Boundary layer classification with machine learning

Thomas Rieutord, Alexandre Paci, Alain Dabas
CNRM, Météo-France/CNRS (UMR 3589), Toulouse, France

ISTP11 – 23/05/2019

Outline

Outline

Fields campaigns in atmospheric sciences

Our problematic

Measured variables

- Temperature
- Humidity
- Wind
- Aerosols

User information

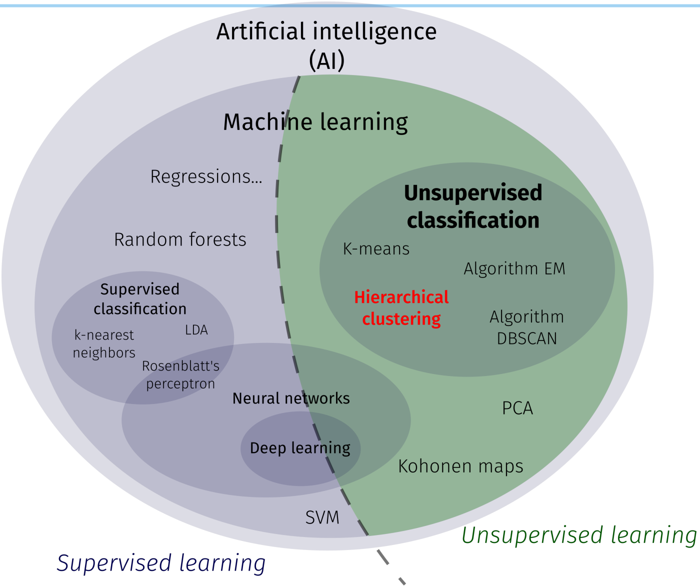
- Warnings (shear, fog, turbulence...)
- Tracking (clouds, BL top...)
- Summary (identification of BL types...)
- Others (outlier identification and removal, uncertainty quantification...)

Problem: no reference for user information...

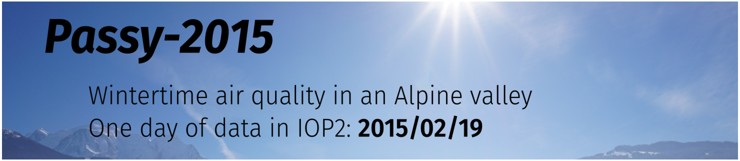
$\Rightarrow$ **Unsupervised learning**

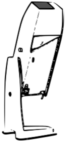
Outline

Machine learning algorithm

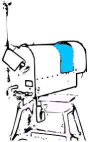


Data used for this study





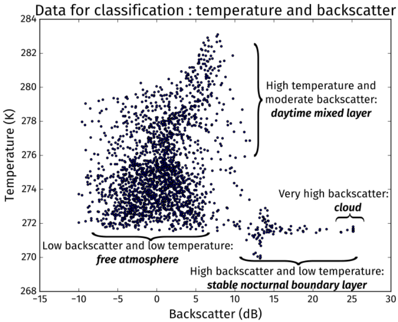
Ceilometer: vertical profiles of **aerosol backscatter**
Δt=15 seconds. Δz=15 meters.

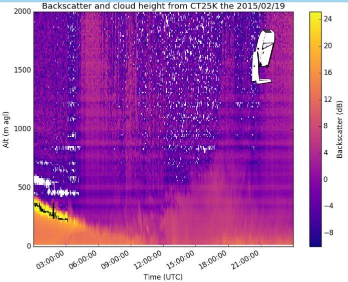


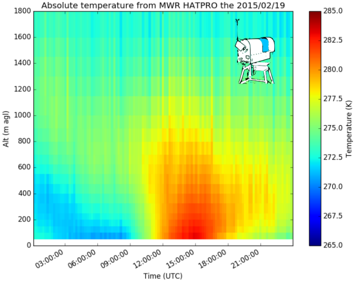
Microwave radiometer: vertical profiles of **temperature**
Δt=12 minutes. Δz=50m at 100m and 300m at 2500m (variable).

