Unsupervised learning algorithms for boundary layer study

Thomas Rieutord*
*thomas.rieutord@meteo.fr

CNRM (UMR 3589), Météo France & CNRS, Toulouse

Presented at *Journée scientifique du SIRTA*, 15/06/2018

Abstract

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 boudary layer classification: *K-means*, *Agglomerative* and *DBSCAN*.

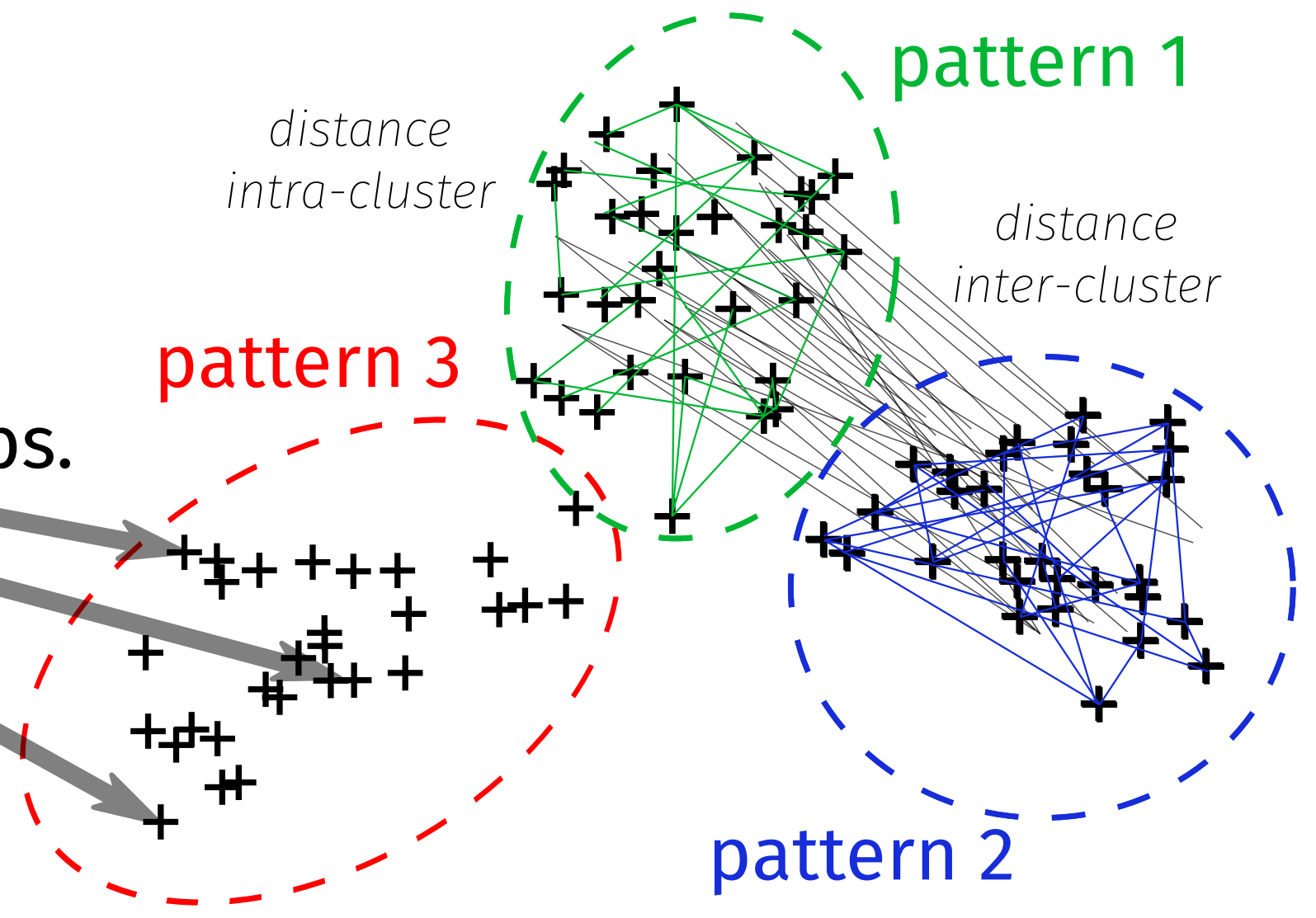
Data are from radiosoundings in the 2nd IOP of the Passy-2015 field experiment (alpine valley, wintertime). One can see a stable layer, a mixed layer and the free atmosphere.

