REDE
Revista Expressão Da Estácio

ARTIGO **REDE – 2019; 1:43-51.**

APLICAÇÃO DA ANÁLISE DE AGRUPAMENTOS EM DADOS DE COLEÇÕES CIENTÍFICAS BOTÂNICAS /

APPLICATION OF CLUSTER ANALYSIS IN DATA OF SCIENTIFIC BOTANICAL COLLECTIONS

Silva, L. A. E.[1,2]; Oliveira, F. A. [1];
Ribeiro, R.[1]; Costa, L. F. S.[2];
Souza, L. C. S.[2]

[1]Instituto de Pesquisas Jardim Botânico - Rio de Janeiro/RJ (IPJBRJ)
[2]Universidade Estácio de Sá (UNESA)

Resumo

Sistemas de informação para gerenciamento de coleções biológicas científicas são fundamentais para o conhecimento da natureza e produção de pesquisas. Estes sistemas evoluíram para tornar disponíveis os dados e imagens de exsicatas (amostras desidratadas) e suas coleções relacionadas. Neste trabalho é apresentada uma abordagem de análise de agrupamentos para a identificação de outliers na coleção de campo de coordenadas geográficas. O método proposto foi aplicado com sucesso no sistema de informações Jabot, usado para gerenciar as coleções científicas do herbário do Instituto de Pesquisas Jardim Botânico do Rio de Janeiro, que é o maior banco de dados da flora nacional.

Palavras-chave: coleções científicas biológicas, banco de dados de herbários, análise de agrupamentos.

Abstract

Information systems for scientific biological collections management are key to the knowledge of nature and research production. These systems have evolved to make available the data and images of plant specimens (samples dehydrated) and its related collections. In this paper is presented an approach of cluster analysis for the identification of outliers in field collection of geographical coordinates. The proposed method has been successfully applied in the Jabot, information system used to manage the scientific collections of the Herbarium of the Instituto de Pesquisas Jardim Botânico do Rio de Janeiro, which is the largest database of national flora.

Keywords: *scientific biological collections, collections database, cluster analysis*

Introdução

Bancos de dados da biodiversidade são cada vez mais consultados para diversos estudos e pesquisas. Eles auxiliam na produção de conhecimento sobre a flora, facilitam o monitoramento e a elaboração de ações de conservação na biodiversidade (Pougy *et al.*, 2014), na produção de modelos preditivos de distribuição de espécies (Barros *et al.,* 2012), na geração

de listagens de plantas ameaçadas de extinção (Martinelli e Moraes 2013; Martinelli *et al.*, 2014), na análise de co-ocorrências de espécies (Silva *et al.,* 2016), entre outras possibilidades. Objetivando facilitar o acesso e a integração desses acervos, diversos sistemas de informação específicos para o gerenciamento de herbários e suas coleções científicas foram desenvolvidos nos últimos anos (Neto *et al.,* 2013). Por meio desses sistemas, o conteúdo das coleções botânicas torna-se disponível para muitos pesquisadores e estudantes, além de promover a conservação do acervo quanto ao manuseio, tendo em vista que as imagens das amostras são disponibilizadas em alta resolução. As coleções botânicas, mantidas em herbários, são compostas por amostras que são testemunhos da ocorrência natural de espécies e, certificam a riqueza da flora de uma determinada região (Bridson & Forman 1992, Peixoto & Morim 2002).

O herbário do Instituto de Pesquisas Jardim Botânico do Rio de Janeiro mantém a maior coleção de amostras de plantas do país (Peixoto & Morim 2002) e, para gerenciar esse acervo, foi desenvolvido o sistema de gerenciamento de coleções científicas *Jabot*[1] (Silva, 2017). A primeira versão do *Jabot* foi desenvolvida entre 2003 e 2005 na própria instituição (Gonzalez, 2009) utilizando *software* livre, com a linguagem de programação PHP e o sistema gerenciador de banco de dados *postgreSQL*. O sistema auxilia na curadoria dos acervos do herbário, que se encontra totalmente informatizado e digitalizado, e das coleções vivas. Seu uso facilita o gerenciamento dos serviços tradicionalmente prestados por herbários, além do armazenamento e disponibilização de dados e imagens de exsicatas *online*. Atualmente o banco de dados tem aproximadamente 760.000 registros e um volume médio anual de entrada de novas amostras no banco na ordem de 30.000 coletas, apesar do volume, observa-se, porém, a necessidade de uma melhora na qualidade dos dados. Dessa forma foi realizado um estudo consistente sobre os tipos de erros presentes nesses bancos de dados.