Full campaign data: 148Go | Used for this case study: 5.8Mo

Raw data examination

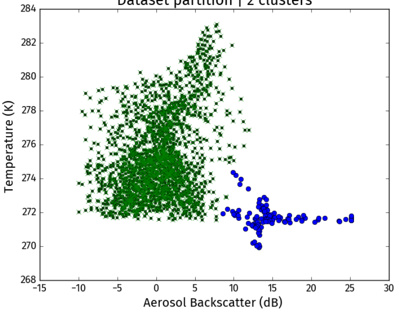




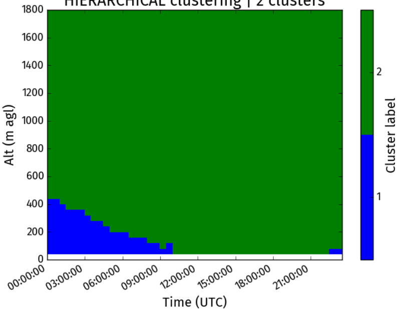


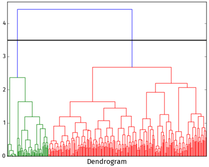
Results of hierarchical clustering

Clusters in feature space



Clusters in z, t space

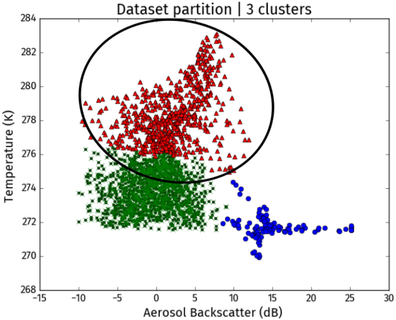




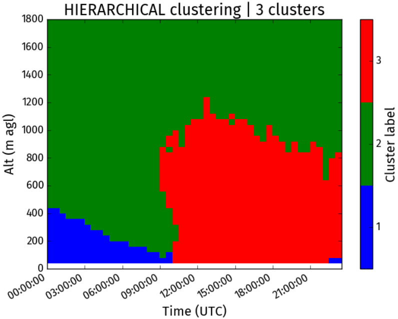
First, the algorithm distinguishes the **stable boundary layer** (blue) for the rest (green).

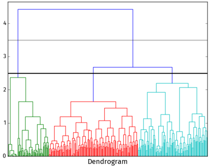
Results of hierarchical clustering

Clusters in feature space



Clusters in z, t space

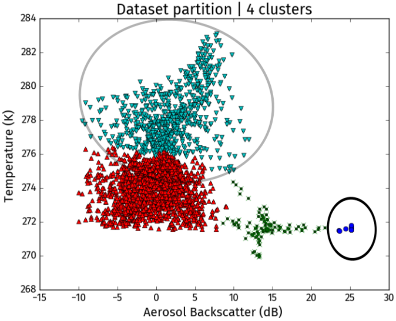




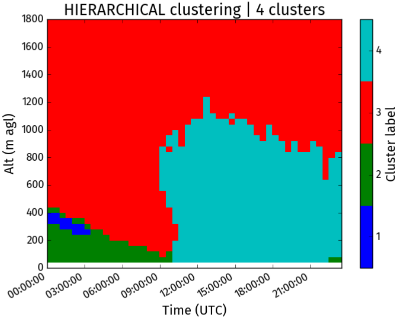
Then, an area corresponding roughly to the **mixed layer** (red) is separated from the large area (green).

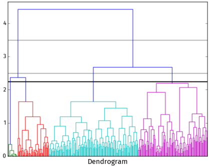
Results of hierarchical clustering

Clusters in feature space



Clusters in z, t space

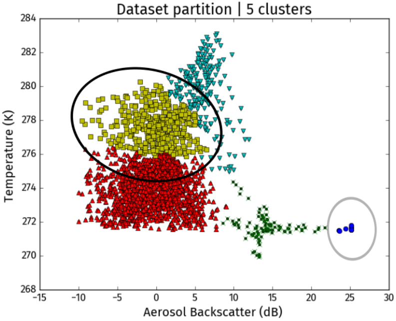




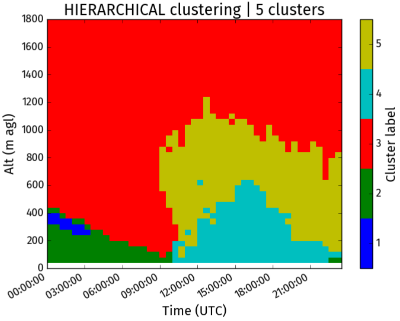
Inside the stable boundary layer (green), it distinguishes a **cloud** (blue) characterized by very high backscatter.

