CLUSTERMAP: PLUGIN DE VISUALIZAÇÃO DE DADOS MULTIVARIADOS EM MAPAS COROPLÉTICOS

Tiago P. Silvano[1], Bryan M. Correa[1], Philipe Borba[2], Ivanildo Barbosa[1]

[1]Seção de Ensino de Engenharia Cartográfica - Instituto Militar de Engenharia (IME)
Rio de Janeiro – RJ – Brazil

{tiago.silvano,bryan.correa,ivanildo}@ime.eb.br

[2] Instituto de Geociências – Universidade de Brasília (UnB)
Brasília – DF – Brazil

philipe.borba@unb.br

Abstract. *The ClusterMap plugin was developed to be used in the QGIS environment to create clusters from multivariate numerical data regarding georeferenced features. Some clustering methods were embedded to enable user to visualize the spatial distribution of the features within each resulting cluster, as well as to analyze their behavior based on decision rules. It is also available a resource to suggest to the user the optimal number of clusters, in the case he finds it useful. User is free to combine how many numeric variables to consider, how many clusters must be created as well as the clustering methods to use, and receive a new choropletic map, quality indicators for clusters, and decision rules.*

Resumo. *O plugin ClusterMap foi implementado para o ambiente QGIS com o objetivo de criar clusters a partir de um conjunto de variáveis, do tipo numérico, associadas a objetos georreferenciados. Foram implementados diferentes métodos de clusterização, permitindo ao usuário visualizar a distribuição espacial dos clusters formados e entender, por meio das regras de decisão, quais as características de cada um deles. Foi disponibilizada uma funcionalidade para sugerir ao usuário o número ótimo de clusters, caso ele julgue útil. O usuário possui a liberdade de testar a quantidade de variáveis, a quantidade de clusters e o método de clusterização, obtendo um novo mapa coroplético, indicadores de qualidade de clusters e regras de decisão.*

1. Introdução

O objetivo deste trabalho é descrever as funcionalidades de um *plugin*, desenvolvido para o ambiente QGIS, que efetua a clusterização a partir de um conjunto de dados georreferenciados, em formato vetorial, com variáveis numéricas, entregando ao usuário um mapa coroplético baseado nos grupos gerados. No rodapé desta página há um *link* para um vídeo que demonstra as funcionalidades da solução desenvolvida.

A motivação para este trabalho é a importância dos mapas coropléticos para analisar a distribuição espacial de uma variável. Uma opção para analisar múltiplas variáveis simultaneamente é a produção de mapas em miniatura (*small multiple maps*), que permitem ao usuário extrair alguma informação das variáveis representadas lado a

Link para video demonstrativo https://youtu.be/XU8C9e9WNd8

lado. Entretanto, para analisar múltiplas variáveis simultaneamente, os autores optaram por empregar a clusterização para agrupar os objetos de acordo com a similaridades de seus atributos, ou seja, objetos de um mesmo *cluster* são similares entre si, ao mesmo tempo que se distinguem de objetos de *clusters* vizinhos.

Um exemplo de análise viabilizada com o *plugin* desenvolvido é a comparação entre indicadores de composição demográfica, atividade econômica ou a evolução do número de notificações de alguma patologia agregados por municípios. A análise pode identificar padrões de dependência espacial de um conjunto de variáveis caso as feições pertencentes a um *cluster* sejam vizinhos [Silvano, Correa e Barbosa 2020].

Alguns requisitos foram elencados a fim de aprimorar a experiência do usuário:

a) Os parâmetros de cada método são informados pelo usuário: método de agrupamento, métrica de distância, medidas de dissimilaridades, e número de *clusters*;
b) A camada gerada é carregada no QGIS automaticamente, categorizada pelo número do *cluster* associado a cada feição;
c) A fim de subsidiar a interpretação do significado implícito na formação de cada *cluster*, foram extraídas regras baseadas em árvores de decisão, que podem ser interpretadas pelo usuário ou utilizadas para a geração de legenda;
d) É possível calcular o número ótimo de *clusters*;
e) É possível calcular as larguras médias de silhueta, geral e por *cluster*;

A combinação dos parâmetros, a definição do número de classes e a interpretação dos resultados fica a cargo do usuário, de modo que o *plugin* seja apenas uma ferramenta de apoio, flexível para aplicações em diversas áreas de atuação.