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.

1 Introduction

p predictors
(atm. variables)

$$X = \begin{pmatrix} x_1^1 & \dots & x_p^1 \\ \vdots & & \vdots \\ x_1^N & \dots & x_p^N \end{pmatrix} \quad N \text{ obs.}$$



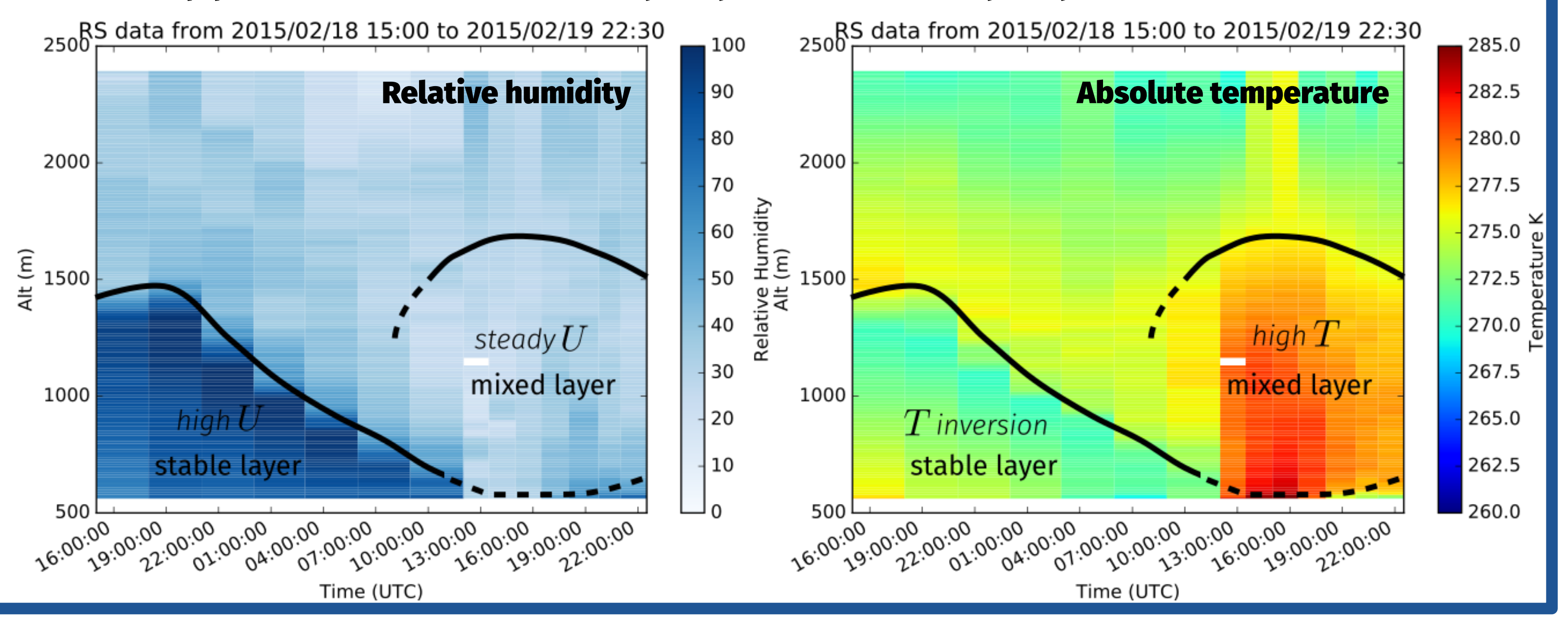