A análise da qualidade de dados é fundamental para o desenvolvimento de pesquisas científicas. Dados com qualidade permitem que os resultados dos trabalhos sejam mais facilmente alcançados e com melhor qualidade. Com isto em foco, existe no *Jabot* um módulo específico para a identificação e tratamento da qualidade de dados. Ele foi desenvolvido com base em estudos anteriores sobre qualidade de dados em coleções botânicas (Chapman 2005a; Silva *et al.,* 2010; Dalcin *et al.,* 2012). O módulo é dividido em três categorias, que foram criadas com base nos tipos de erros comumente encontrados nas coleções botânicas, que são os erros taxonômicos, de georreferenciamento e nos dados de identificação das coletas. O *Jabot* não

[1] http://jabot.jbrj.gov.br

avalia apenas o nível de qualidade de dados existentes. Há uma preocupação em impedir a entrada de novos erros, diminuindo o tempo em atividades de limpeza de dados (Chapman 2005b). Para tal, foi desenvolvida um sistema de importação de planilhas[2]. Nesse componente do sistema, 81 validações são realizadas antes da importação, indicando os erros encontrados nos dados de coletas como, por exemplo, o comparativo dos táxons em relação à lista de espécies da Flora 2020 (Forzza 2010), adotada como dicionário oficial.

Quanto à localização geográfica, a confiabilidade dos dados passou a ser fundamental para os botânicos, permitindo monitorar e modelar o ambiente da distribuição geográfica dos indivíduos. Dessa maneira, buscando diminuir a possibilidade de erros de digitação dos atributos geoespaciais, a entrada de novos registros botânicos no *Jabot* passou a ser feita com o uso de filtros que definem a pertinência de valores para latitude, longitude e denominação de municípios. Tais filtros fazem a comparação dos valores informados pelo usuário com as coordenadas dos limites municipais constantes da base vetorial BC250-IBGE2014. Os limites municipais da referida base fizeram parte do escopo do Projeto de atualização permanente da Base Cartográfica Contínua do Brasil na escala 1:250.000 (BC250) (Azevedo & Neto 2011), um conjunto de dados geoespaciais de referência, estruturados em bases de dados digitais, permitindo uma visão integrada do território nacional. Para a implementação das validações geográficas, o sistema *Jabot* utiliza a extensão geoespacial *PostGIS* com o *PostgreSQL* para o suporte ao tratamento e análise de dados espaciais (Urbano & Cagnacci 2014). Em sistemas que envolvem dados associados à localização geográfica, a integração do *PostGIS* com sistemas de informações geográficas (SIG) (Longley, 2013) e *WebGIS* possibilita diversos recursos para consulta, visualização e análise geoespacial. Na Tabela 1 são apresentados os tipos de erros identificados nos dados geográficos e suas possíveis causas.

Objetivos

Neste trabalho é apresentada uma abordagem de análise de agrupamentos para a identificação de outliers na coleção de campo de coordenadas geográficas.

[2] http://jabot.jbrj.gov.br/v2/validarplanilha_externo.php

Tabela 1. Descrição dos erros encontrados e das causas.