2. Implementação do *Plugin*

O *plugin* desse projeto foi desenvolvido para o software *QGIS*. O código implementado na linguagem *Python* com base no *Processing Framework* buscando mesmos padrões e funcionalidades dos algoritmos de processamento do *QGIS.* As principais bibliotecas *Python* utilizadas nesse projeto são a biblioteca *Sklearn* para implementação dos métodos de agrupamento e classificação, *Numpy* para o processamento de *array* e gerenciamento de matrizes, *Matplotlib* com o objetivo de gerar gráficos bidimensionais para análise do usuário.

As interfaces foram customizadas com auxílio do *Qt Designer* e da biblioteca *PyQt,* conexão entre *Python* e o *Qt,* de forma a atender as funcionalidades do *plugin.*

3. Funcionalidades do *Plugin*

O *plugin* permite ao usuário a escolha de dois métodos de agrupamento, um não hierárquico (*k-means)* e um hierárquico aglomerativo. Os valores dos parâmetros são passados pelo usuário para cada método por intermédio de interfaces próprias.

Caso o usuário opte pelo método *k-means,* o *plugin* permite a escolha da camada vetorial dentre aquelas pré-carregadas no QGIS. Depois que a camada é selecionada, os atributos de tipo numérico são apresentados para que o usuário possa selecionar quais serão utilizados no processo de agrupamento. Além disso, o usuário necessita selecionar o número de clusters *k* a serem gerados no processamento [Camilo e da Silva 2009]. Para

auxiliar o usuário nesta escolha, foi implementada uma ferramenta de análise do número ótimo de cluster com os métodos *Elbow e Silhouette* [The scikit-yb developers 2019].

Por outro lado, caso o usuário selecione o método hierárquico, as mesmas funcionalidades da escolha da camada e atributos são mantidas. Em seguida, o usuário possui a opção de escolha das medidas de dissimilaridade entre cluster *Ward, Single Linkage, Complete Linkage* e *Average Linkage,* bem como as métricas *Euclidiana e Manhattan* para o cálculo das distâncias [Camilo e da Silva 2009]. Assim como no método *k-means,* o usuário necessita selecionar o número *k* de *clusters* a serem criados.

O principal produto gerado pelo *plugin* (*output)* é uma camada vetorial temporária com os mesmos atributos da camada selecionada para o processo de agrupamento, acrescido de um atributo que registra o número do *cluster* a que pertence cada feição. O *plugin* apresenta ao final do processamento as larguras médias de silhueta, geral e dividida por *cluster*, como indicador de qualidade da clusterização. Também são disponibilizadas as regras de decisão extraídas do resultado da etapa anterior, a fim de auxiliar o usuário na compreensão da semântica implícita na geração dos *clusters.*

4. Resultados

O plugin ClusterMap está sendo preparado para publicação no repositório oficial do QGIS. Usuários de diversas áreas de atuação instalaram o *plugin* a partir do arquivo compactado fornecido pelos autores para fins de teste e identificação de oportunidades de melhoria. Esses usuários vislumbraram possíveis aplicações para o plugin, uma vez que a visualização das geometrias dos objetos pertencentes a um mesmo *cluster* pode sugerir a dependência espacial de um conjunto de variáveis, ou indicar algum padrão de comportamento baseado na experiência do usuário.

Na primeira página deste trabalho, está indicado o *link* para um vídeo que ilustra uma aplicação do *plugin*.

5. Conclusão

Neste artigo foi apresentado o *plugin* clusterMap, utilizado para realizar clusterização e visualização de um conjunto de dados multivariados georreferenciados, em formato vetorial, com variáveis do tipo numérico. O plugin também disponibiliza ao usuário métricas para avaliação da qualidade da clusterização e um conjunto de regras de decisão que permitem ao usuário interpretar a lógica de formação de cada *cluster*.

Referências

Camilo, C. O. e da Silva, J. C. (2009) "Mineração de Dados: Conceitos, Tarefas, Métodos e Ferramentas: Relatório Técnico RT-INF_001-09", Universidade Federal de Goiás.

Silvano, T. P., Correa, B. M. e Barbosa, I. (2020) Análise da distribuição espacial de indicadores sociais e demográficos: uma abordagem baseada em mineração de dados. In. *Revista Brasileira de Cartografia*, v. 72. n. 1, páginas 67-80.

The scikit-yb developers (2019) "Clustering Visualizers", https://www.scikit-yb.org/en/latest/api/cluster/index.html, Setembro.