel Core 2 Duo T7100 de 1.8GHz, memória 2GB RAM. O APRIORI é um método clássico de mineração de regras de associação baseado no princípio da antimonotonicidade do suporte, ou seja, "um *k-itemset* somente pode ser frequente se todos os seus *(k-1)-itemsets* forem frequentes" (GOLDSCHMIDT e PASSOS, 2005). Este algoritmo é um modelo não-paramétrico que tem por premissa a crença de que as relações que ocorrem de forma consistente no conjunto de dados repetir-se-ão em observações futuras. O método não exige conhecimento profundo do fenômeno a modelar (BAÇÃO e PAINHO, 2003, p. 5), sendo, dessa maneira, oportuno o seu uso na base utilizada neste trabalho.

Devido ao *software* WEKA ler arquivos em arff, foi necessário realizar duas conversões. Primeiramente de xls para csv e depois de csv para arff. A conversão de xls para csv foi realizada pelo comando salvar do Microsoft Excel®. A conversão de csv para arff foi realizada pelo site do pesquisador esloveno Marko Tkalčič (2008), o qual oferece este serviço gratuitamente.

Antes das duas conversões, os espaços e as vírgulas contidas no arquivo xls foram substituídos por "." (um ponto), uma vez que a leitura em arff entende que os espaços e vírgulas separam um atributo do outro. No site que permite a conversão, os atributos foram selecionados como nominal conforme definido na seção 'Categorização dos dados'.

O arquivo arff gerado foi salvo e importado para o WEKA 3.7.1 a fim de se aplicar o APRIORI. Para a definição do suporte mínimo fez-se necessário analisar como se daria a primeira etapa dos cálculos do algoritmo em relação aos atributos do canto (PC1 e PC2), em que há a exclusão de categorias inferiores ao valor definido. Para evitar que fossem consideradas poucas categorias, ou até nenhuma, calculou-se qual era a porcentagem de ocorrência (ou seja, o suporte mínimo) das categorias de cada atributo do canto em relação ao total de linhas (82), conforme indica a Tabela 1.

**Tabela 1 – Porcentagem de ocorrência das categorias dos atributos PC1 e PC2**

| PC1 | Núm. Ocorrências | Proporção rel. ao total | PC2 | Núm. Ocorrências | Proporção rel. ao total |
|---|---|---|---|---|---|
| -4a-2 | 15 | 0,183 | -4a-2 | 1 | 0,012 |
| -2a0 | 34 | 0,415 | -2a0 | 24 | 0,293 |
| 0a2 | 17 | 0,207 | 0a2 | 49 | 0,598 |
| 2a4 | 11 | 0,134 | 2a4 | 8 | 0,098 |
| 4a6 | 5 | 0,061 | Total | 82 | 1 |
| Total | 82 | 1 | | | |

Fonte: os autores.

A análise da Tabela 1 aponta que o suporte mínimo de 0,1 garante que um maior número de categorias seja considerado nos cálculos posteriores do algoritmo. Esta escolha possibilita a busca de um maior número de padrões entre os atributos do canto e os outros atributos da base de dados. Com este suporte mínimo somente as categorias 4a6 do PC1, 2a4 e -4a2 do PC2 serão desconsideradas. Aumentar o suporte, para 0,15 ou 0,2, por exemplo, excluiria categorias do PC1 com um número de ocorrência considerável, como 11 e 15 vezes.

Dessa forma, foi estipulado o suporte mínimo de 0,1 (ou 10%), o qual foi também considerado no segundo cálculo do algoritmo, em que verifica a ocorrência das categorias de todos os atributos da base em conjunto. E para o último cálculo, em que é considerado as ordens seqüenciais desses conjuntos acima da confiança mínima definida, foi utilizada a métrica de 0,9 (ou 90%).

