geocmeans: A R package for spatial fuzzy c-means

20 Avril 2021

# Summary

Unsupervised classification methods like *k-means* or the *Hierarchical Ascendant Classification* (HAC) are widely used in geography even though they are not well suited for spatial data. Yet, recent development has been proposed to include the geographical dimension into clustering. As an example, ClustGeo (Chavent et al. 2018) is a spatial extension of the HAC, available in the R package with the same name. We propose in the R package geocmeans a spatial extension of the *Fuzzy C-Means* (FCM) algorithm to complete the growing toolbox with a fuzzy approach. The package provides also several helper functions to assess and compare quality of classifications, select appropriate hyperparameters, and interpret the final groups.

# Statement of need

Traditional clustering algorithms do not account for the geographical dimension of data causing two main concerns:

* First, an important part of the information related to observations’ locations and geographical organization is disregarded. Often, in social, environmental and economical sciences the geographical dimension is highly structuring, in particular in the context of positive spatial autocorrelation.
* Second, in many applications (like the development of regional policy), it is desirable that close observations have more chances to belong to the same group. In this regard, it is common to observe with traditional unsupervised classification algorithms “holes”: observations attributed to a certain group and surrounded by observations attributed to another. Often, the difference of the semantic attributes between those observations does not justify such spatial inconsistency.

A first approach for spatial clustering explored from the 1990s was to impose to the classification a spatial constraint based on contiguity (also called aggregation based approach). The most probably known methods are the AZP (Openshaw 1977), SKATER(Assunção et al. 2006) and AMOEBA (Aldstadt and Getis 2006). However, this approach can only yield spatially continuous groups and can be too strict to obtain meaningful clusters. That is why a new approach has been proposed: including the spatial dimension. The spirit is: space should not act as a constraint in the classification but like supplementary data.

The recent method ClustGeo (Chavent et al. 2018) has received some attention from geographers. However, it yields a hard clustering and may hide situations were observations are located at the border of two groups. There is thus a need for available unsupervised spatial fuzzy clustering methods.

Such methods have been discussed and applied in the brain imagery segmentation field (Cai, Chen, and Zhang 2007; Zhao, Jiao, and Liu 2013) leading to the development of the Spatial Fuzzy C-means (SFCM) algorithm. The package geocmeans is proposed to make this tool available to researchers and professionals working with spatial datasets. A first application to construct a socio-residential and environmental taxonomy in Lyon has outlined the potential of the method (Gelb and Apparicio 2021).

# Core functionality

The geocmeans package has been built to be self-sufficient and minimize the coding need from users. Thus, several helper functions are available along the main clustering algorithms to asses classification quality, select hyper parameters and interpret results.

## Main algorithm

geocmeans provides four fuzzy unsupervised classification algorithms:

* CMeans, the original c-means algorithm, requiring two hyper parameters, *m* (fuzzyness degree) and *k* (number of groups).
* GFCMeans, the so-called generalized c-means algorithm. It is known to accelerate convergence and yield less fuzzy result by adjusting the membership matrix at each iteration. It requires an extra *β* parameter controlling the strength of the modification. The modification only affects the formula updating the membership matrix.

with:  
  
  
 the probability for observation *k* to belong to cluster *i*  
 the observation *k* in the dataset *x*  
 the cluster *i*  
*m* the fuzzyness parameter

* SFCMeans, the SFCM algorithm, requiring two more parameters *W* and *α*. *W* is a spatial weight matrix used to calculate a spatially lagged version of the dataset *x*. *α* is used to control the weight of the spatially lagged dataset. If then SFCM degenerates to a simple FCM. If the same weight is given to the original and lagged dataset. If then the spatially lagged dataset has a weight doubled in comparison with the original dataset, and so on… The integration of the spatially lagged dataset modifies the formula updating the membership matrix and the formula updating the centers of clusters.

with:

the cluster *i*  
 the spatially lagged version of *x*

As the formula suggests, the SFCM can be seen as a spatially smoothed version of the FCM and *α* controls the degree of spatial smoothness. This smoothing can be interpreted as an attempt to reduce spatial overfitting of the FCM.

* SGFCMeans, the SGFCM algorithm, combining SFCM and SGFCM and thus requiring the definition of three extra parameters *W*, *α* and *β*. Only the formula to calculate the membership matrix is different from the SFCM.