Target: Minimise intra-cluster distance

All methods look for high density areas with a ***dissimilarity metric***

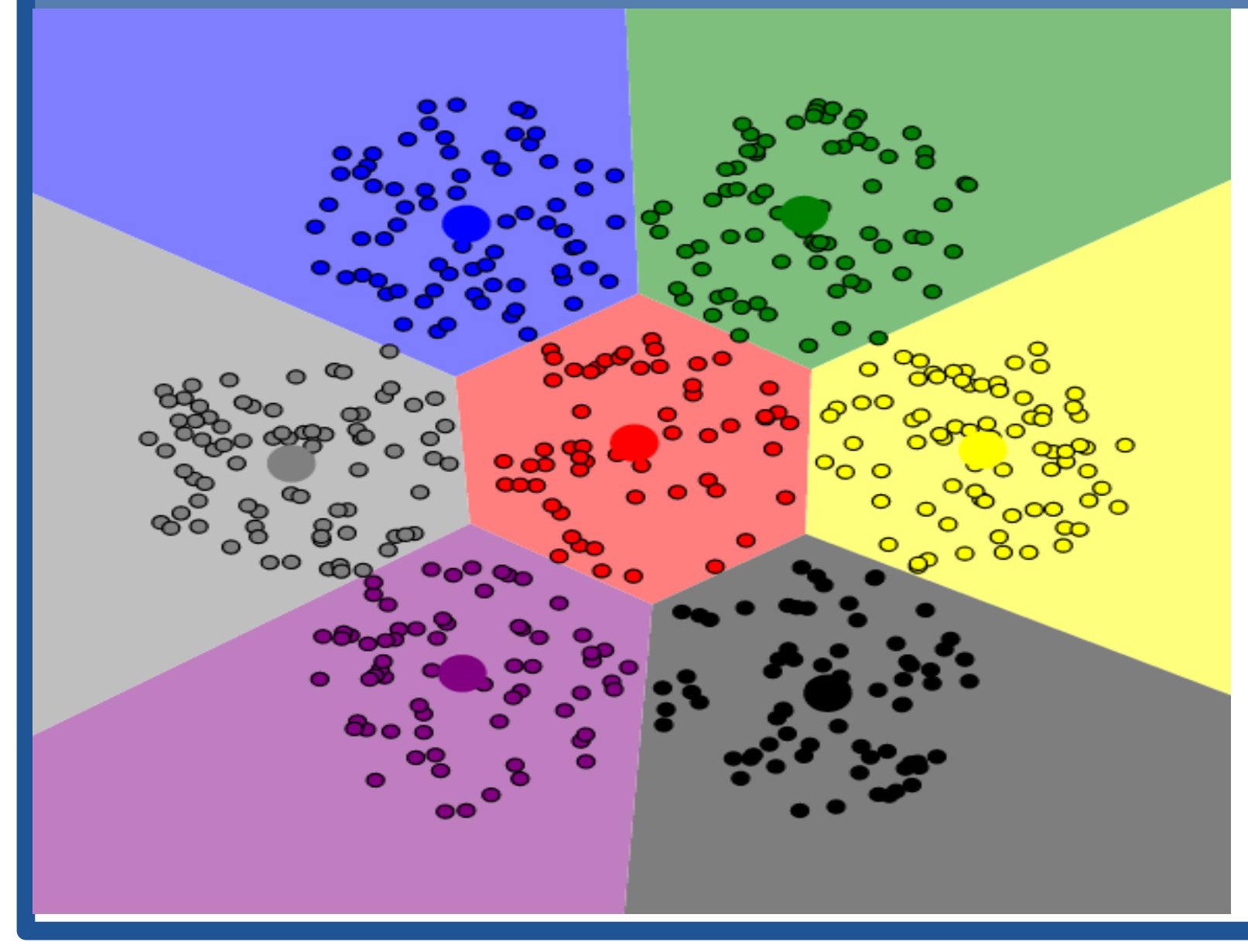
problem-dependent questions

- ***Which predictors?***
  Here: normalized z, t, T, U
- ***Which dissimilarity?***
  Here: squared gap

