**Python scripts used for performing Partition Around Medoids (PAM) clustering of L2-v3 AERONET measurements**

Martin Montes, SSAI-NASA

Lanham, April 30, 2020

**mainPAM.py**

This script is the main and it is used to perform PAM clustering based on an AERONET L2-v3 global database having SSA records (globe\_aeronet\_SSArecordsID.csv). The main is has two main routines**: PAMbinder.py** and **PAMalgo.R**, thus R open source soft along with libraries must be installed before running mainPAM from a console. The database is open using a csv reader and split in even and uneven record ids in order to generate two subdatasets with 75% and 25% of data (globe\_aeronet\_SSAtrain2.csv and globe\_aeronet\_SSAtest2) for clustering and validation, respectively. The following step is the data stratification based on different aerosol loadings. In our application, only measurements with AOD(0.44) 0.4 are used. Also each AOD(0.44) value corresponds to the mean value between coincident and non-coincident values. Stratified data for clustering are backup in trainglobe75\_3params2.cv.

Before clustering, aerosol properties for clustering are selected (single scattering albedo at a wavelength of 0.44 m, SSA(0.44), and Angstrom exponents for the spectral range 0.44-0.87 m, AEe and AEa, respectively). The next step is to call the R-python wrapper which has two methods: pam\_clus2 and check\_kmax. The PAM clustering is done twice, the first time with an intitial guess in terms of the maximum number of clusters (Kmaxi) and second time using the final estimate or Kmax. The method pam\_clus2 has 2 arguments: kmaxi and flag\_run. In our case, kmaxi=30 and flag\_run indicates if it is the first or the second clustering with initital (init) or optimized (opt) classification (1 or 2, respectively).

In **PAMbinder.py**, a subprocess is built based on a command composed by 4 arguments: type of R call, path of R script, Kmaxi or Kmax anf flag\_run. The subprocess calls the **PAMalgo.R**

Script and the package ClusterR is loaded. The PAM clustering is carried out by Cluster\_Medoids method. This method is set with the Mahalanobis distance. PAMalgo.R output is divided in 3 files storing the cluster indices (ci), the medoid centers (cc) and the silhouette index values (si).

In **PAMbinder.py**, check\_kmax method is called wth arguments defining the number of initial maximum number of clusters (nclus), the silhouette index threshold (sil\_thre) and the number of records per cluster with a Silhouette threshold (nsamples). By default, these values are 30, 0.51 and 10. Values for si are derived from clus\_silinit.csv and used to determine Kmax, the outuput of check\_kmax. The final processing step in **PAMbinder.py is the** the optimization of Mahalanobis distances using Kmax instead of Kmaxi , thus pam\_clus2 is called again but using the flag\_run =2.