Erro	Causa do erro
a) Coordenadas da coleta apresentam-se com o valor zero ('Lat/Long = zero').	1) Coletas históricas; 2) A coordenada não foi preenchida no banco de dados; 3) Não consta na ficha de identificação, apesar da coleta ser recente
b) Coletas no mar para espécimes terrestres ou na terra para espécimes marinhas.	1) A latitude ou longitude foi preenchida de forma incorreta no banco de dados; 2) A descrição da localidade não foi descrita de maneira clara; 3) A coordenada foi preenchida de forma incorreta na ficha; 4) As coordenadas encontram-se invertidas
c) Coleta em região de fronteira terrestre (municípios, estados ou países) ou fronteira terrestre - marítima.	1) Baixa precisão das coordenadas; 2) Localidade imprecisa no ato da coleta da coordenada
d) Coordenadas apresentam-se incompletas, insuficientes e erradas.	1) Preenchimento de latitude e longitude errada no banco de dados ou na ficha de descrição da coleta
e) O nome de país, estado ou município não corresponde com a coordenada.	1) A coordenada foi preenchida de forma incorreta; 2) Atualização do banco de dados para os processos de emancipação de estado ou município; 3) Preenchimento incorreto no banco nos campos de país/UF e município

Material e Métodos

Material

Com a possibilidade de os dados da biodiversidade apresentarem um alto número de erros devido aos fatores apresentados anteriormente e, com o aumento do volume de dados; verificou-se a necessidade de técnicas para agilizar a identificação dos pontos duvidosos nas coletas. Assim, foi considerado o uso da mineração de dados (Han *et al.,* 2011), por meio da análise de agrupamentos (Han *et al.,* 2001) e com o algoritmo *k-means*. Assim, de modo a avaliar o método desenvolvido nesse trabalho, foi considerada a premissa na qual as coletas realizadas por um coletor e, em um determinado dia, devem normalmente, estarem próximas. Tendo em vista que um coletor não tem tempo e condições para se deslocar em longas distâncias em um

período curto. Consequentemente, as coletas muito distantes são consideradas como suspeitas e indicadas para revisão.

Método proposto

O método proposto consiste na aplicação do algoritmo da análise de agrupamentos *k-means* à unidade formada por um conjunto de coletas de um coletor específico, de acordo com a premissa detalhada na Seção Material. Para a avaliação da eficiência do método, um teste aplicado a uma base contendo 21 coletas foi realizado com dados reais presentes na base de dados do *Jabot*. Os detalhes da aplicação do método proposto são apresentados na Figura 1.

Algoritmo detalhando o método proposto para a análise dos dados de coletas

1: **Defina** *parâmetros* <- *nome_coletor*, *dia_coleta*, *mes_coleta*, *ano_coleta*;
2: **Enquanto** não-fim tabela
  2.1: **Se** existem registros nos parâmetros indicados **então**
    2.1.1: **Obtenha** coordenadas {*latitude*, *longitude*} para *parâmetros* indicados;
    2.2.1: **Leia** próximo registro;
  2.2: **Fim-se**;
3: **Fim-enquanto**;
4: **Transforme** registros em data.frame;
5: **Defina** o número de *clusters* para a aplicação do *k-means*;
6: **Efetue** a aplicação do *k-means* para o nr de *clusters* informados;
7: **Verifique** o centróide dos agrupamentos;
8: **Calcule** a distância euclidiana: dist.euclid<- function (long1, lat1, long2, lat2)
  8.1: **Se** dist.euclid > 5 km **então**
    8.1.1: **Indique** a coleta como suspeita;
  8.2: **Fim-se**;

Figura 1. Detalhamento do método proposto com o algoritmo *k-means*

O método foi implementado com o uso *software* R (R Development Core Team 2017), i ntegrado ao PostgreSQL. A análise não-supervisionada aqui aplicada tem justificativa no grand e volume de dados analisado. Como resultado da aplicação do algoritmo, foram encontrados 3 *clusters*, apresentados na Tabela 2. Observa-se claramente que o *cluster* de número 1 destoa q uanto ao valor da longitude. O valor aproximado da distância desse *cluster* para os outros dois fica na distância de 4966,9 quilômetros.

