Identificação de Espécies de Caramujos aplicando Aprendizado Baseado em Instâncias

Gustavo Silva Semaan[1], Marcelo Nocelle Almeida[1], Raphael Magno de Souza Lage[1], Marcos Vinicius Naves Bedo[1], Debora Alvernaz Corrêa[2], José André de Moura Brito[3]

[1] Instituto do Noroeste Fluminense de Educação Superior da Universidade Federal Fluminense (INFES/UFF) – Santo Antônio de Pádua, RJ, Brasil

[2]Centro Universitário Descomplica UniAmérica – Rio de Janeiro, RJ, Brasil

[3]Escola Nacional de Ciências Estatísticas do Instituto Brasileiro de Geografia e Estatística (ENCE/IBGE) – Rio de Janeiro RJ, Brasil

{gustavosemaan, mnocelle, marcosbedo}@id.uff.br

Abstract. *Decision-support systems benefit from hidden patterns (semi-)automatically extracted from massive digital data. In the specific domain of gastropod characterization, morphological data and measurements can support biologists in the identification of land snails by patterns. Although slugs and snails can be easily identified by their excretory and reproductive systems, the mollusk body is commonly inaccessible because of either soft material deterioration or shell fossilization. This study aims at investigating the behavior of a distance-based classifier algorithm over a handcrafted dataset of morphological features of land snail shells. Experimental evaluations indicate a fine-tuned distance-based classifier achieved a hit ratio up to 99% in the task of snail identification.*

Resumo. *Os sistemas de apoio à decisão beneficiam-se de padrões ocultos (semi-)extraídos automaticamente de grandes bases de dados. No domínio específico da caracterização de gastrópodes, dados e medidas morfológicas podem apoiar os biólogos na identificação de caramujos terrestres. Embora lesmas e caramujos possam ser facilmente identificados por seus sistemas excretório e reprodutivo, o corpo do molusco é comumente inacessível por causa da deterioração do material macio ou da fossilização da concha. O presente artigo tem, como objetivo, investigar a performance de um classificador baseado em instâncias, quando aplicado sobre um conjunto de dados com características morfológicas de conchas de caracóis terrestres. Avaliações experimentais indicam que um método proposto alcançou acurácia superior a 99% na identificação de espécies de caramujos.*

1. Introdução

O avanço da tecnologia da informação resulta no armazenamento crescente e intenso de dados em repositórios, que possibilita desde obtenção de informações simples até o apoio na tomada de importantes decisões. Para alcançar objetivos em que *simples* consultas a Banco de Dados não são suficientes, pode-se utilizar o processo de Descoberta de Conhecimento em Bases de Dados (do inglês *Knowledge-Discovery in Databases* (KDD)), que

aborda diversas áreas como aprendizado de máquina, estatística, otimização e inteligência artificial [Han et al. 2012][Tan et al. 2018].

Em um contexto geral, o reconhecimento de espécies é de grande importância para o conhecimento da biodiversidade. A ampliação desse conhecimento fornece elementos que podem ser utilizados como ferramentas para, por exemplo, a conservação das espécies ameaçadas de extinção ou mesmo para formular estratégias de controle de doenças transmitidas por animais. Em ambas as situações, a correta identificação das espécies é uma premissa fundamental [Prévot et al. 2013][Almeida et al. 2021].

Em particular, no que concerne à caracterização e identificação dos gastrópodes terrestres, toma-se por base, essencialmente, a análise de sua morfologia interna, sobretudo dos aparelhos excretor e genital dos animais. Entretanto, tanto em materiais depositados em coleções científicas quanto em coletas realizadas na natureza, as partes moles dos moluscos são comumente inacessíveis.

O objetivo geral desta pesquisa é propor um método capaz de identificar, automaticamente, algumas espécies de caramujos da Família *Subulinidae* com base, apenas, na morfologia de suas conchas. Para isso, foi utilizado o processo de KDD e suas etapas, desde o pré-processamento, os algoritmos de mineração de dados ao pós-processamento, com a apresentação dos resultados. O trabalho está estruturado da seguinte maneira: a seção 2 apresenta o problema da identificação de espécies de caramujos; a seção 3 apresenta a metodologia utilizada; a seção 4 relata os experimentos computacionais e por fim a seção 5 trata das considerações finais e as propostas para trabalhos futuros.