Foram encontrados 172 padrões, das quais 42 apresentaram um dos dois atributos do canto (PC1 e PC2) utilizados para a mineração. Os campos vazios do banco de dados foram

preenchidos com 0 (zero) a fim de não serem considerados para análise. A versão 3.7.1 do WEKA oferece, para o algoritmo APRIORI, comandos que permitem a exclusão dos padrões retirados da base importada que relacionam os campos considerados vazios, quando preenchidos por zero. Esses comandos são os "*removeAllMissingCols*" e "*treatZeroAsMissing*".

## Pré-processamento dos dados

A base de dados continha, inicialmente, informações de 200 espécies distribuídas em 200 linhas. O Quadro 1 apresenta os atributos presentes inicialmente na base e suas respectivas descrições.

**Quadro 1 – Descrição resumida dos atributos originais da tabela**

| Atributo | Descrição |
|---|---|
| **Família** *Thamnophilidae* | Esta coluna lista as espécies da família *Thamnophilidae* que por sua vez são relacionadas aos parâmetros do canto e da biologia das espécies e características de seu micro-habitat. |
| **Parâmetros do canto** | PC1 e PC2 correspondem aos dois eixos principais que representam o canto de cada uma das espécies. Esses eixos foram obtidos da Análise de Componentes Principais, realizada a partir da média das seguintes variáveis do canto: frequência maior (Hz), frequência menor (Hz), variação na frequência (Hz), frequência de pico (Hz), duração do canto (s), entropia agregada (s) e número de notas (unidades). O PC1 representa os parâmetros relacionados à freqüência e entropia dos cantos, enquanto o PC2 representa os parâmetros temporais do canto: número de notas e duração do canto. |
| **Morfologia do bico** | Neste atributo, parte dos dados da biologia das espécies, foi excluído, uma vez que não continha nenhum campo preenchido. |
| **Dados da biologia** | Peso, tamanho, micro-habitat, tipo de dieta, modo de forrageio, altitude, movimento, bando misto e dimorfismo de plumagem. |

Fonte: os autores.

No atributo "dados da biologia", o peso (g) e o tamanho (cm) indicados estão relacionados à estrutura física de cada espécie. O micro-habitat aponta características mais específicas da vegetação em que cada espécie é encontrada. A dieta indica a alimentação de cada espécie. O modo de forrageio apresenta o agrupamento das espécies na busca por alimentos. A altitude aponta a elevação do terreno máxima em que cada espécie comumente é encontrada. O movimento indica se a espécie é residente ou possui hábito migratório. O bando misto diz respeito ao comportamento de forrageio da espécie e, por fim, o dimorfismo de plumagem aponta se há diferença de coloração na plumagem entre macho e fêmea de uma mesma espécie.

Os dados referentes à coluna Movimento continham a informação residente para todas as espécies, pois a família não é migratória, dessa forma esta coluna foi excluída. Algumas espécies não possuíam os atributos de canto preenchidos, assim as linhas referentes foram excluídas também. Ao final dessas exclusões a planilha continha 82 linhas (82 espécies) e 15 colunas (1 coluna que relaciona as espécies da família e 14 colunas que relacionam os dados dos atributos).

A categorização dos dados realizada para preparar a base para o processo de mineração de dados está descrito por atributo.

a) Família

Como não havia espécie repetida, os termos que indicavam a espécie foram suprimidos, restando somente os termos que designam o gênero da família. Por mais que fosse importante permanecer a indicação da espécie, se os dados não fossem padronizados não seria possível encontrar padrões por agrupamento ao utilizar o algoritmo APRIORI.

As espécies foram agrupadas em 32 gêneros diferentes: *Batara*, *Biatas*, *Cercomacra*, *Clytoctantes*, *Cymbilaimus*, *Dichrozona*, *Drymophilla*, *Dysithamnus*, *Epinecrophylla*, *Formicivora*, *Frederickena*, *Gyminocichla*,

*Herpsilochmus*, *Hypocnemis*, *Hypocnemoides*, *Hypoedaleus*, *Mackenziaena*, *Microrhopias*, *Myrmoborus*, *Myrmochanes*, *Myrmorchilus*, *Myrmotherula*, *Neoctantes*, *Pygiptila*, *Pyriglena*, *Sakesphorus*, *Sclateria*, *Taraba*, *Terenura*, *Thamnistes*, *Thamnomanes*, *Thamnophilus*.

