GGSOM: FERRAMENTA DE VISUALIZAÇÃO BASEADA EM MAPAS AUTO-ORGANIZÁVEIS

Felipe Carvalho de Souza[1], **Rafael Duarte Coelho dos Santos**[1], **Karine Reis Ferreira**[1]

[1]Instituto Nacional de Pesquisas Espaciais (INPE)
São José dos Campos – SP – Brazil

{felipe.carvalho,rafael.santos,karine.ferreira}@inpe.br

Abstract. *Analysis of multidimensional and time series data is useful and pertinent to several different applications, being a challenge due to the volume and complexity of the data. A possible approach for analysis of this kind of data is to use clustering algorithms to reduce the dimensionality of the data. This paper presents a tool for clustering and visualization of data, called ggsom, which uses a technique for data dimensionality reduction through projection of the data in a smaller number of dimensions by the Kohonen's Self-Organizing Map. The tool is evaluated with data from time series of vegetation coverage from Bahia state.*

Resumo. *A análise de dados multidimensionais e séries temporais é útil e aplicável em diferentes contextos, porém é um desafio dado o seu volume e complexidade. Uma possível abordagem para análise deste tipo de dados é através do uso de algoritmos de agrupamento para redução da dimensão dos dados. Este trabalho apresenta uma ferramenta de agrupamento e visualização de dados, denominada ggsom, a qual usa a técnica de redução da dimensionalidade através da projeção dos dados em menos dimensões por meio do algoritmo de Mapas Auto-Organizáveis de Kohonen. A ferramenta é avaliada com os dados de séries temporais de cobertura do solo da região da Bahia.*

1. Introdução

Dados de séries temporais são agentes de descobertas científicas em diferentes domínios, por exemplo, Astronomia [Rebbapragada et al. 2009], Biologia [Fujita et al. 2012], Medicina [Wismüller et al. 2002]. Posto que avanços científicos têm se realizado com o grande volume de dados de séries temporais disponíveis, ainda assim, é uma tarefa complexa explorá-los, pelo fato da alta dimensionalidade contida nos mesmos. Dimensionalidade refere-se ao número de atributos de um conjunto de dados.

Os problemas ocasionados pela alta dimensionalidade são descritos por [Verleysen and François 2005], os quais são divididos em duas partes: conceitual e tecnológica. O problema conceitual refere-se à contra-intuição em entender o espaço geométrico multidimensional, pela dissemelhança de propriedades conhecidas de espaços de duas ou três dimensões. Na parte tecnológica, os autores mencionam a ausência de ferramentas para análise de dados com alta dimensão. Levando em consideração os problemas apresentados, este trabalho apresenta uma ferramenta de visualização de dados, denominada *ggsom*[1], que utiliza a técnica de redução de dimensionalidade por meio do

[1]https://CRAN.R-project.org/package=ggsom

algoritmo de Mapas Auto-Organizáveis (SOM) visando auxiliar tarefas de análise exploratória de dados (EDA).

2. Área de Estudo

A área de estudo compreende as cidades do oeste da Bahia, norte de Goiás e sul de Tocantins. A região de estudo foi escolhida com base no conjunto de 275 amostras coletadas em campo, com as seguintes classes: Algodão, Área Urbana, Milho, Vegetação Arbustiva, Cerradão, Florestal Ciliar, Pastagem Arbustiva, Pastagem Herbácea, Soja e Solo Exposto. A paleta de cores foi definida manualmente de forma que, as classes mais parecidas espectralmente compreendam cores mais próximas.

Os dados usados neste estudo foram extraídos do sensor *MultiSpectral Instrument* a bordo do satélite Sentinel-2A desenvolvido pela ESA[2]. Para nosso estudo, as séries temporais extraídas correspondem ao ano agricola de agosto de 2017 a abril de 2018, após a extração foi calculado o Índice de Vegetação por Diferença Normalizada (NDVI).

3. Desenvolvimento

A ferramenta desenvolvida neste trabalho baseia-se em dois pacotes da linguagem de programação R: Kohonen[3] e ggplot2[4]. O pacote Kohonen é usado para treinar o SOM e o ggplot2 para a criação do gráfico de coordenadas paralelas. Desta forma, a ferramenta *ggsom* opera como um utilitário entre os dois pacotes supracitadas, de forma a modelar o dado gerado pelo Kohonen e visualizá-lo no ggplot2.

4. Resultados

Com o objetivo de avaliar a ferramenta, várias configurações do SOM foram geradas: topologia retangular e variações de 3x3, 6x6, 9x9 e 12x12 de neurônios. Através de uma análise visual, o melhor resultado foi o SOM 6x6, apresentado na Figura 1. O número localizado no canto superior esquerdo mostra a quantidade de observações associadas a cada neurônio.