2. Identificação de Espécies de Caramujos

Conforme relatado na introdução, a identificação de espécies é fundamental para o conhecimento da biodiversidade, e pode ser utilizada como ferramenta no apoio à análise e resolução de diversos problemas, como conservação de espécies e controle de doenças. Por exemplo, as espécies da Família *Subulinidae* são potenciais hospedeiros intermediários de parasitos humanos e de animais domésticos [Almeida and Mota 2011]. Embora seja possível diferenciar as espécies consideradas neste trabalho com base na Figura 1 (pela cor, forma e/ou textura), de maneira geral, as conchas obtidas na natureza não estão preservadas como os exemplares de uma coleção [Almeida et al. 2021].

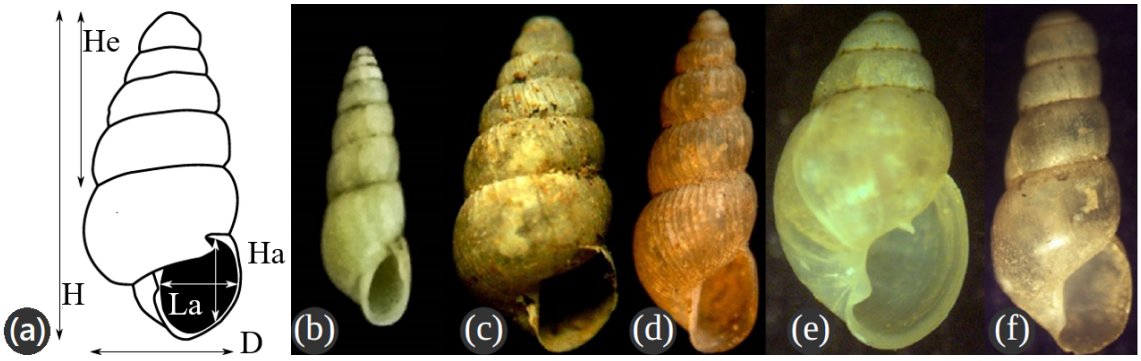

Figura 1. Atributos dos caramujos e espécies (adaptado de [Almeida et al. 2021]).

O trabalho utilizou o *dataset* `HELIX`[1], que possui 518 conchas, coletadas manu-

[1] *Dataset* `HELIX` disponível em https://doi.org/10.5281/zenodo.5500215.

almente em um período de um ano na região metropolitana da cidade de Juiz de Fora (MG). Os atributos indicados na Figura 1(a) são: Altura (H), Diâmetro (D), Altura da espira (He), Altura da abertura(Ha) e Largura da abertura (La). As conchas são de cinco espécies de caramujos terrestres, entre as mais comuns no Brasil, e possuem uma ampla variedade de indivíduos, com diferentes gêneros e idades. Seguem as espécies e seus quantitativos: (b) *Allopeas gracilis* (50); (c) *Beckianum beckianum* (149); (d) *Dysopeas muibum* (110); (e) *Leptinaria unilamellata* (62); e (f) *Subulina octona* (147).

A maioria dos gastrópodes terrestres possui uma concha externa secretada por uma superfície glandular (manto), composta por carbonato de cálcio [Leme 1995]. Os primeiros pesquisadores naturalistas utilizavam a concha, quase que exclusivamente, como caráter taxonômico para descrever as espécies [Colley et al. 2012]. Para grupos como bivalvas e gastrópodes marinhos, a maioria das espécies possuem conchas diferentes o suficiente para a diferenciação entre as espécies. Contudo, entre os gastrópodes terrestres ocorre o inverso, e as características externas, sobretudo a concha, podem não fornecer elementos suficientes e confiáveis para a distinção entre as espécies [Prévot et al. 2013].

De modo adicional, os gastrópodes pulmonados terrestres são altamente polimórficos, o que dificulta ainda mais a diagnose específica [Backeljau 2001]. Assim, caracteres exclusivamente conquiliológicos podem não ser suficientes para análise, sendo a morfologia interna a mais adequada para esse fim [Colley et al. 2012]. Porém, não há como estudar a morfologia interna do animal quando se dispõe apenas da concha, o que ocorre com frequência tanto em coletas na natureza quanto em coleções secas.

A taxonomia moderna direciona-se a uma abordagem integrativa, que considera aspectos como a morfologia da concha, das partes moles, estudos moleculares e citogenéticos, dados ecológicos e comportamentais. Em diversas famílias de moluscos gastrópodes terrestres as espécies são muito semelhantes. No Brasil ocorrem muitas espécies de caracóis que são muito semelhantes entre si, o que dificulta a identificação correta, sobretudo, quando se tem disponível apenas as conchas [Almeida and Mota 2011].