b) Parâmetros do canto

PC1 e PC2 – os dados estavam indicados por números exatos[3]. O menor valor encontrado do PC1 foi -3,224 e o maior 4,370. Para se obter maior números de ocorrências iguais, optou-se por padronizá-las em intervalos de 2. Ex: -4 a -2; -2 a 0 e assim por diante. Para o PC2 o menor valor foi -2,120 e o maior 3,454. A padronização deste eixo seguiu o mesmo procedimento do PC1.

c) Dados da biologia

**Peso** (g) – os dados foram retirados de Dunning (2008) e indicam a média do peso por espécie em número exato. Algumas espécies não possuíam informação, enquanto outras não indicavam a média do peso, apesar de apresentá-lo em intervalo de mínimo e máximo ou pelo sexo (macho e fêmea). Para fins de padronização todos os pesos foram convertidos para o seu valor médio. O menor peso encontrado foi 7 g e o maior 131 g. Não havia concentrações entre o menor e o maior valor deste atributo, assim optou-se por padronizá-los em intervalos de 5 g. Ex.: 6-10; 11-15. As médias com casas decimais que coincidiam com o início ou fim de um intervalo foram incluídas no intervalo padronizado anterior. Ex: 35,8 no intervalo 31-35.

**Tamanho** (cm) - os dados estavam indicados por intervalos e por números exatos. O menor tamanho encontrado foi 6 cm e o maior 28 cm. Não havia grandes concentrações entre o menor e o maior valor deste atributo, assim optou-se por padronizá-los em intervalos de 5 cm. Ex: 6-10; 11-15. Os números exatos com casas decimais contidos no início ou no final de um intervalo padronizado, como 10,5, e intervalos como 9-11, igualmente presentes em dois intervalos subseqüentes, foram enquadrados no intervalo anterior, nestes casos no 6-10.

**Micro-habitat** - como os dados deste atributo estavam descritos por extenso em diversas seqüências, optou-se por padronizá-los em siglas que posteriormente foram distribuídas em cinco diferentes atributos, apresentados a seguir:

Dados relacionados ao estrato da floresta:

- DS – Dossel, Sub-dossel e estrato superior;
- SB – Sub-bosque;
- EM – Estrato médio.

Dados relacionados à floresta em montanha:

- FM – Floresta montana;
- FSM – Floresta sub-montana e de Pé-de-montanha.

Dados relacionados ao ambiente:

- Aberto – caatinga, savana, cerrado, chaco, campo aberto, campinas, clareiras, deserto, mangue, áreas abertas e semi-abertas;
- Fechado - floresta tropical úmida sempre verde, floresta úmida, floresta semi-úmida, floresta árida, floresta seca, floresta de pinus, floresta ripária, floresta de transição, floresta de araucária, floresta decídua, floresta semi-decídua, floresta de terra firme, floresta de galeria, floresta alagada, buritizal, floresta com bambus, floresta temperada, floresta de várzea, floresta de igapó, capões.

Dados relacionados ao status de conservação da floresta:

- VP – vegetação primária;
- VS – vegetação secundária;

[3] Neste trabalho são considerados exatos os valores que não são indicados na forma de intervalos. Exemplo: 566,67 ou 35 são considerados exatos, já 155-655 ou 44,3-53 são considerados intervalos.

- AA – áreas antropizadas como plantações, manejo de madeira, parques e jardins.

Bambu:

- BA - espécies que são fortemente associadas à vegetação com a presença de bambus.