Alguns fatores podem justificar tal erro, mas a principal possibilidade é um simples err o de digitação. Porém, como consequência, esse dado permanece no banco até o momento qu e um pesquisador atente para essa grande diferença. Esse dado pode ainda ser descartado em um processo criterioso de seleção de dados para pesquisas, caso o sistema que gerencia os da dos da coleção científica não possua

Tabela 2. Valores dos clusters encontrados com o *k-means*

Cluster	Latitude	Longitude
1	-6.058889	-5.246667
2	-6.058889	-50.129722
3	-6.058889	-50.277824

rotinas para a melhora da qualidade de dados. A Figura 2 apresenta o resultado da aplicação do *k-means* e a Figura 3 destaca a distância do *cluster* 1 para os demais.

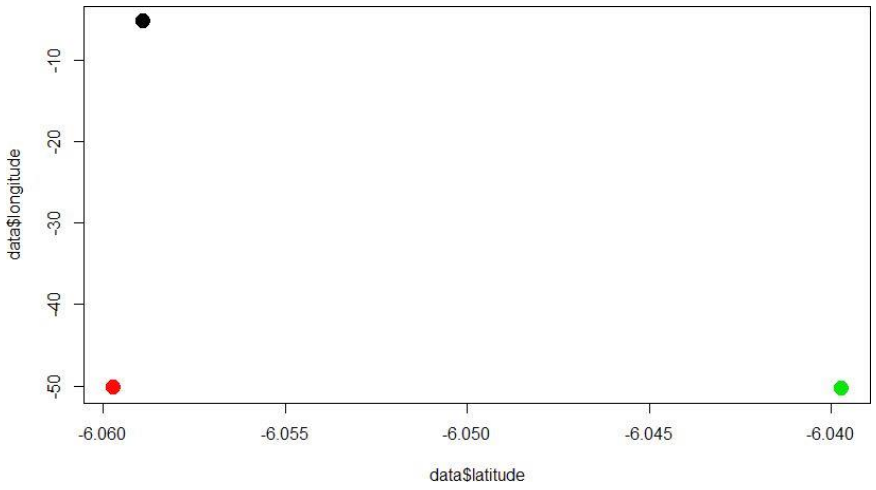

Figura 2. Resultado da aplicação do k-means com 3 clusters.

Conclusão

Diante dos grandes desafios para a conservação dos seus recursos naturais, a geração de conhecimento de grandes bases de dados da biodiversidade tem despertado a atenção da comunidade científica, pois bancos de dados da biodiversidade são fundamentais para pesquisas taxonômicas e de conservação, entre outras. Tais aplicações possibilitam conclusões aos

Figura 3. Mapa dos centróides das coletas.

tomadores de decisão acerca das políticas e ações que devem ser colocadas em prática para a conservação dos ecossistemas. Neste artigo foi apresentado um estudo e a proposta de um método a ser aplicado em dados de coletas presentes em bancos de dados de herbários, com o objetivo de detectar coletas com coordenadas suspeitas. O método inovou quando aplicou o tradicional algoritmo *k-means*, da análise de agrupamentos, considerando, como unidade de estudo o nome do coletor associado ao local da coleta, em um dia específico.

Foi possível perceber que a mineração de dados pode contribuir eficientemente para identificar coletas suspeitas, em largos volumes de dados, com um maior grau de dificuldade de identificação, o que dificilmente poderia ser realizado com outros métodos.

Em relação a trabalhos futuros, considera-se o estudo comparativo para utilização de outros algoritmos da análise de agrupamentos buscando uma possível maior eficiência; a avaliação da aplicação com outros níveis de granularidade e, ainda; com a aplicação a partir de outros atributos. Considera-se também aplicar o método na detecção e alerta no momento do cadastro das coletas buscando evitar novas entradas suspeitas no banco de dados.

Agradecimentos

Agradeço ao Programa Pesquisa Produtividade da Universidade Estácio de Sá pelo apoio.

Referências Bibliográficas

AZEVEDO, J. B., NETO, W. J. S. **Índice de Nomes Geográficos** - Base Cartográfica Contínua do Brasil ao Milionésimo - BCIM. 2011.

BARROS, F. S. M., SIQUEIRA, M. F., COSTA, D. P. Modeling the potential geographic distribution of five species of Metzgeria Raddi in Brazil, aiming at their conservation. **The Bryologist**, v. 115, n. 2, p. 341–349. 2012.