Pelo site Conquiliologistas do Brasil [CdB 2021] é possível selecionar duas espécies dentre as disponíveis para compará-las visualmente, onde também são disponibilizadas informações relevantes como o *habitat*, a alimentação, a frequência (ocorrência), o tamanho médio e os estados do Brasil em que se encontram. Por exemplo, a espécie *Leptinaria unilamellata* é encontrada nos estados de SP e RJ, possui tamanho médio de 10-13 mm, seu habitat é a terra e é encontrada em uma frequência alta (abundante).

3. Metodologia

A Figura 2 apresenta o método utilizado para as análises realizadas, modelado com o objetivo de organizar e executar diferentes configurações em subconjuntos de dados, denominados instâncias virtuais (IVs). Para isso, além de iterativamente ocorrerem operações de pré e pós-processamento, na etapa de Mineração é utilizado o algoritmo k-NN (do inglês k *Nearest Neighbor*) [Tan et al. 2018] [Han et al. 2012], que usa conceitos de aprendizado baseado em instâncias [Aha et al. 1991].

A Figura 2 (a) ilustra o conjunto de dados de entrada X. Ele é submetido a um pré-processamento que, mediante reduções verticais e horizontais, forma novas instâncias (IVs), cujo nome deve-se ao fato de serem armazenadas apenas na memória principal.

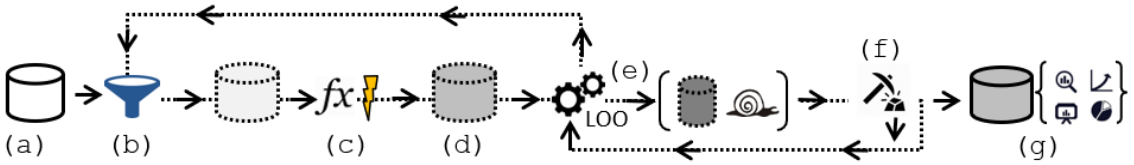

Figura 2. Metodologia utilizada com base no processo de KDD.

Cada IV corresponde a um subconjunto de dados $X^l \in X$, que possui um subconjunto de atributos (redução vertical) e/ou de registros (reduções horizontais 1 e 2) de X.

A etapa de pré-processamento possui basicamente três funções: (i) redução vertical (RV); e (ii) duas reduções horizontais (RH1 e RH2) (Figura 2 (b)); e (iii) a padronização dos dados da IV gerada (Figura 2 (c)). A RV consiste em gerar todas as combinações de atributos. A RH1 possibilita limitar a quantidade de exemplares por espécie da instância, com o intuito de balancear as classes das IVs. Dado o contexto de aprendizado baseado em instâncias (vizinhos), a classificação pode ser tendenciosa, favorecendo as espécies com maior quantidade de representantes. A RH2 ignora, iterativamente, cada uma das classes, com o objetivo de aumentar a taxa de acerto ao minimizar confusões entre individuos de classes que possuem características semelhantes.

Por fim, cada instância IV precisa ter seus dados padronizados. Formalmente, dado um conjunto $X = \{x_1, x_2, ..., x_n\}$ em que cada objeto x_i possui p atributos, $x_i = \{x_i^1, x_i^2, ..., x_i^p\}$, deve-se formar o conjunto Z (Figura 2 (d)) com valores adimensionais. Para isso, considera-se as Equações 1 e 2, em que as medidas μ_h e σ_h indicam, respectivamente, a média e o desvio padrão do atributo h.

$$\sigma_h = \sqrt{\sum_{i=1}^{n}(x_i^h - \mu_h)^2} \qquad h = 1, ..., p \tag{1}$$

$$z_i^h = \frac{x_i^h - \mu_h}{\sigma_h} \qquad h = 1, ..., p \qquad i = 1, ..., n \tag{2}$$

Após a formação e a padronização de uma dada instância IV, ela é submetida ao método *Leave-One-Out* - LOO (Figura 2 (e)), um caso especial do *k-fold cross-validation* cujo $k - fold$ possui tamanho um. Iterativamente, o LOO remove cada registro de IV ($Z^* = Z - z_i$) e submete Z^* ao algoritmo k-NN para que o registro z_i seja classificado (Figura 2 (f)). Após as execuções do algoritmo minerador, armazena-se a matriz confusão e a acurácia (Figura 2 (g)) [Han et al. 2012][Tan et al. 2018].