Application to boundary layer: boundary layer classification



2 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

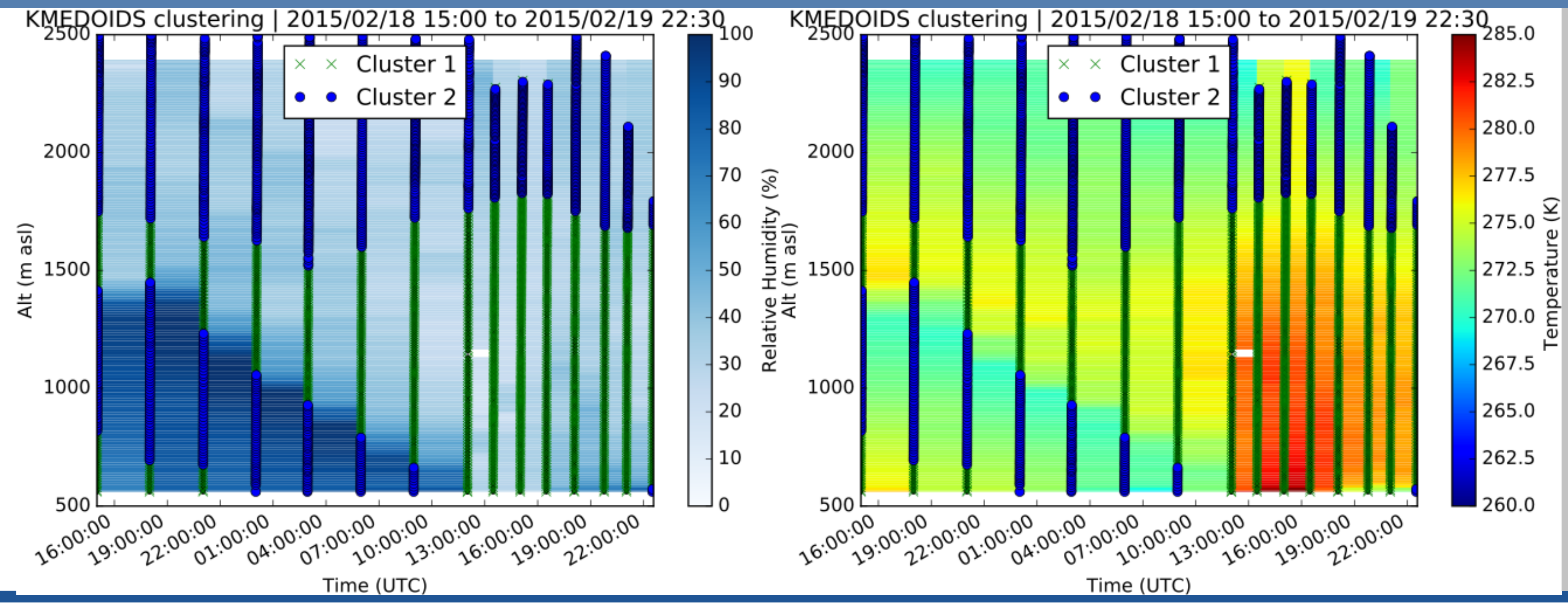
⊕

- Fast convergence (few 10 iterations)
- Different strategies of initialisation

⊖

- Converges toward a local minimum
- Initialisation highly influences the result
- Choice of the number of groups?

