COMPARISON OF CLUSTERING METHODS TO DERIVE ECOSYSTEM FUNCTIONAL TYPES

**Ecosystem Functional Types: Clustering Methods**

This is a comparison of the Ecosystem Functional Types (EFTs) obtained using two different clustering methods: ISODATA and K-means. The EFTs map from ISODATA presented here (Figure 1) is directly the one published by Ivits & Cherlet (2013), while the map result of the K-means method (Figure 2) is the one just produced in 2020.

The methodology to prepare the data to be clustered is (or should be) exactly the same. This is a three steps process: (1) the 5 phenological variables are checked against multicollinearity, removing those highly correlated (|r| > 0.7); (2) a first PCA (“screening” PCA) with the less-correlated variables is run in order to know the optimal number of variables (PC axes) to be used in a subsequent PCA, as well as the phenological variable most associated to that PC; and (3) a final PCA is run with the results of the step 2. After that, the clustering process is made using the rotated components of the final PCA instead of with the phenological variables because the former are normalized with zero mean and 1 SD, which is statistically meaningful for iterative clustering of pixels using minimum distance techniques (Ivits et al., 2013).

However, some remarks need to be done at this point:

- The time series used for each method are different. For the ISODATA map they are 1982-2010, while for K-means are 1999-2012.

- On the one hand, the ISODATA method is able to optimize the final number of clusters, which in this case was 14. On the other hand, the K-means method is not able to optimize the number of clusters, it needs to be done externally. In this case, the optimal number of clusters was calculated to be around 20, using the “scree plot” method. Such method is based on running several K-means clustering with different number of clusters each in order to assess how the quality of the models change with the number of clusters. Then, a plot is made with the results and an “elbow” indicates where the quality of the model no longer improves substantially as the number of clusters (i.e. model complexity) increases. See Figure 3 for the results of this assessment. Even though the optimal number of clusters was calculated to be around 20, in order to compare both methods, a K-means clustering map was produced with 14 clusters (Figure 2). However, the map with 20 clusters is also included in this report, see Figure 4.

- As previous tests of K-means with up to 100 iterations were not converging, the maximum number of iterations was set to 500. Within this limit, the process achieved convergence in all the tests made. In contrast, it is reported that the ISODATA clustering was run with 5 iterations. This low number is probably due to the extreme slowness of the method. However, being a global study, it seems unlikely that this process achieve any convergence with such a low number of iterations.

- The colour palettes used in the three maps do not correspond to clusters placed in similar areas

**Conclusions**

Both ISODATA and K-means are iterative algorithms and probably both have pros and cons. However, currently ISODATA is not among the most used algorithms for clustering data, while in contrast, K-means is widely used in data analysis. This fact makes that several R packages can be found to run the latter, and they are optimized in terms of processing speed. Contrarily, an ISODATA algorithm could not be found in any R package and, therefore, it had to be developed *ad hoc*.

The “scree plot” method implemented in this study, undoubtedly has some level of subjectivity. There exists several numerical methods to calculate the optimal number of clusters that remove such subjectivity, but which take also some statistical assumptions. They could be explored if a higher level of accuracy is believed as necessary.

Although the general pattern of both maps are quite similar, some significant differences can be appreciated. These dissimilarities might be due to the differences mentioned above (i.e. different time series, issue with the convergence) or also to the fact that both algorithms are highly dependent on the random initialization of centroids at the beginning of the process. In some cases it is advised to run the algorithm several times using different initialization centroids and pick the results of the run producing the lower sum of squared distance. This approach could also be implemented, although it would slow down significantly the whole process.

In conclusion, the results obtained with K-means seem meaningful and the algorithm fully adequate to derive the Ecosystem Functional Types within the framework of the Land Productivity Dynamics indicator.

**Figures**

|  |
| --- |
|  |
| Figure 1: ISODATA classification from Ivits & Cherlet, 2013 |

|  |
| --- |
|  |
| Figure 2: K-means classification made with the methodology developed for the Land-Productivity Dynamics indicator, but setting the number of clusters suboptimally to 14 EFTs |

|  |
| --- |
|  |
| Figure 3: “Scree plot” method used to calculate the optimal number of clusters. The “elbow” indicates where the quality of the model no longer improves substantially as the number of clusters (i.e. model complexity) increases |

|  |
| --- |
|  |
| Figure 4: K-means classification made with the methodology developed for the Land-Productivity Dynamics indicator, setting the number of clusters to the optimal 20 EFTs |

**Bibliography**

- Ivits, E.& Cherlet, M. (2013) Land-Productivity Dynamics Towards integrated assessment of land degradation at global scales title; EUR 26052; doi:10.2788/59315

- Ivits, E., Cherlet, M., Horion, S. and Fensholt, R. (2013) Global biogeographical pattern of ecosystem functional types derived from earth observation data. Remote Sensing, 5, 3305–3330.