4. Experimentos Computacionais

Para os experimentos computacionais foi utilizado um computador dotado de um processador Intel i7 2,7GHz, com 8 GB de memória RAM e Sistema Ubuntu 14.04. As implementações foram realizadas utilizando softwares gratuitos, sejam eles: Java 8, IDE Eclipse Oxygen 3, MariaDB 10, phpMyAdmin e R 3.3. Conforme [Tan et al. 2018], o desempenho dos algoritmos baseados na regra do vizinho mais próximo pode ser mensurado com base: (i) na quantidade de objetos da instância; (ii) no tempo computacional necessário para classificar um objeto; e (iii) na acurácia obtida.

Uma vez que o classificador não armazena informações que podem ser reaproveitadas, ele pode demandar tempo um computacional razoavelmente alto para classificar um dado objeto em instâncias de grande porte. Especificamente neste trabalho, os tempos computacionais não foram relatados, mas para a instância completa (HELIX) e um dado caramujo, sua identificação ocorre na ordem de milésimos de segundo. Em relação aos cálculos das acurácias, foi utilizado o LOO, conforme apresentado na seção 3.

Em um experimento preliminar, com o intuito de avaliar se as classes (como grupos) são homogêneas, com base na ideia de uma iteração do clássico algoritmo k-Means [Jain 2010], foi calculado o centroide de cada espécie. Em seguida, foi verificado para cada caramujo a qual centroide ele está mais próximo e, quando esse centroide corresponde a sua própria classe, atribui-se um acerto. O percentual de acerto geral foi de 76,5% (122 erros), embora as especies *Beckianum beckianum* e *Dysopeas muibum* tenham se destacado com 99,3% (1 erro) e 98.2% (2 erros), respectivamente. De fato, em uma Análise de Componentes Principais apresentada em [Almeida et al. 2021], essas classes estão bem separadas, enquanto os objetos das demais classes estão amontoados. A classe *Subulina octona* obteve o pior resultado, com apenas 40,1% de acerto (88 erros).

O clássico algoritmo k-NN pode ser aplicado na classificação de objetos (caramujos) desconhecidos, baseando-se na comparação com objetos semelhantes que foram previamente classificados. Assim, ele verifica as classes (espécies) dos k objetos mais próximos (mais semelhantes ou menos diferentes) com base na distância Euclidiana, e identifica a espécie do objeto submetido à classificação.

De acordo com a metodologia ilustrada na Figura 2, os parâmetros utilizados foram: (i) Análise de vizinhos: $k \in \{1, 2, ..., 9, 10\}$; (ii) RV: fazer todas as combinações entre os cinco atributos dos caramujos da instância HELIX (31 combinações); (iii) RH1: quantidade de exemplares por espécie limitadas em $\{10, 20, 30, 40, 50\}$, selecionados de maneira aleatória; (iv) RH2: Ignorar uma classe: considera todas as espécies e também ignora cada uma das 5 espécies; (v) Padronização: devido às RV e RHs, a padronização deve ocorrer no processo para cada IV. A partir do *dataset* HELIX, as combinações de parâmetros resultaram em cerca de 9,5 mil Instâncias Virtuais.

Com base nos cerca de 95 mil resultados gerados a partir das IVs, da aplicação do LOO e dos valores de k, análises são apresentas em diferentes perspectivas. A Tabela 1 apresenta os resultados com a RH2. Nesse sentido, ao remover (ignorar) os registros busca-se reduzir as *confusões* entre pares de classes semelhantes. Destaca-se o aumento na acurácia em todas as colunas ao ignorar, separadamente, as classes com Ids 2, 3 e 5. Observa-se, também, reduções na coluna Mínimo ao ignorar as espécies 1 e 4, que são identificadas com mais facilidade [Almeida et al. 2021]. A Figura 3 apresenta um gráfico *boxplot* referente também às acurácias com o uso da RH2.