Results

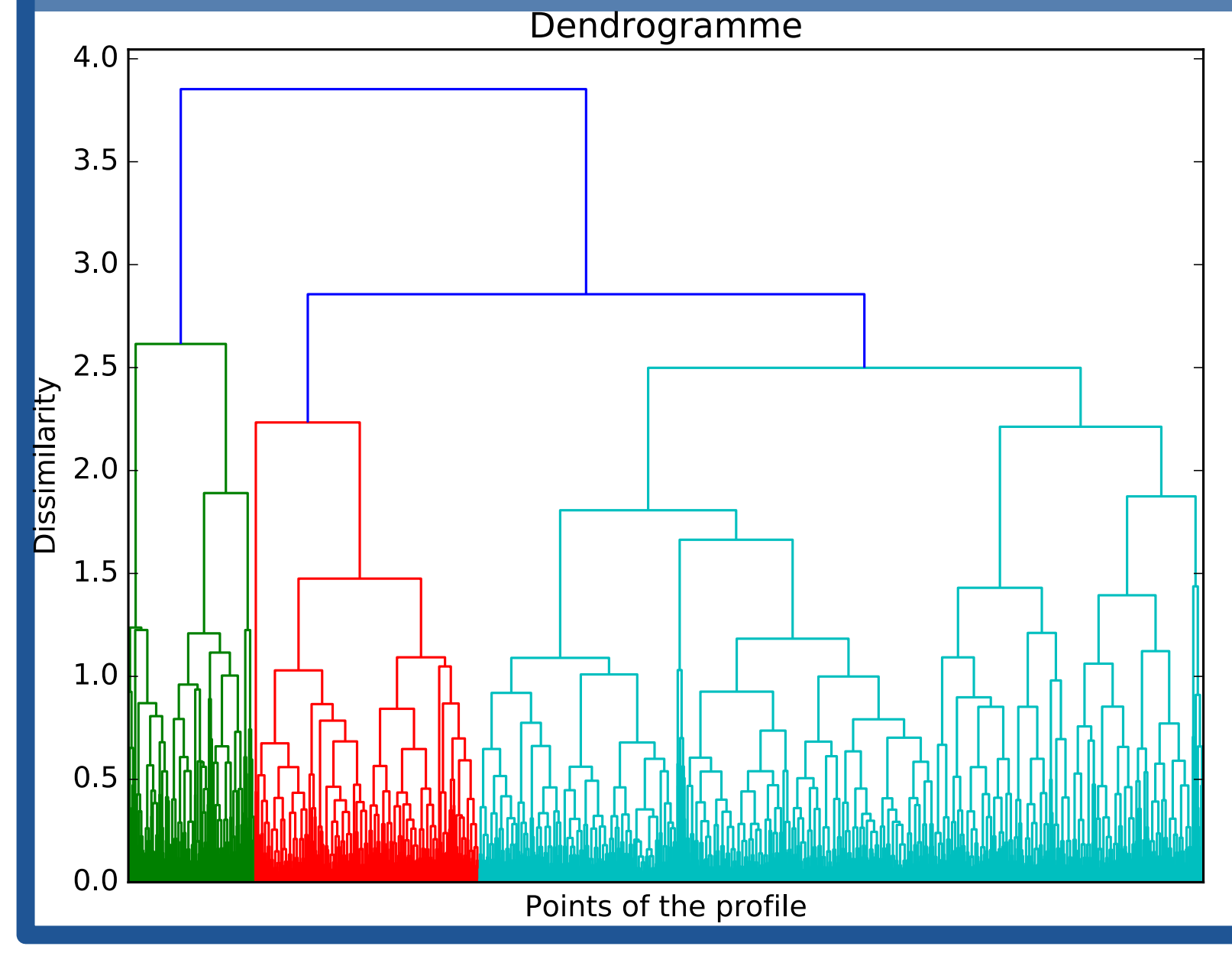


Conclusion

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.

👎 Results 👍 Prospects

3 Agglomerative clustering



- All points are considered as a group of 1 element

The 2 most similar groups are merged

Distances between groups are updated

- One group contains all. Result is the dendrogram (merging tree)