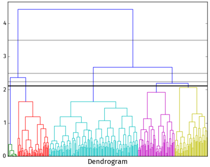
Results of hierarchical clustering

Clusters in feature space



Clusters in z, t space

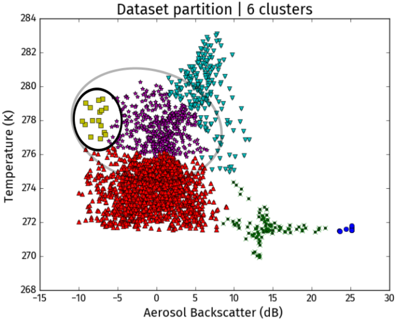




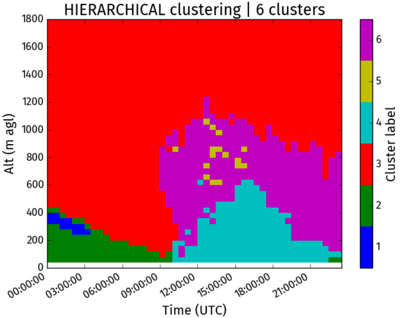
The mixed layer is split into a lower layer (cyan, hot and dusty) and an upper layer (yellow, mild and pure). It reflects the decoupling between temperature and backscatter.

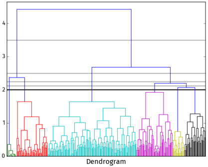
Results of hierarchical clustering

Clusters in feature space



Clusters in z, t space

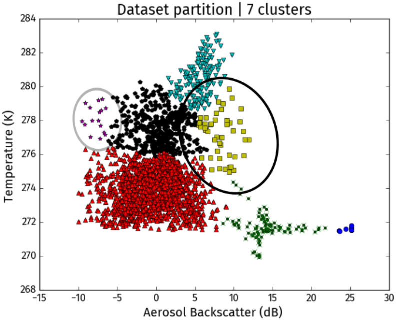




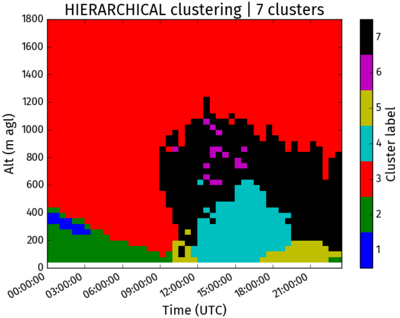
The next cluster (blue) is taken from the **upper layer** (pink) and seems to be made of **outliers**.

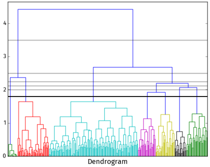
Results of hierarchical clustering

Clusters in feature space



Clusters in z, t space



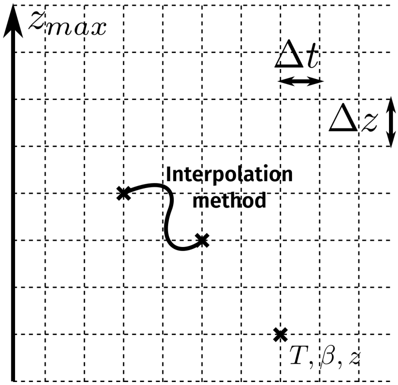


The next cluster (yellow) is taken from the lower layer (cyan) and depicts **transition layers**.

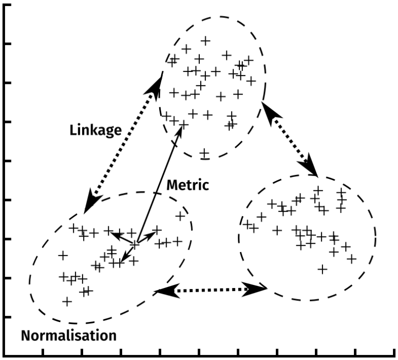
Outline

Inventory of parameters

Parameters of the dataset



Parameters of the classification



9 parameters, 184 datasets, 216 classification for each

Total: 39 744 classifications to perform.

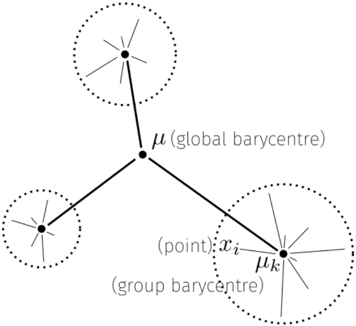
Measure of performance

Ideally, the quality of unsupervised classification must be assessed by a human expert. But impossible to manually check 39744 classifications...