**Dieta** – os dados deste atributo foram padronizados em insetos, artrópodes, pequenos vertebrados, moluscos, sementes, frutos e ovos e filhotes de aves, conforme abaixo:

- Insetos: larvas, pupas e adultos de insetos pertencentes às diversas famílias das ordens *Orthoptera*, (larvas e adultos de) *Lepidoptera*, *Coleoptera*, *Dermaptera*, *Hymenoptera*, *Hemiptera*, *Homoptera*, *Isoptera*, *Diptera*, e *Formicidae*;
- Artrópodes: quilópodes, escorpião, outros artrópodes, aranha, e pseudoescorpião;
- Pequenos Vertebrados: sapo, lagarto, pequenas cobras, lagartixas, lagartos do gênero *Anoles*, e ovos e filhotes de aves;
- Moluscos: lesma e caramujo;
- Sementes: pequenas sementes e sementes em geral;
- Frutos: frutos de qualquer tipo.

**Modo de forrageio** – neste atributo havia informações redundantes contidas em outros atributos. Assim, optou-se por excluir as informações redundantes e padronizar os dados em solitário, par e familiar, conforme abaixo:

- Solitário: solitário;
- Par: em par, em par próximo ao solo e em par na vegetação densa;
- Familiar: pequenos grupos familiares e grupos familiares;

Algumas informações anexas na descrição do modo de forrageio foram excluídas: raramente, frequentemente, no solo, estrato médio.

**Altitude** (m) – os dados estavam indicados por intervalos identificados por números exatos. A menor altitude encontrada foi 0 m e a maior 3050 m. Optou-se por informar a altitude máxima considerando intervalos de 1000 em 1000 metros, obtendo a seguinte padronização para este atributo: 1 (de 0 a 1000m), 2 (de 1001 a 2000m) e 3 (de 2001-3000m).

**Bando Misto** – os dados foram padronizados em N, NA, SO e S, conforme detalhamento a seguir:

- N – não segue bando;
- NA – não há informação, sem informação precisa, sem informação;
- SO – ocasionalmente seguidor de correição, raramente associado, segue raramente, segue às vezes, segue eventualmente, segue ocasionalmente;
- S – segue, segue frequentemente, segue regularmente, segue rotineiramente.

**Dimorfismo Plumagem** – os dados indicados eram: forte, fraco e não. Por sua simplificação, a padronização deste atributo foi considerada como: forte, fraco e N para não.

## Resultados

Os padrões que consideram os atributos do canto estão distribuídos ao longo dos 172 padrões totais recuperados, Deste total, foram encontrados 42 padrões que relacionaram os atributos do canto aos outros atributos da base. O número de ocorrência em conjunto dos atributos de cada padrão variou entre 8 a 13 vezes das 82 possíveis, conforme indicado na Tabela 2.

Foram recuperados 32 padrões relacionados ao PC2 e dez ao PC1. Contudo, esses atributos não ocorrem em conjunto em nenhum dos padrões. As categorias do canto presentes são -2 a 0 do atributo PC2, em especial, e as -4 a -2, 0 a 2 e 2 a 4. A Tabela 2 resume as informações encontradas em 24 dos 42 padrões recuperados de forma decrescente: do padrão com maior confiança (100%) para o de menor (90%).