Ainda com base na Tabela 1, o total de caramujos por espécies está desbalanceado. A Figura 3 apresenta um gráfico *boxplot* referente aos percentuais de acerto com o uso da RH1, em que diferentes quantitativos de indivíduos por espécie foi utilizado. Com base nesse gráfico, com o aumento da quantidade de registros por classe ocorre também aumento no percentual de acerto. Foi estabelecido, então, o uso da quantidade máxima de registros por espécie o valor da quantidade da espécie com menos indivíduos na instância original (neste experimento 50 unidades). Destaca-se, novamente, que os indivíduos são

Classes de HELIX			Acurácia (%)		
Id	Espécie ignorada	Qtde	Max.	Min	Média
1	*Beckianum beckianum*	149	97,3	*59,6*	86,3
2	*Dysopeas muibum*	110	99,5	96,3	98,2
3	*Allopeas gracilis*	50	99,8	98,5	99,1
4	*Leptinaria unilamellata*	62	98,5	*71,7*	89,8
5	*Subulina octona*	147	99,7	97,8	98,7

Tabela 1. Acurácia do k-NN com a RH2.

selecionados de maneira aleatória e, por isso, houveram 10 execuções do algoritmo.

A Tabela 2 apresenta as acurácias quando não houve RH, mas foi aplicada a RV. É possível observar acurácias altas, com o valor máximo de 99,4%, e também na coluna mínimo, com valores superiores a 95%. Isso indica que, nesse caso em específico, é possível reduzir o tempo de processamento ao utilizar apenas um atributo. De modo adicional, a acurácia aumentou com a redução do valor de k.

	Acurácia (%)		
k	Max.	Min.	Média
1	99,4	97,5	99,2
2	99,4	97,5	99,2
3	98,8	96,7	98,6
4	98,6	96,7	98,6
5	97,9	96,3	97,8
6	97,9	96,3	97,8
7	97,3	95,9	97,2
8	97,3	95,9	97,2
9	96,9	95,8	96,6
10	96,7	95,6	96,6

Tabela 2. Acurácia do k-NN com a RV (combinações de atributos).

Com o intuito de fornecer um panorama geral dos resultados, gráficos *boxplot* são apresentados na Figura 3. Nessa figura o *boxplot* (a) refere-se à acurácia obtida com base em cada classe ignorada (RH2), também quando alguma classe foi ignorada. Pode-se observar que, ao desconsiderar os registros da classe 1, os resultados foram inferiores, com mediana inferior a 90% de acurácia. Em contrapartida, ao ignorar os registros das classes 2 e 3, as acurácias medianas foram próximas a 100%.

A Figura 3 (b) destaca os melhores resultados obtidos ao balancear as classes, estabelecendo como valor a quantidade de indivíduos da classe com menos animais. Por fim, a Figura 3 (c) apresenta um *boxplot* referente à acurácia para todos os experimentos realizados e da melhor configuração identificada: valores de $k \in \{1, 2, \ldots, 9, 10\}$, não ignorar nenhuma classe e 50 registros por classe nas instâncias virtuais.

5. Conclusões e Trabalhos Futuros

Conforme apresentado no trabalho, a identificação de gastrópodes terrestres baseia-se, principalmente, na morfologia interna, com foco no aparelhos excretor e genital dos ani-