## Selecting parameters

As stated above, up to five hyper parameters have to be selected by the user. Finding the best combination is facilitated by the function selectParameters calculating the classifications for all the possible combinations of parameters in specified ranges and returning several metrics of classification quality.

## Interpreting the results

To interpret the results, four functions are provided:

* summarizeClusters: returning summary statistics for each cluster for in-depth analysis.
* spiderPlots: plotting a spider chart to quickly differentiate the clusters.
* violinPlots: plotting one violin plot split by cluster for each variable in the dataset.
* mapClusters: mapping the membership matrix and the most likely cluster for the observationss.
* calcqualityIndexes: returning several quality indexes for a classification.
* spatialDiag: performing a complete spatial diagnostic to determine if the inclusion of space in the classification is justified.

# Example

The data used for the socio-residential and environmental taxonomy in Lyon are included in the package. The following example uses this data to demonstrate the basic functionality of the package. More details are given in the vignettes of the package.

library(geocmeans)  
library(spdep)  
  
data(LyonIris)  
  
#selecting the columns for the analysis  
AnalysisFields <-c("Lden","NO2","PM25","VegHautPrt","Pct0\_14",  
 "Pct\_65","Pct\_Img","TxChom1564","Pct\_brevet","NivVieMed")  
   
#creating a spatial weights matrix  
Neighbours <- poly2nb(LyonIris,queen = TRUE)  
WMat <- nb2listw(Neighbours,style="W",zero.policy = TRUE)  
  
#rescaling the columns  
Data <- LyonIris@data[AnalysisFields]  
for (Col in names(Data)){  
 Data[[Col]] <- scale(Data[[Col]])  
}  
  
#considering k = 4 and m = 1.5, find an optimal value for alpha  
DFindices\_SFCM <- selectParameters(algo = "SFCM", data = Data,  
 k = 4, m = 1.5, alpha = seq(0,2,0.05),  
 nblistw = WMat, standardize = FALSE,  
 tol = 0.0001, verbose = FALSE, seed = 456)  
  
#keeping alpha = 0.7  
SFCM\_results <- SFCMeans(Data, WMat, k = 4, m = 1.5, alpha = 0.7,  
 tol = 0.0001, standardize = FALSE,  
 verbose = FALSE, seed = 456)  
  
#calculating some quality indexes  
calcqualityIndexes(Data, SFCM\_results$Belongings)  
  
#mapping the results  
mapClusters(LyonIris, SFCM\_results$Belongings)

# Acknowledgements

We are grateful to Professor Philippe Apparicio for his comments and suggestions on the package, its documentation and this article.

The project was partially funded by the research chair of Canada on environmental equity and the city (950-230813).

# References

Aldstadt, Jared, and Arthur Getis. 2006. “Using AMOEBA to Create a Spatial Weights Matrix and Identify Spatial Clusters.” *Geographical Analysis* 38 (4): 327–43.

Assunção, Renato M, Marcos Corrêa Neves, Gilberto Câmara, and Corina da Costa Freitas. 2006. “Efficient Regionalization Techniques for Socio-Economic Geographical Units Using Minimum Spanning Trees.” *International Journal of Geographical Information Science* 20 (7): 797–811.

Cai, Weiling, Songcan Chen, and Daoqiang Zhang. 2007. “Fast and Robust Fuzzy c-Means Clustering Algorithms Incorporating Local Information for Image Segmentation.” *Pattern Recognition* 40 (3): 825–38.

Chavent, Marie, Vanessa Kuentz-Simonet, Amaury Labenne, and Jérôme Saracco. 2018. “ClustGeo: An r Package for Hierarchical Clustering with Spatial Constraints.” *Computational Statistics* 33 (4): 1799–1822.

Gelb, Jérémy, and Philippe Apparicio. 2021. “Apport de La Classification Floue c-Means Spatiale En géographie: Essai de Taxinomie Socio-résidentielle Et Environnementale à Lyon.” *Cybergeo: European Journal of Geography*.

Openshaw, Stan. 1977. “A Geographical Solution to Scale and Aggregation Problems in Region-Building, Partitioning and Spatial Modelling.” *Transactions of the Institute of British Geographers*, 459–72.

Zhao, Feng, Licheng Jiao, and Hanqiang Liu. 2013. “Kernel Generalized Fuzzy c-Means Clustering with Spatial Information for Image Segmentation.” *Digital Signal Processing* 23 (1): 184–99.