BRIDSON, D., FORMAN, L. **The Herbarium Handbook.** 3rd. ed. London: Royal Botanic Gardens, Kew. 1992.

CHAPMAN, A. D. **Principles of Data Quality.** 2005a. http://imsgbif.gbif.org/CMS_ORC/?doc_id=3135&download=1.

CHAPMAN, A. D. **Principles and Methods of Data Cleaning**: Primary Species e Species-Occurrence Data. 2005b. http://www2.gbif.org/DataCleaning.pdf.

DALCIN, E. C., SILVA, L. A. E., ZIMBRÃO, G., *et al.* Data Quality Assessment at the Rio de Janeiro Botanical Garden Herbarium Database and Considerations for Data Quality Improvement. In In: **8th International Conference on Ecological Informatics**. 2012.

FORZZA, R. C. *et al.,* **Lista de Espécies da Flora do Brasil.** 2010. http://floradobrasil.jbrj.gov.br, (accessed on Feb 12).

GONZALEZ, M. **Quantificação de custo e tempo no processo de informatização das coleções biológicas brasileiras**: a experiência do herbário do Instituto de Pesquisas Jardim Botânico do Rio de Janeiro. Rodriguésia, v. 60, p. 1–11. 2009.

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

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

LONGLEY, P. A. *et al.,* **Sistemas e ciência da informação geográfica**. 2. ed. Porto Alegre: Bookman. 2013.

Martinelli, G., Messina, T., Filho, L. S. **Livro Vermelho das Plantas do Cerrado**. Rio de Janeiro: Instituto de Pesquisas Jardim Botânico do Rio de Janeiro. 2014.

MARTINELLI, G., MORAES, M. **Livro vermelho da flora do Brasil**. Rio de Janeiro: Instituto de Pesquisas Jardim Botânico do Rio de Janeiro. 2013.

NETO, P. C. G., LIMA, J. R., BARBOSA, M. R. V., BARBOSA, M. A., MENEZES, M. **Manual de Procedimentos para Herbários**. Recife: Editora Universitária - UFPE. 2013.

PEIXOTO, A. L. AND MORIM, M. P. Coleções Botânicas: documentação da biodiversidade brasileira. **Flora**, v. 55, n. 3, p. 21–24. 2002.

POUGY, N., MARTINS, E., VERDI, M., *et al.,* Urban forests and the conservation of threatened plant species: the case of the Tijuca National Park, Brazil. **Natureza & Conservação**, v. 12, n. 2, p. 170–173. 2014.

R DEVELOPMENT CORE TEAM. R: A Language and Environment for Statistical Computing. In R Foundation for Statistical Computing Vienna Austria. , Tertiary **R: A language and**

environment for statistical computing. R Foundation for Statistical Computing. 2017. http://www.r-project.org, (accessed on Nov 4).

SILVA, L.A.E.; FRAGA, C.N.; ALMEIDA, T.M.H.; GONZALEZ, M.; LIMA, R.O.; ROCHA, M.S.; BELLON, E.; RIBEIRO, R.S.; OLIVEIRA, F.A.; CLEMENTE, L.S.; MAGDALENA, U.R.; MEDEIROS, E.V.S.; FORZZA, R. C. Jabot - Botanical Collections Management System: the experience of a decade of development and advances. **Rodriguésia - Revista do Jardim Botânico do Rio de Janeiro**, v. 68, n. 2, p. 391–410. 2017.

SILVA, L. A. E., BARROS, R. O., DALCIN, E., ZIMBRÃO, G. S., SOUZA, J. 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. 2010.

SILVA, L. A. E., SIQUEIRA, M. F., PINTO, F. S., *et al.,* Applying data mining techniques for spatial distribution analysis of plant species co-occurrences. **Expert Systems with Applications**, v. 43, p. 250–260. 2016.

URBANO, F., CAGNACCI, F. **Spatial database for GPS wildlife tracking data**—a practical guide to creating a data management system with PostgreSQL/PostGIS and R. 1. ed. Switzerland: Springer International Publishing. 2014.