Unsupervised learning algorithms for boundary layer study
Thomas Rieutord

To cite this version:

HAL Id: meteo-02465168
https://hal-meteo.fr.archives-ouvertes.fr/meteo-02465168
Submitted on 3 Feb 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Unsupervised learning aims to derive high level information from data without reference. This work shows an example of how it can be used to derive user information from field campaign measurements. Three algorithms have been tested on their ability to make a good boundary layer classification: K-means, Agglomerative and DBSCAN.

Data are from radiosoundings in the 2nd IOP of the Passy-2015 field experiment (alpine valley, wintertime). One can see a stable layer, a mixed layer and the free atmosphere.

Agglomerative gives the best results and has promising prospects. K-means and DBSCAN give clusters not corresponding to visual examination, but both have many ways of improvement.

**K-means clustering**
- Initialisation of centroids
  - Points attributed to the closest centroid
  - Centroid updated to better represent the group
  - Minimum of intra-cluster variance reached
- Fast convergence (few 10 iterations)
- Converges toward a local minimum
- Initialisation highly influences the result
- Choice of the number of groups?

**Agglomerative clustering**
- Can highlight a "natural" number of groups
- Nested clusters (identify smaller scale structures?)
- No parameter to tune
- Graphical summary of results in dendrogram
- Gives hierarchical structure anyway, regardless whether it is relevant
- Small changes in data can lead to different dendrogram
- Choice of linkage?
- Prohibitive cost when large dataset

**DBSCAN clustering**
- Automatically find the number of groups
- Clusters can be of any shape
- Resilient to outliers (can even identify them)
- Edge points connected to more than one cluster can change assignment depending on their ordering
- Clusters must be of similar density
- Choice of the parameters $m$ and $\epsilon$?

**Results**
- K-means clusters have the good borders but they are not consistent with visual examination. More work is required on initialisation (start from meaningful centroids?) and predictors.
- DBSCAN clustering gives 7 clusters, which is too much. It appears to be very sensitive to settings values which are hard to correctly set. More advanced variants (e.g. OPTICS) might be better.