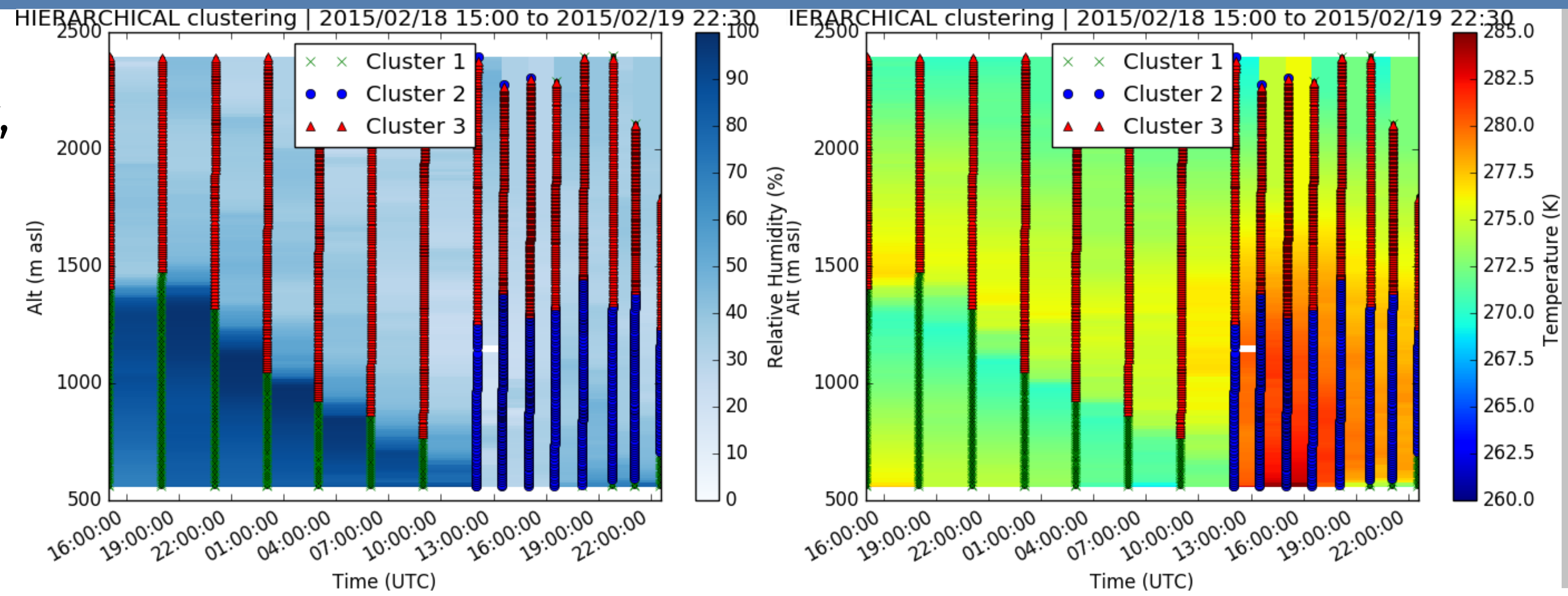
⊕

- 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, regarless whether it is relevant
- Small changes in data can lead to different dendrogram
- Choice of linkage?
- Prohibitive cost when large dataset

Results

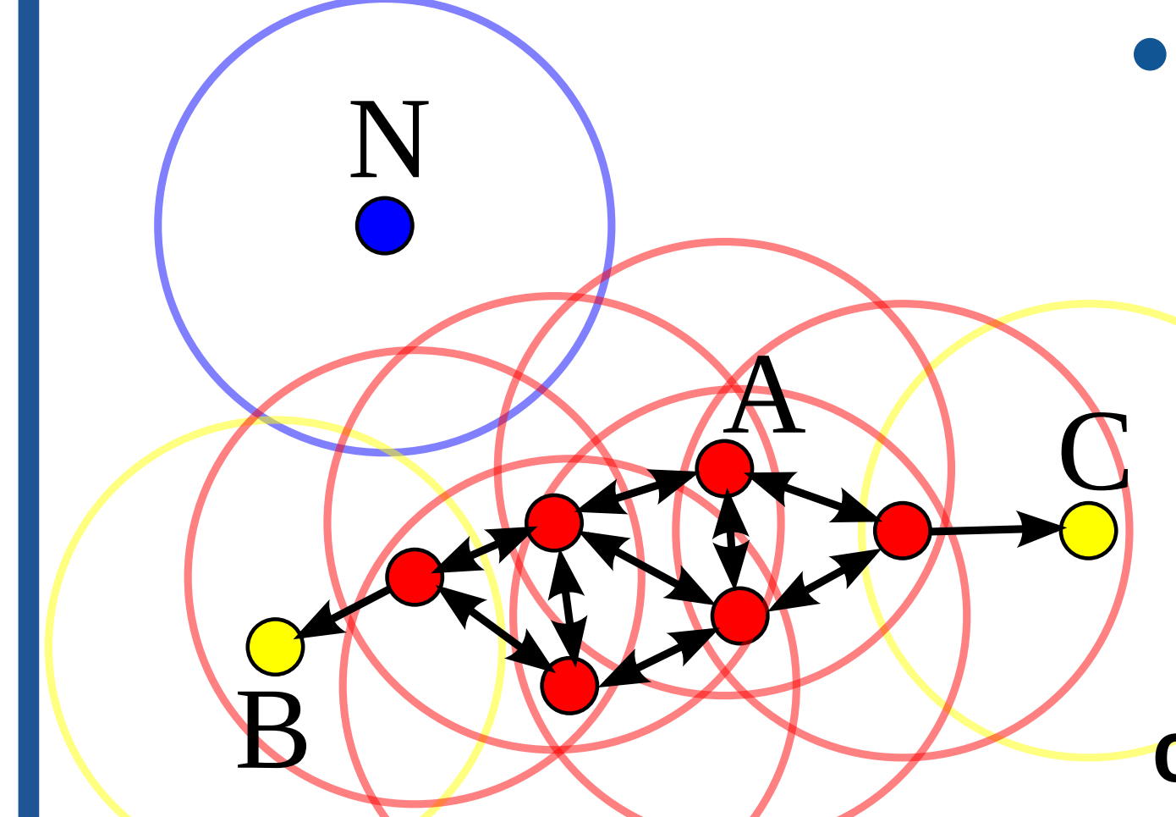


Conclusion

Agglomerative clustering finds well the visible layers. Moreover, the dendrogram gives an even more precise information (still to be examined) and many improvements are possible.