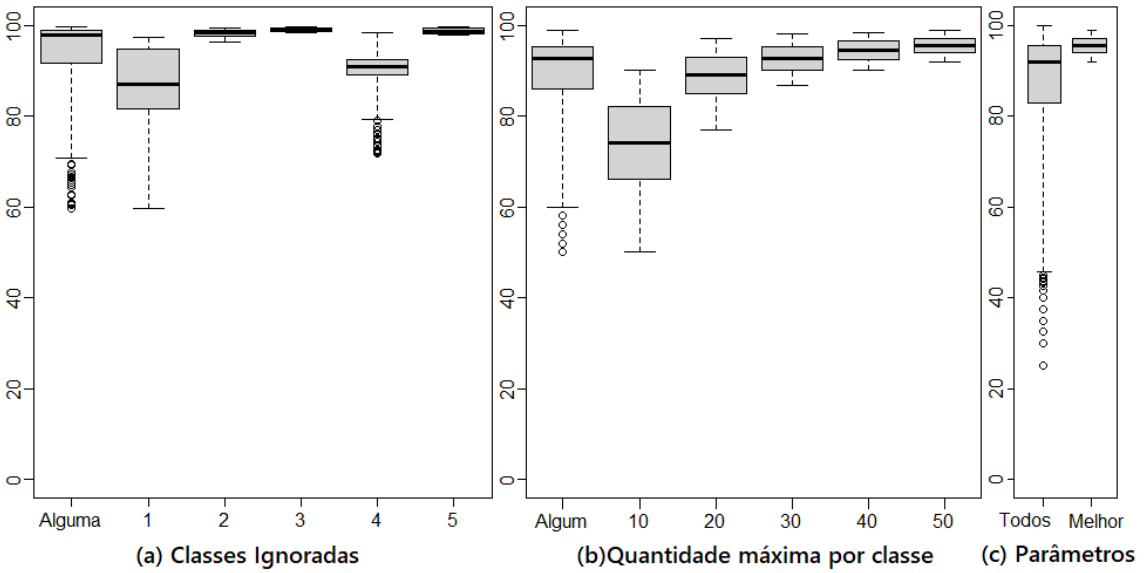

Figura 3. Gráficos *boxplot* - acurácias para RH2, RH1 e panorama geral.

mais. Entretanto, em coleções científicas e nas coletas realizadas na natureza, dificilmente as partes moles dos moluscos estão disponíveis. Assim, como objetivo geral foi utilizado um método baseado no processo de KDD capaz de identificar, automaticamente, a espécie de caramujos com base apenas na morfologia de suas conchas.

A partir do *dataset* `HELIX` (seção 2) que possui 518 conchas de cinco espécies e cinco atributos, o método utilizado gerou aproximadamente 9,5 mil Instâncias Virtuais, e cerca de 95 mil resultados de experimentos computacionais. A classificação de um dado exemplar no *dataset* original (completo) ocorre na ordem de milésimos de segundo e com acurácia média superior a 99,1%. É importante destacar a simplicidade do método, que utiliza algoritmos clássicos da literatura e a possibilidade de aplicação em problemas semelhantes. De fato, a proposta remete aos primeiros naturalistas que utilizavam as conchas para descrever as espécies, em uma abordagem que utiliza um ferramental da Tecnologia da Informação.

Com base nos experimentos realizados e nos resultados obtidos, pode-se concluir que o método proposto é uma alternativa eficiente e eficaz para a resolução do problema. Como trabalhos futuros pretende-se realizar experimentos: (i) com dados de outras espécies de caramujos da mesma família; e (ii) utilizar o algoritmo de agrupamento DBSCAN (*Density-Based Spatial Clustering of Applications with Noise*)[Tan et al. 2018], uma vez identificado o sucesso na classificação com o uso do k-NN e os altos percentuais de erros relatados no experimento preliminar com o k-Means, o que indicam que os grupos (classes) podem ter formatos arbitrários.

Referências

Aha, D. W., Kibler, D., and Albert, M. K. (1991). Instance-based learning algorithms. *Mach. Learn.*, 6(1):37–66.

Almeida, M. and Mota, G. (2011). Conquiliomorfometria, ciclo de vida, crescimento alómetrico da concha (subulina octona bruguière, 1789) (pulmonata, subulinidae) em condições de campo. *Biofar*, 5(1):141–151.

Almeida, M., Olmes, L., Semaan, G., Oliveira, D., Santos, L., and Bedo, M. (2021). Helix: A data-based characterization of brazilian land snails. In *Brazilian Symposium on Databases (SBBD)*.

Backeljau, T.; Baur, B. (2001). The biology of terrestrial molluscs / edited by g.m. barker. In *The biology of terrestrial molluscs*, pages 383–412. CABI Pub.

CdB (2021). Conquiliologistas do Brasil. www.conchasbrasil.org.br. Acesso: 21/10/2021.

Colley, E., e Simone, L., and Silva, J. (2012). Uma viagem pela história da malacologia. *Estudos de Biologia*, 34(83).

Han, J., Kamber, M., and Pei, J. (2012). *Data mining concepts and techniques, third edition*. Morgan Kaufmann Publishers, Waltham, Mass.

Jain, A. K. (2010). Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, 31(8):651 – 666.

Leme, J. (1995). Sistemática e biogeografia. In Barbosa, F., editor, *Tópicos em malacologia médica*, pages 12–49. Editora Fiocruz, Oxford.

Prévot, V., Jordaens, K., Sonet, G., and Backeljau, T. (2013). Exploring species level taxonomy and species delimitation methods in the facultatively self-fertilizing land snail genus rumina (gastropoda: Pulmonata). *PLOS ONE*, 8(4):1–18.

Tan, P.-N., Steinbach, M., Karpatne, A., and Kumar, V. (2018). *Introduction to Data Mining (2nd Edition)*. Pearson, 2nd edition.