**Tabela 2 – Padrões encontrados pelo APRIORI**

| Se | Então | Confiança |
|---|---|---|
| PC2=-2a0 e Ambiente=A.F e Dieta=artropodes.insetos | Tam(cm)=11--15 | 13/13 ou 100% |
| PC2=-2a0 e Dieta=artropodes.insetos e Alt(m)=2 | Tam(cm)=11--15 | 12/12 ou 100% |
| PC2=-2a0 e Peso(g)=16-20 | Tam(cm)=11--15 | 10/10 ou 100% |
| PC2=-2a0 e Peso(g)=6—10 | MF=solitario.par.familiar | 9/9 ou 100% |
| PC2=-2a0 e Peso(g)=16-20 e Dieta=artropodes.insetos | Tam(cm)=11--15 | 9/9 ou 100% |
| PC1=0a2 e Estrato=SB | Bambu=0 | 8/8 ou 100% |
| PC1=2a4 e MF=solitario.par.familiar | Dieta=artropodes.insetos | 8/8 ou 100% |
| PC1=-4a-2 e Tam(cm)=11—15 e Bambu=0 | Genero=Thamnophilus | 8/8 ou 100% |
| PC1=-4a-2 e Tam(cm)=11—15 e MF=solitario.par | Genero=Thamnophilus | 8/8 ou 100% |
| Genero=Thamnophilus e PC1=-4a-2 e Alt(m)=2 | MF=solitario.par | 8/8 ou 100% |
| PC2=-2a0 e Peso(g)=6—10 e Dieta=artropodes.insetos | MF=solitario.par.familiar | 8/8 ou 100% |
| PC2=-2a0 e MF=solitario.par.familiar e DP=2 | Bambu=0 | 8/8 ou 100% |
| PC2=-2a0 e StConservacao=VS e Dieta=artropodes.insetos | Tam(cm)=11--15 | 12/11 ou 92% |
| PC2=-2a0 e Dieta=artropodes.insetos e BM=SO | Tam(cm)=11--15 | 12/11 ou 92% |
| PC1=2a4 | Dieta=artropodes.insetos | 11/10 ou 91% |
| Genero=Thamnophilus e PC1=-4a-2 | MF=solitario.par | 11/10 ou 91% |
| PC1=-4a-2 e Ambiente=A.F | MF=solitario.par | 11/10 ou 91% |
| PC1=0a2 e Dieta=artropodes.insetos | MF=solitario.par.familiar | 11/10 ou 91% |
| PC2=-2a0 e Tam(cm)=11—15 e DP=2 | Bambu=0 | 11/10 ou 91% |
| PC1=-4a-2 e Tam(cm)=11—15 | Genero=Thamnophilus | 10/9 ou 90% |
| PC2=-2a0 e Peso(g)=16-20 | Dieta=artropodes.insetos | 10/9 ou 90% |
| StConservacao= VS e Alt(m)=1 | PC2=-2a0 | 10/9 ou 90% |
| PC2=-2a0 e Peso(g)=16-20 e Tam(cm)=11—15 | Dieta=artropodes.insetos | 10/9 ou 90% |
| PC2=-2a0 e Peso(g)=16-20 | Tam(cm)=11—15 e Dieta=artropodes.insetos | 10/9 ou 90% |

Fonte: os autores.

Tomando como exemplo o primeiro padrão da Tabela 2, o resultado deve ser interpretado da seguinte forma: se PC2 = -2a0, o ambiente= A.F (aberto e fechado) e a dieta= artrópodes.insetos, então o tamanho deverá ser 11-15 (entre 11 e 15 cm) com 100% de confiança. Considerando que os atributos PC2, ambiente e dieta ocorreram 13 vezes juntos, e nessas 13 vezes, o atributo tamanho foi encontrado em todas essas ocorrências.

Dos padrões com 100% de confiança encontram-se as razões de ocorrência em conjunto dos atributos: 13/13, 12/12, 10/10, 9/9 e 8/8; com 92% a razão 12/11; com 91% a razão 10/11; e com 90% a razão 10/9.

Vale ressaltar que os padrões que apresentaram o atributo Bambu estão todos indicando a informação 0 (zero). Isto se explica pelo fato do atributo ter sido categorizado em duas opções somente: 0 ou BA e, destes, 61 serem 0. Apesar da baixa ocorrência do BA, esta categoria foi incluída na segunda etapa de cálculos do algoritmo por ter confiança mínima de 0,25. Contudo, não se apresentou suficientemente em conjunto com outros atributos para ter a confiança mínima de 0,9. Ao contrário da opção 0, a qual está presente em grande parte dos padrões encontrados, apesar de terem sido utilizados os comandos "*removeAllMissingCols*" e "*treatZeroAsMissing*".

## Considerações finais

A grande porcentagem de ocorrência de algumas categorias do canto não garantiu sua recuperação nos 42 padrões encontrados, apesar do baixo

suporte mínimo (0,1) escolhido. Tanto a categoria -2 a 0 do atributo PC1 (41%) e a 0 a 2 do atributo PC2 (59%) são relacionadas de forma muito variável aos atributos de biologia e de micro-habitat das espécies.