👍 Results 👍 Prospects

4 DBSCAN clustering



- For all points in the dataset
  - Neighbours = points closer than ϵ

More than m neighbours? NO / YES

Core point same cluster as its neighbours.

Any core point in the neighborhood? NO / YES

Outlier No cluster attribution.

Edge point same cluster as its neighbours.

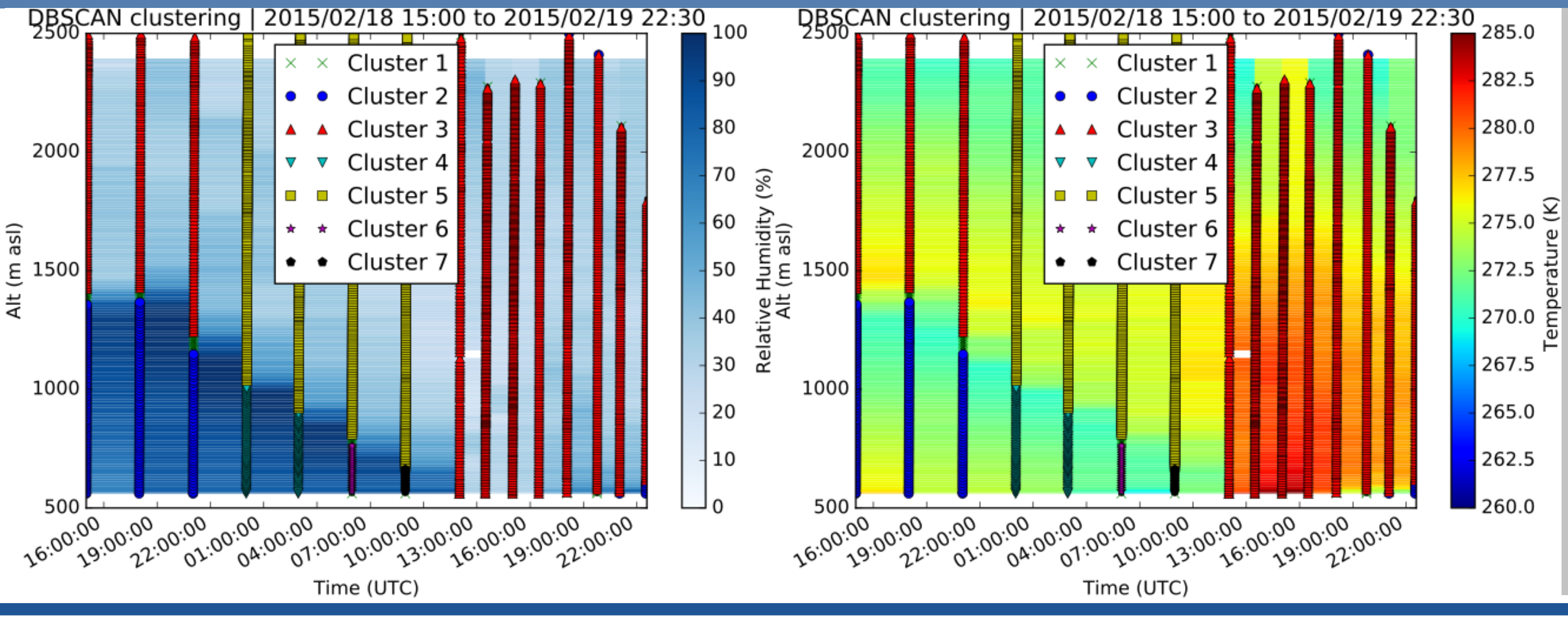
⊕

- 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 assignation depending on their ordering
- Clusters must be of similar density
- Choice of the parameters m and ϵ?

Results



Conclusion

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.

👎 Results 👍 Prospects