De acordo com a Figura 1, apenas alguns grupos alcançaram uma separação de classes totalmente homogênea, por exemplo: Soja (6x3) e Pastagem Herbácea (1x2). Aconteceram algumas confusões esperadas, por conta da similaridade espectro-temporal, como: Milho com Soja (5x5) e Vegetação Arbustiva com Florestal Ciliar (4x6). Os grupos com os piores resultados são Soja com Solo exposto (6x2) e Área Urbana com Vegetação Arbustiva e Herbácea (4x1).

A partir da análise feita, é possível concluir que tais quedas na série temporal não pertencem aos períodos de colheita, pois diversos neurônios confundiram classes espectralmente distintas, por exemplo Cerradão com pastagem Herbácea e Arbustiva e Área Urbana.

5. Conclusão

Neste trabalho foi apresentado a ferramenta *ggsom*, usada para realizar a análise exploratória com redução de dimensionalidade do conjunto de dados de cobertura do solo,

[2]https://sentinel.esa.int/web/sentinel

[3]https://CRAN.R-project.org/package=kohonen

[4]https://CRAN.R-project.org/package=ggplot2

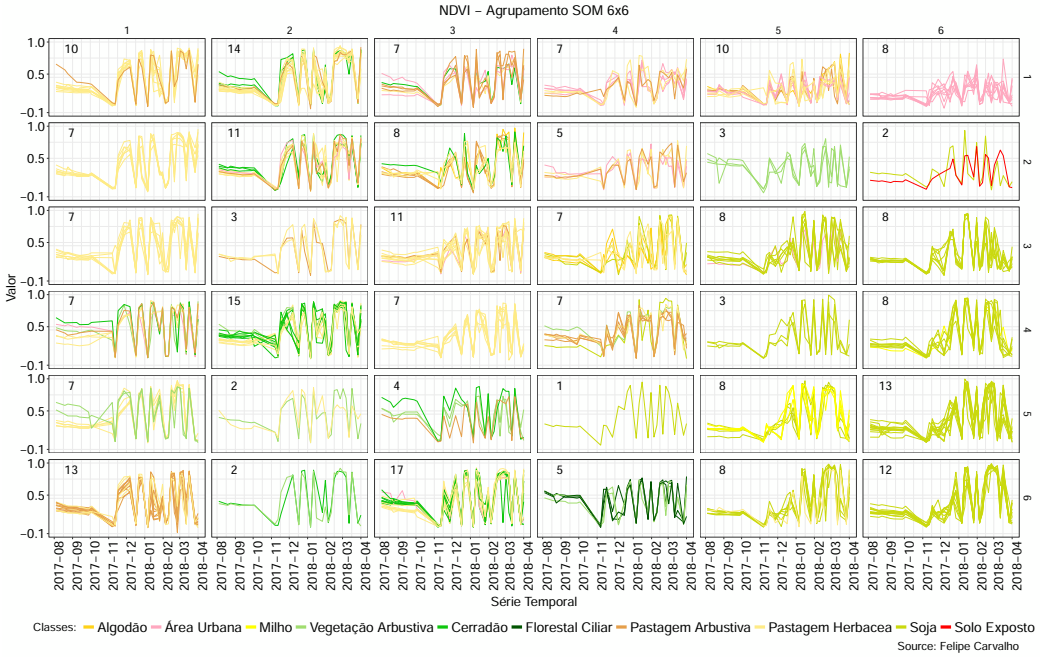

Figura 1. Visualização em coordenadas paralelas em matriz produzida pela ferramenta *ggsom* para o agrupamento da rede SOM 6x6

usando como técnica de visualização coordenadas paralelas. Através do uso da ferramenta foi possível identificar padrões na série temporal, assim, avaliando o comportamento espectral de cada classe e concluindo que os picos e as quedas apresentados na mesma representam nuvens. Outra informação obtida foi a homogeneidade de algumas classes, por exemplo Soja, informação útil para futuramente utilizar algoritmos de classificação.

Referências

Fujita, A., Severino, P., Kojima, K., Sato, J. R., Patriota, A. G., and Miyano, S. (2012). Functional clustering of time series gene expression data by granger causality. *BMC systems biology*, 6(1):137.

Rebbapragada, U., Protopapas, P., Brodley, C. E., and Alcock, C. (2009). Finding anomalous periodic time series. *Machine learning*, 74(3):281–313.

Verleysen, M. and François, D. (2005). The curse of dimensionality in data mining and time series prediction. In *International Work-Conference on Artificial Neural Networks*, pages 758–770. Springer.

Wismüller, A., Lange, O., Dersch, D. R., Leinsinger, G. L., Hahn, K., Pütz, B., and Auer, D. (2002). Cluster analysis of biomedical image time-series. *International Journal of Computer Vision*, 46(2):103–128.