Devido todos os padrões encontrados relacionando o atributo bambu terem apresentado informação zero, este atributo deve ser desconsiderado nas interpretações. A inclusão deste atributo foi feita com o intuito de estabelecer a relação da presença de bambu no ambiente com algum atributo do canto.
Apesar de o atributo PC2 ter ocorrido 32 vezes no total dos padrões obtidos, as relações geradas não foram as comumente esperadas ao considerar as hipóteses de evolução do canto. Entretanto, os 10 padrões recuperados para o PC1 permitem, em parte, essa associação. Esse atributo do canto representa os parâmetros de frequencia e entropia, e sua associação com a segunda menor classe de tamanho (11-15 cm) e o gênero *Thamnophilus* sugerem um aprofundamento no estudo dessas relações.

A forma da padronização dos dados da base e o seu tamanho, em linhas, não permitiram que fossem gerados padrões conclusivos ou que indicassem tendências marcantes em relação à evolução do canto da família *Thamnophilidae*. Porém, a padronização dos dados não pôde ser reduzida para menos categorias, caso contrário haveria perda significativa de informações dos atributos, e a amostra utilizada relaciona já um número considerável de espécies da família.

Entretanto, o APRIORI se mostrou útil como uma ferramenta exploratória, pois foi possível identificar indícios de interessantes relações entre os atributos do canto e os da biologia e micro-habitat para uso em novos estudos. Além disto, foi também verificado que há uma possível tendência no direcionamento distinto dos atributos do canto PC1 e PC2, uma vez que não foram encontrados em conjunto em nenhum dos padrões resgatados.

Dessa forma, apresenta-se como sugestão a aplicação do APRIORI em base de dados com maior número de linhas e que seja possível a padronização em um menor número de categorias, recomendando-se a sua aplicação, na macroecologia, para estudos exploratórios, a fim de aprofundar os estudos das relações entre atributos ao utilizar outros algoritmos, como os de análise multivariada.

## Referências

AGRAWAL, R.; IMIELINSKI, T.; SWAMI, A. Mining association rules between sets of itens in large databases. ACM SIGMOD Conference Management of Data, 1993. **Proceedings...** Disponível em: <http://www.rakesh.agrawal-family.com/papers/sigmod93assoc.pdf>. Acesso em: 10 nov. 2010.

BAÇÃO, F.; PAINHO, M. Aspectos metodológicos da utilização do data mining no âmbito da geografia. **Finisterra**, v. 38, n. 75, p. 135-147, 2003.

BEGON, M.; TOWNSEND, C. R.; HARPER, J. L. **Ecology**: from individuals to ecosystems. 4 ed. Oxford: Blackwell, 2006.

BEKKER, R. M. et al. Long term datasets: from descriptive to predictive data using ecoinformatics. **Journal of Vegetation Science**, v. 18, n. 4, p. 457-462, 2007. Disponível em: <http://onlinelibrary.wiley.com/doi/10.1111/j.1654-1103.2007.tb02559.x/ abstract>. Acesso em: 10 nov. 2010.

BLACKBURN, T. M. **Method in macroecology**. Disponível em: <http://wolfweb.unr.edu/~ldyer/classes/blackburn.pdf>. Acesso em: 20 abr. 2011.

CHARIF, R. A.; WAACK, A. M.; STRICKMAN, L. M. **Raven Pro 1.4 user`s manual**. Cornell Lab of Ornithology, Ithaca, NY. 2010. Disponível em: <http://www.birds.cornell.edu/brp/raven/RavenOverview.html>. Acesso em: 14 set. 2010.

DEL HOYIO, J.; ELLIOT, A.; CHRISTIE, D. A (Ed.). **handbook of the birds of the world**. 8 v. (Broadbills to Tapaculos). Barcelona: Lynx, 2003.

DUNNING, J. B (Ed.). **CRC hanbook of avian body masses**. 2 ed. Boca Raton: CRC, 2008.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. **Knowledge discovery and data mining**: towards a unifying framework. 1996. Disponível em: <http://www.aaai.org/home.html>. Acesso em: 10 out. 2010.