⇒ use of a score: the Calinski-Harabazs index[1].

$$CH = \frac{\text{between-groups dispersion}}{\text{within-group dispersion}}$$

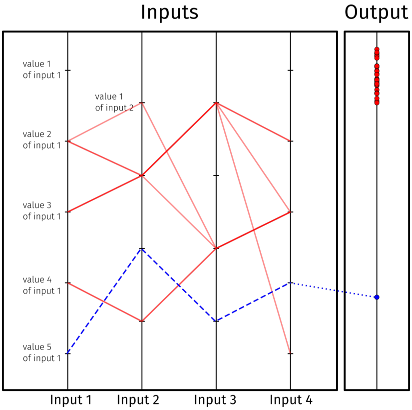
$$CH = \frac{\frac{1}{K-1}\sum_{k=1}^{K}\|\mu_k - \mu\|^2}{\frac{1}{N-K}\sum_{k=1}^{K}\sum_{i\in I_k}\|x_i - \mu_k\|^2}$$



The higher CH, the better the classification.

[1] Calinski T. and Harabaz J. (1974) A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1), 1-27.

Overview of sensitivity analysis: cobweb plots



Read cobweb plots[1,2]:

- Each thread links all inputs that gave one of the output.

  (ex: blue dashed thread)

- We select few best/worse outputs.

  (red threads and bullets)

- Repeated passage show what set of input values give good/bad outputs.

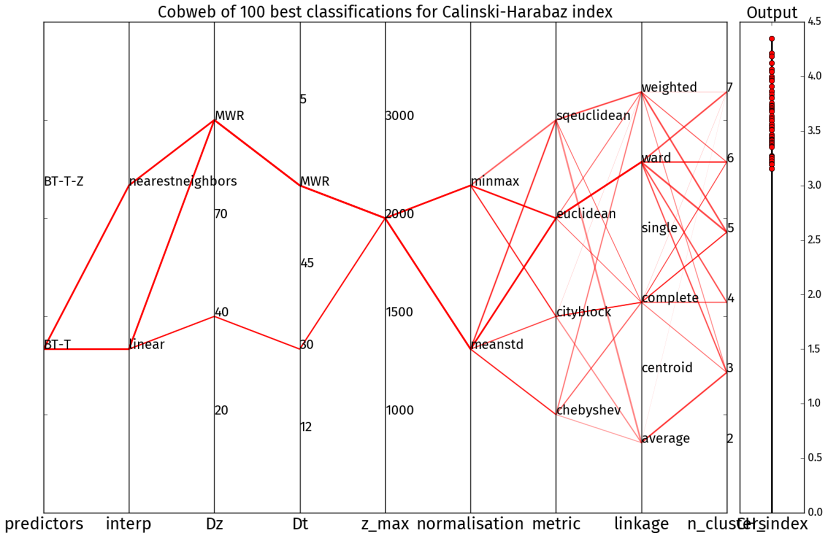
  (input1=value3+input2=value2+input3=value1;

  input3=value3+input4=value3...)

[1] Kurowica & Cooke (2006). *Uncertainty analysis with high dimensional dependence modelling*. John Wiley & Sons. (page 193)

[2] Iooss & Lemaître (2015). A review on global sensitivity analysis methods. In Uncertainty management in simulation-optimization of complex systems (pp. 101-122). Springer.

The 100 best classifications



Outline

Conclusions

- **It is possible to make boundary layer classification with unsupervised clustering.**

- **Dataset preparation parameters are more influential than classification ones.**

- **Scores help to find best settings:**
  - Medium resolution, below 2000m
  - Avoid single and centroid linkage

- **But expertise on the data and the processes are still required to make a really relevant classification.**

Prospects

- **Extend the experiment to larger dataset.**
- **Publish an open source spinnet of code with data.**
  → `https://github.com/ThomasRieutord/bl-classification`
- **Derive more user-oriented products (automatic boundary layer height detection, standard classification product...)**

  → See poster: *Machine learning for boundary layer height detection from aerosol lidars*, Tiago Machado et al.

References

- Calinski & Harabazs (1974): A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1), 1-27.
- Collaud-Coen et al. (2014): Determination and climatology of the planetary boundary layer height above the Swiss plateau by in situ and remote sensing measurements as well as by the COSMO-2 model. *Atmospheric Chemistry and Physics.*
- Iooss & Lemaître (2015): A review on global sensitivity analysis methods. In *Uncertainty management in simulation-optimization of complex systems* (pp. 101-122).
- Kurowica & Cooke (2006): *Uncertainty analysis with high dimensional dependence modelling*. John Wiley & Sons.
- Stull (1988): *An introduction to boundary layer meteorology* (Vol. 13). Springer Science & Business Media.

Météo-France & CNRS

thomas.rieutord@meteo.fr

www.umr-cnrm.fr

www.meteofrance.com | @meteofrance