_____. **From data mining to knowledge discovery in databases**. 1997. Disponível em: <http://www.kdnuggets.com/gpspubs/aimag-kdd-overview-1996-Fayyad.pdf>. Acesso em: 10 out. 2010.

GLUSMAN, G. et al. The olfactory receptor gene superfamily: data mining, classification and nomenclature. **Mammalian Genome**, NY, 2000.

GOLDSCHMIDT, R.; PASSOS, E. **Data mining**: um guia prático. Rio de Janeiro: Elsevier, 2005.

GOTELLI, N. J. Perspectives in biogeography: hypothesis testing, curve fitting, and data mining in macroecology. **International Biogeography Society Newsletter**, v. 6, n. 3, p. 1-7, 2008.

LI, D.; DI, K.; LI, D. Land use classification of remote sensing image with GIS data based on spatial data mining techniques. **Archives of Photogrammetry and Remote Sensing**, Amsterdam, v. 33, parte b3, 2000.

NEVES, R. de C. D das. **Pré-processamento no processo de descoberta de conhecimento em banco de dados**. 2003. Dissertação (Programa de Pós-graduação em Computação) - Instituto de Informática, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2003.

TKALCIC, M. **csv2arff**. 2008. Disponível em: <http://slavnik.fe.uni-lj.si/markot/csv2arff/csv 2a'rff.php?do=instructions>. Acesso em: 11 abr. 2011.

XENO-CANTO. 2005-2010. Disponível em: <http://www.xeno-canto.org/>. Acesso em: 20 abr 2010.

## Data mining: the search for knowledge about the singing evolution of *Thamnophilidae* family

### Abstract

*Introduction: Describes the use of a data mining technique about the song, biology and micro-habitat of the Thamnophilidae bird family in order to find patterns which relate them. Method: A database was built in Excel ® spreadsheet listing 82 species of the family of the bird Thamnophilidae comprising various attributes related to bird calling features, biology and micro-habitat in which they are found. For the analysis it was used the algorithm APRIORI in the WEKA 3.7.1 software.. Results: The association of the different attributes of the 82 different species, considering 10% of minimum support and 90% of minimum confidence, allowed the rescued of 172 patterns, from which 42 contained one of the song's attributes: PC1 e PC2. The patterns which related the attribute PC2 were the most expressive ones due to its relation to the size and gender of the family. Conclusions: The experiment demonstrated that the algorithm could be better suited in larger databases and/or when the data standardization presents a lower number of categories, what could be a limitation in the macroecology field. Nonetheless, it has presented itself as an alternative instrument to the exploratory study of the relations among diverse attributes, which results could serve as objects for further analysis.*

### Keywords

*Datamining; Database; Forest birds; Thamnophilidae (bird); Bird songs.*

Recebido em 9 de maio de 2011
Aceito em 12 de maio de 2011

[1] Graduada em Administração - UEA, Mestre em Ciência, Gestão e Tecnologia da Informação - UFPR. Estudante de pós-graduação (Mestrado) - Bolsista PROF - UFPR.
leticia.csilva01@gmail.com

[2] Bacharel em Informática - UFPR, Mestre em Informática Industrial, Doutora em Engenharia Biomédica. Professor adjunto - UFPR/DeCiGI.
dtsunoda@ufpr.br

[3] Graduada em Ciências Biológicas, Mestre em Ecologia (INPA), Doutoranda pelo Programa de Pós-graduação em Ecologia e Conservação (UFPR). Estudante de pós-graduação (Doutorado) - UFPR.
viviane.deslandes@gmail.com

Como citar este artigo:

COSTA E SILVA, L. da; TSUNODA, D. F.; DESLANDES, V. Mineração de dados: busca de conhecimento sobre a evolução do canto da família Thamnophilidae. **AtoZ**, Curitiba, v. 1, n. 1, p. 61-70, jan./jun. 2011. Disponível em: <www.atoz.ufpr.br>. Acesso em: