XI giornata sulla modellistica in aria(net)
Milano, 11 aprile 2024

Machine Learning non supervisionato per il clustering di giorni meteorologici

Applicazione con le Self Organizing Maps (SOMs)

Umberto Giuriato, Daniela Barbero

The Motivation

Need

*More and more often, customers require air quality assessments on **long periods** of time (5+ years).*
*This implies long execution times for dispersion models, which may become a bottleneck on the **delivery time**.*
*We need a method to **fasten** the generation of long-term statistics (average, percentiles)*

Possible solution

*A **Machine Learning** algorithm that automatically selects **the most representatives days**, which will be the only ones simulated to get the long-time statistics*

Algorithms that perform **clustering** of data, i.e. splitting them in groups electing a representant for each group are a kind of **Unsupervised Machine Learning**

Self Organizing Maps (SOMs) are an example of such algorithms

The Dataset for Learning

The training dataset is a tabular dataset with ***N samples*** (the days to select) and ***D features***.

SAMPLES:
The days to cluster

Date	**WIND SPEED 1h**	**WIND SPEED 2h**	**…**	**TEMPERATURE 23h**	**TEMPERATURE 24h**
01-01-2022	6.52	6.41	…	274.15	275.27
02-01-2022	4.22	4.99	…	276.24	275.65
…	…	…	…	…	…
30-12-2022	2.15	3.00	…	273.21	273.20
31-11-2022	5.65	5.15	…	275.65	276.54

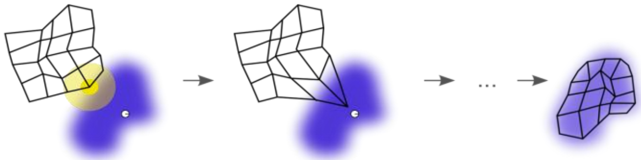
Each day is a vector of meteo features.

Daily periodicity accounted!

FEATURES: Meteorological variables for each hour in the day:
point extractions from meteo fields / observations

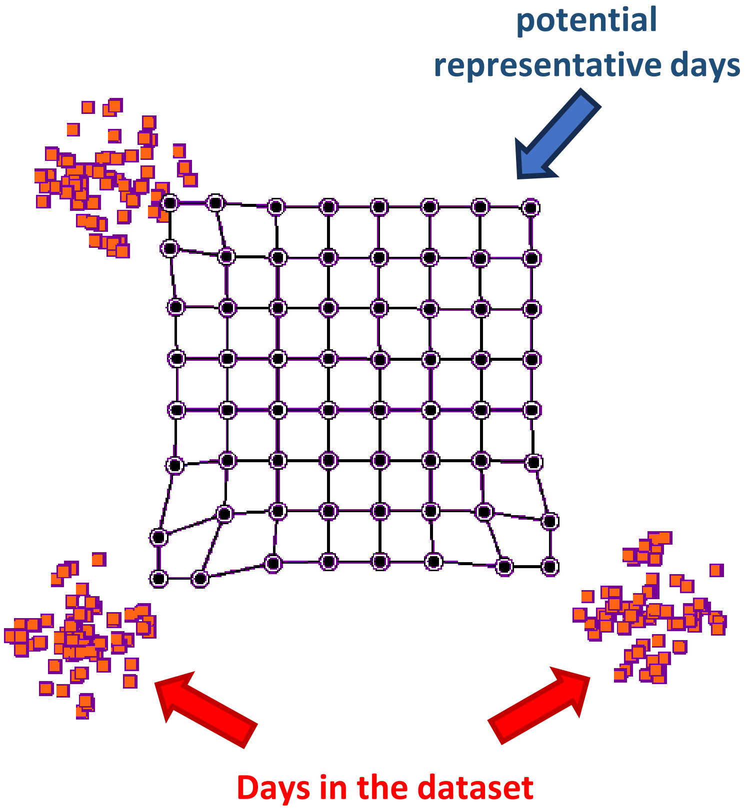
Self Organizing Maps - Learning Process

*Self Organizing Maps (**SOMs**)*
an unsupervised learning technique that can be used for clustering.



*They are **Artificial neural networks** that reduce the dimensionality of a dataset mapping it into a **2D grid***

Each potential
representative day (*vector in feature space*)
is iteratively pushed towards the
days in the dataset (*vector in feature space*)
dragging its neighbors with itself,
until the original dataset is covered

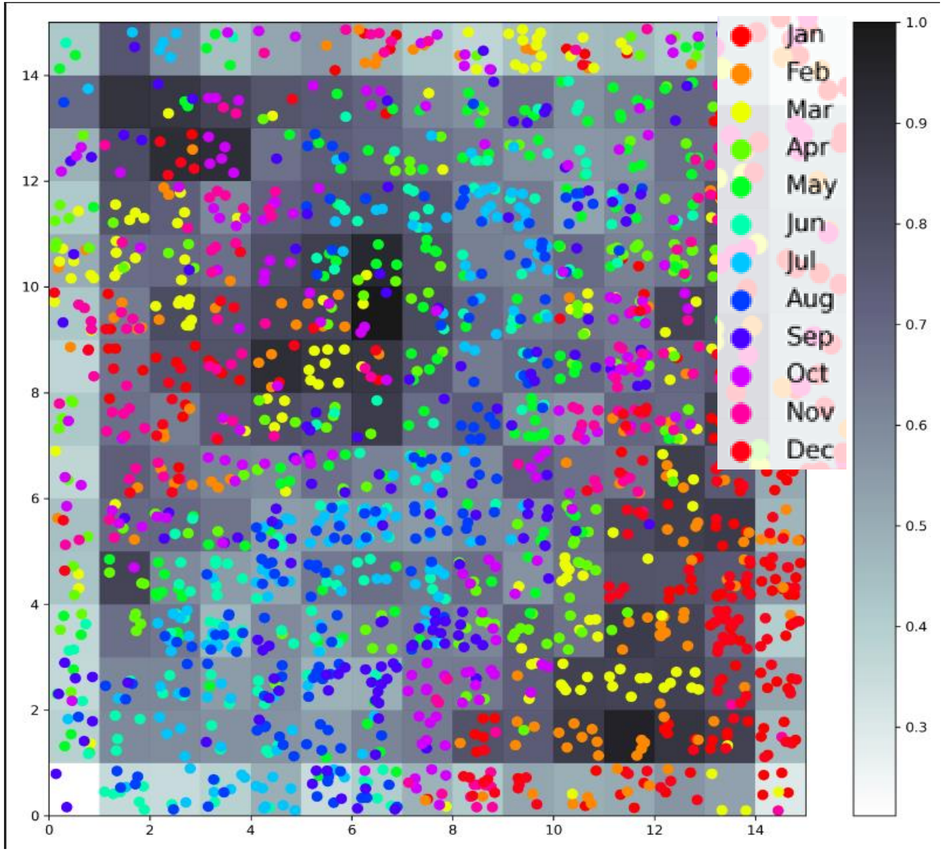


Unsupervised Learning for days clustering

Example of a trained SOM

***Variables**: wind speed, wind direction, temperature, pressure, RH*

DISTANCE UNIT MAP



SOM 15x15

Selection of *225* representative days over *5* years of meteo data at an industrial plant point

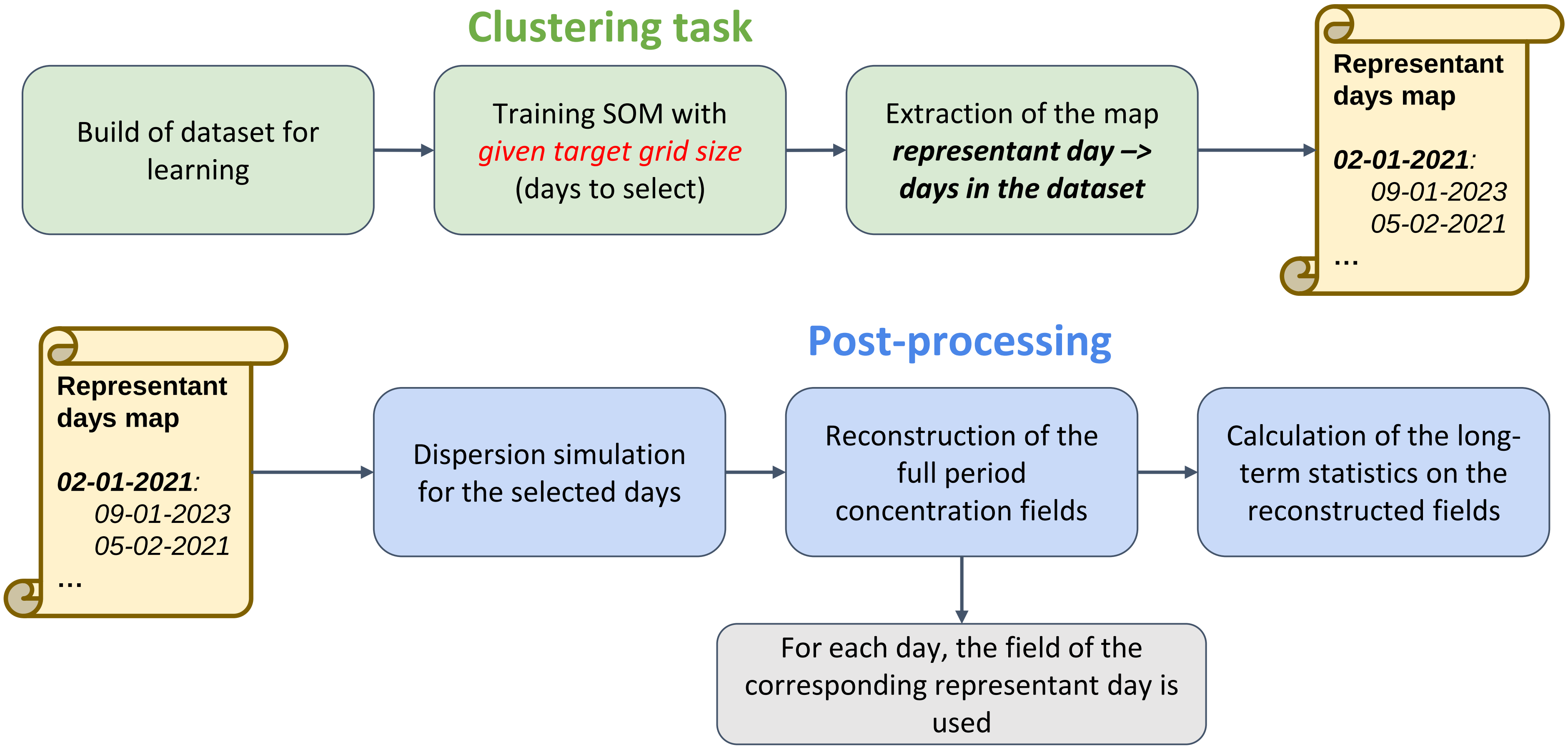
Each cell is the neuron of the representative day

- Days of the same month (also belonging to different years) fall in the same cluster
- Representative days close to each other also belongs to the same season

Meteorological years are similar to each other
Days in a season are similar to each other

Unsupervised Learning for days clustering

Workflow in “production”

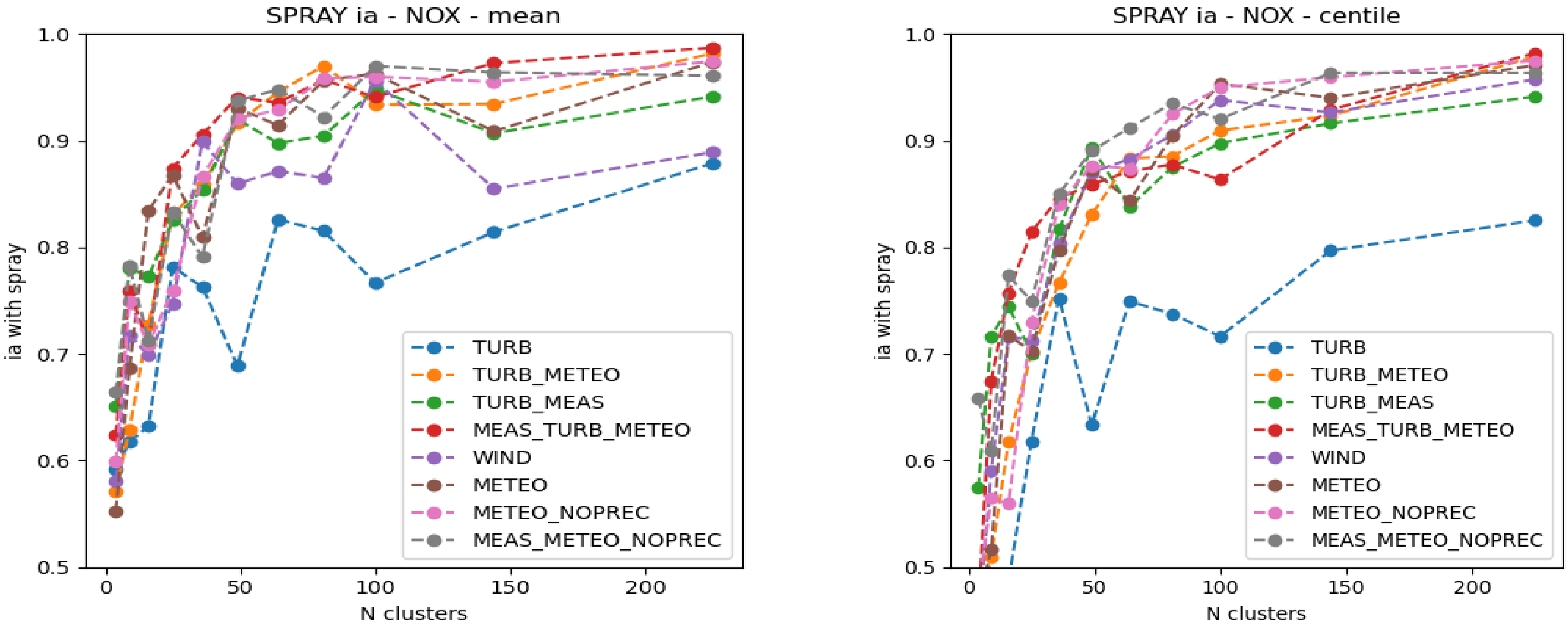


Unsupervised Learning for days clustering

Performance evaluation [NO_X]

ARIANET

Index of Agreement between full SPRAY simulation and SOM reconstruction – 1 year statistics – plant in the Po Valley

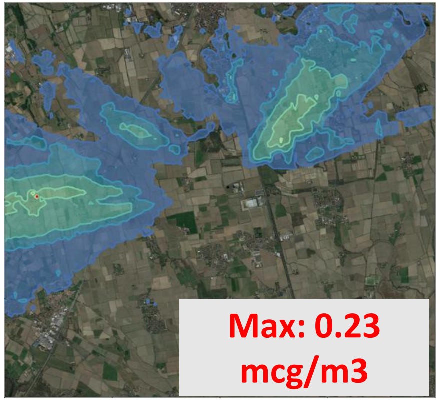


- *No dramatic difference in selecting different feature sets*
- *A variable with yearly periodicity improves the performance (not just wind but also temperature)*
- *Elbow point at around 75 days*

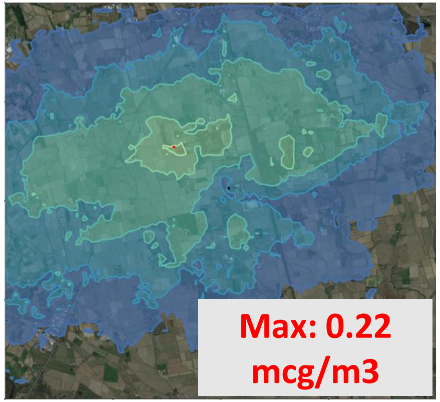
Unsupervised Learning for days clustering

Performance evaluation: Average [NO_X]

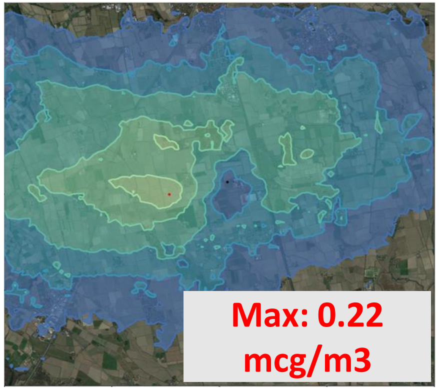
SOM 4x4

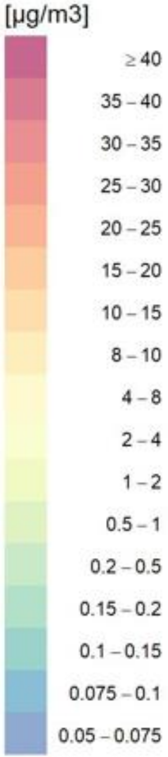


SOM 9x9



SOM 15x15



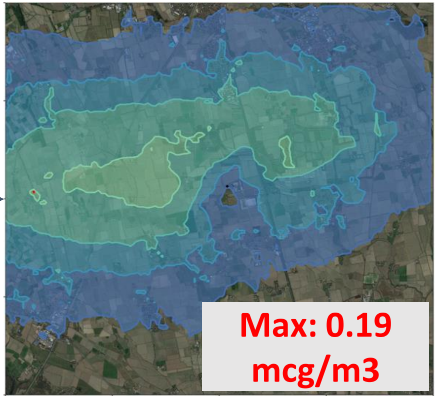


Reconstruction of 1 year simulation

Too few representative days are not able to catch all the plume directions contributing to the mean

After 80 days the difference in the concentration distribution is low

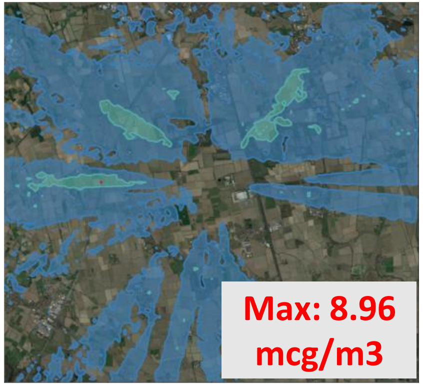
Full SPRAY



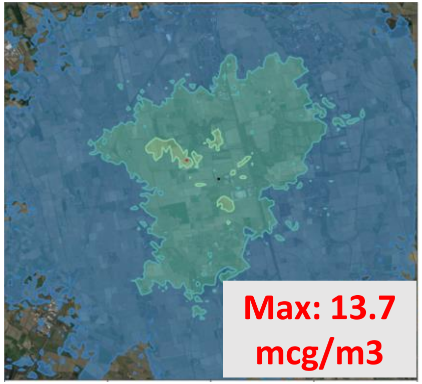
Unsupervised Learning for days clustering

Performance evaluation: 99.8 Hourly Percentile [NO_X]

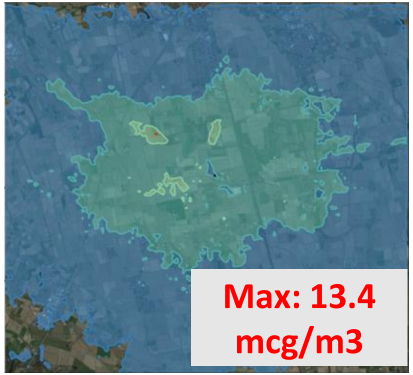
SOM 4x4



SOM 9x9



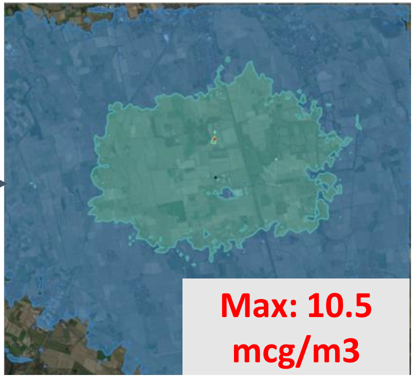
SOM 15x15

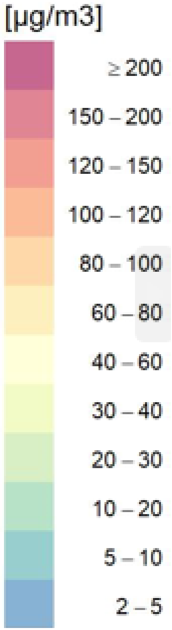


Reconstruction of 1 year simulation

Although increasing the days selected to reconstruct the field, the similarity with the original field is less pronounced compared to the annual average

Full SPRAY

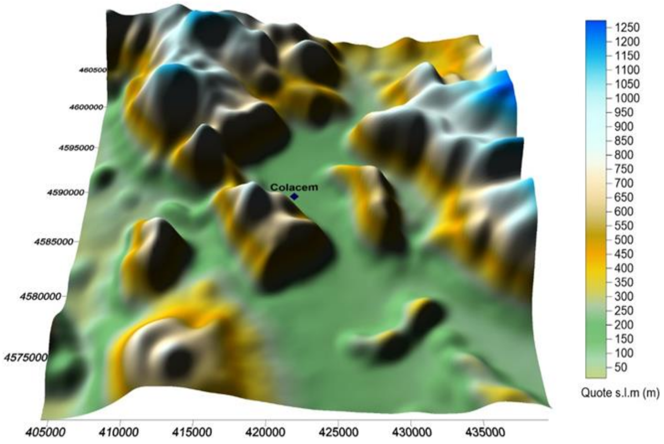


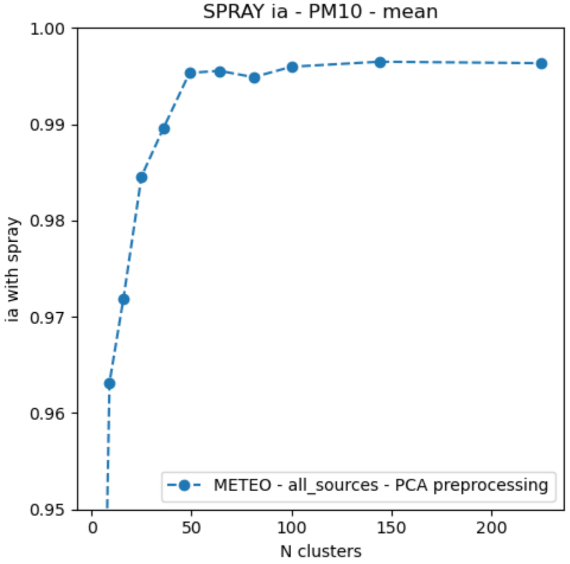


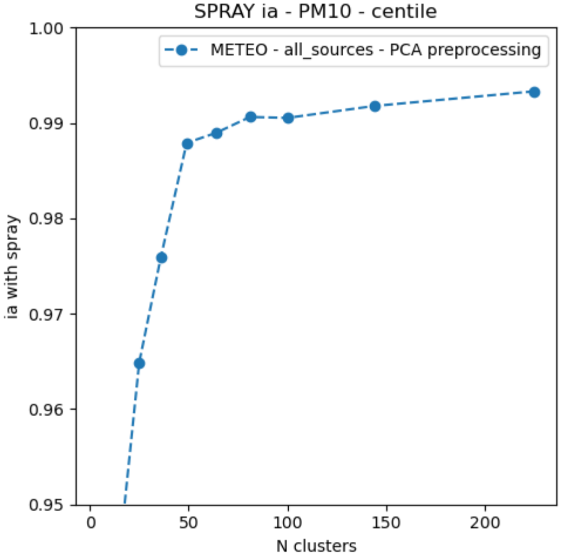
Another case (Sesto Campano)

For comparison, we applied the SOM clustering to another case, with **Different Orography**

Po Valley *vs* ***Hill Territory*** *(the industrial site is in a valley)*



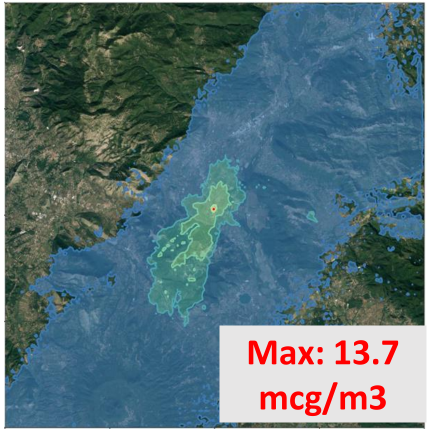




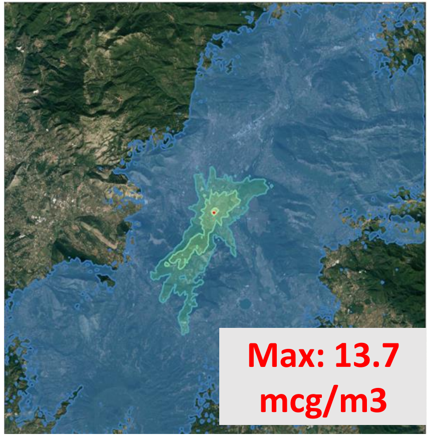
METEO_NOPREC_PCA

Performance evaluation: 90.4 Daily Percentile [PM]

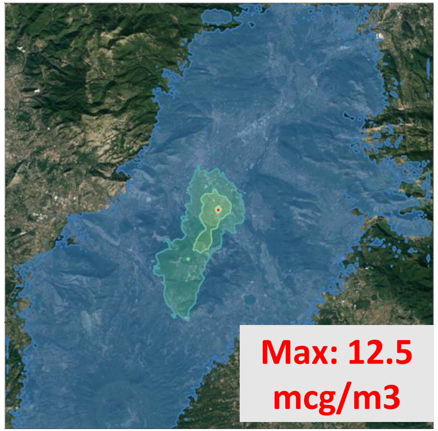
SOM 4x4



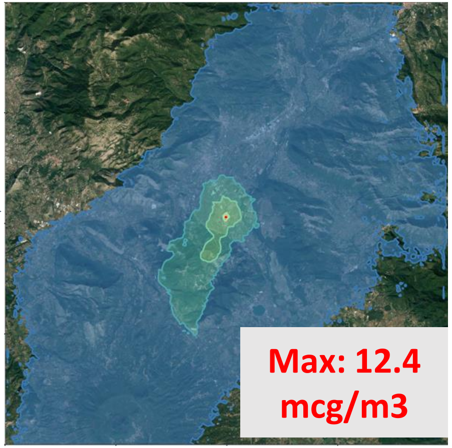
SOM 5x5

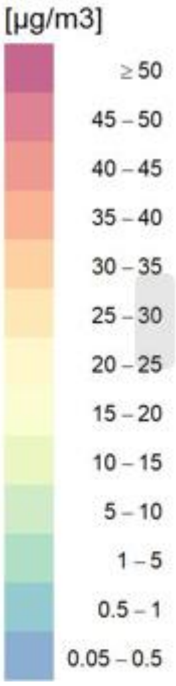


SOM 9x9



Full SPRAY





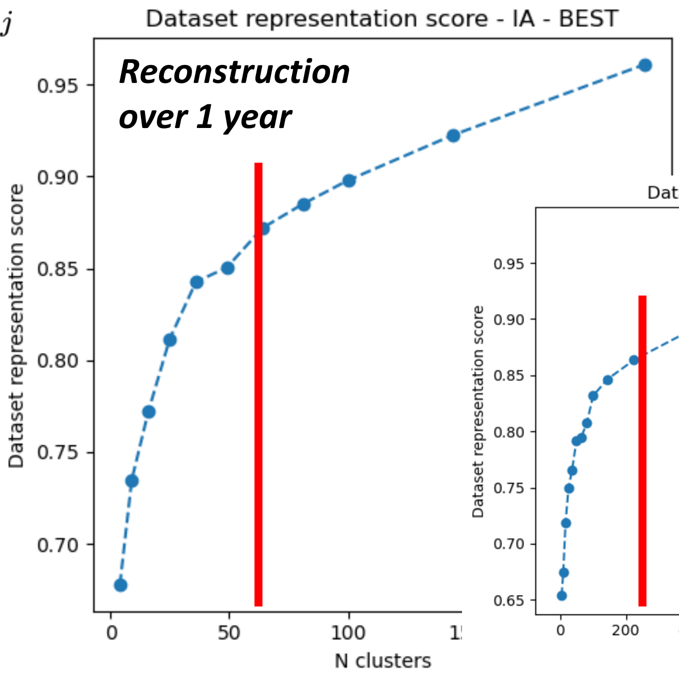
The orography constraints the dispersion on few directions, so that less days are necessary to reproduce the statistics

Dataset representation by clusters

Weighted average of the ***Index of Agreement*** *of reconstructed feature time series, where the weights are the* ***PCA loadings*** *for each feature.*

$$Score = \frac{1}{\sum_{j=1}^{n} PC_{\text{load}}^{j}} \sum_{j=1}^{n} PC_{\text{load}}^{j} \times IA_j$$

- The dataset representation score reproduces the functional form of the Index of Agreement with SPRAY.
- The elbow-point after which is not convenient to add representative days it is consistent with what observed



Conclusions

- **Few meteorological attributes**, but not limited to wind are sufficient to get decent clustering
- Clustering is much better if the **orography** constraints the dispersion in privileged directions
- **Dataset representation score** is a good metric to evaluate the number of days to simulate
- The **yearly periodicity** of meteo attributes makes the method suitable for the selection of days over long-range periods
- **MAIN TO DO**: Find a way to embed full 2D meteo fields as features for SOM clustering (eg use variational autoencoders to embed fields in a low dimensional latent space…)

Thank you for your attention !

Appendix

APPENDIX

Workflow in "production"

Clustering task

Selection of dataset for learning, with **standard scaling** of input feature vectors → Optimization of the SOM **hyperparameters** (*kernel size* and *learning rate*) → Training SOM with *given target grid size* (days to select) → Find the ***representant day***, whose features vector is closer to every neuron weight

Post-processing

Dispersion simulation for the selected days → Reconstruction of the full period concentration fields → Calculation of the long-term statistics on the reconstructed fields

Reconstruction of the full period concentration fields → **BEST**: For each day, the field of the corresponding representant day is used

Reconstruction of the full period concentration fields → **POND**: For each day, the weighted average of the fields of the representant days are used

$$W_{ij} = \exp\left(-d_{ij}\right)$$

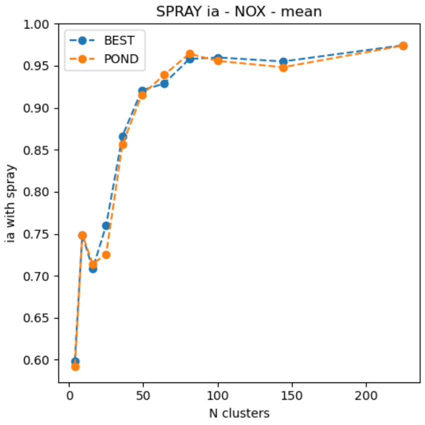
$$i \in [1, N]$$

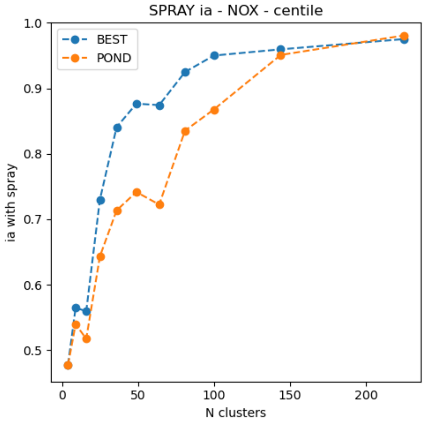
$$j \in [1, K_1 \times K_2]$$

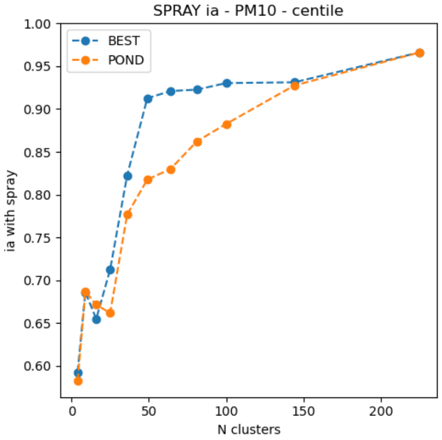
Unsupervised Learning for days clustering

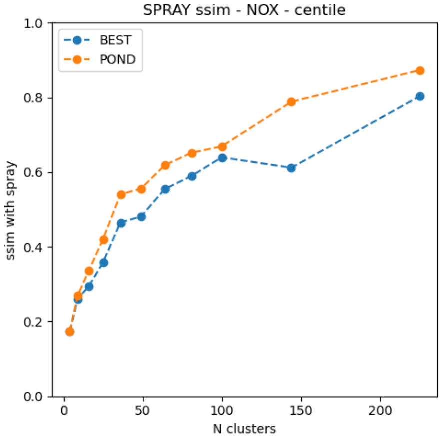
MEAS_METEO_NOPREC_PCA

Performance evaluation: POND vs BEST









- No appreciable difference is found in the annual average reconstruction if POND or BEST method is used.
- The percentiles are better reproduced by BEST reconstruction, according to IA. The reason may be that the average, although ponderate, tend to flatten the values, cutting the tails of the distributions
- However, the similarity among images is slightly better when POND reconstruction is used

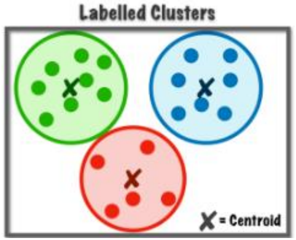
Dataset representation by clusters

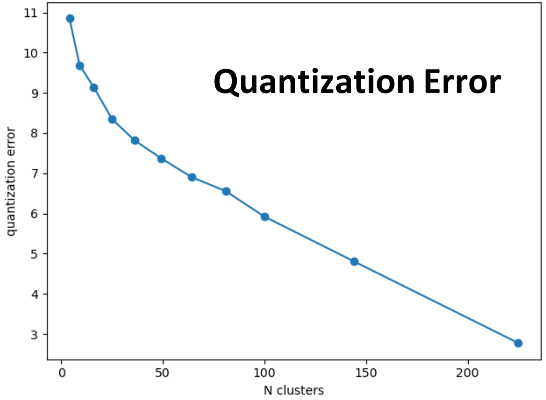
MEAS_METEO

Since the target variable is not known at training time, we need to understand the quality of clustering with just the features alone

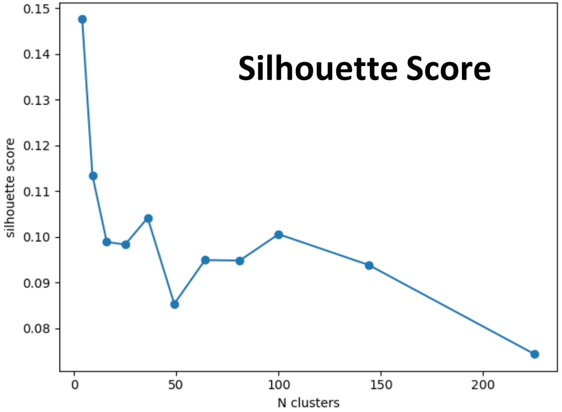
The **quantization error** is an estimate of how well the points in a cluster are well represented by the winning neuron

$$QE = \frac{1}{N}\sum_{i=1}^{N}(\|x_i - m_{ci}\|)$$





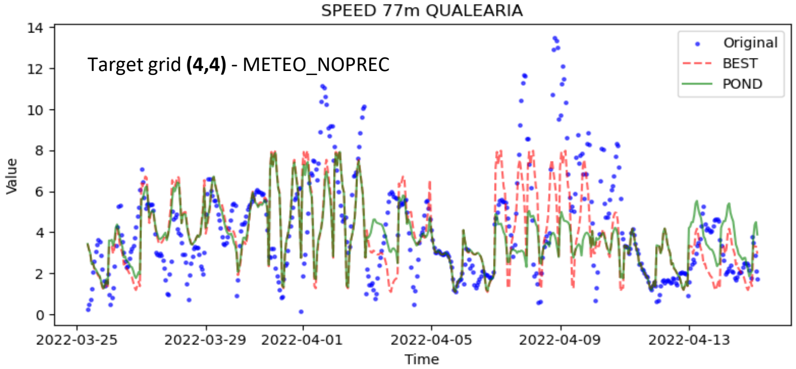
The **silhouette score** quantifies how similar a point is to its own cluster (cohesion) compared to other clusters (separation).

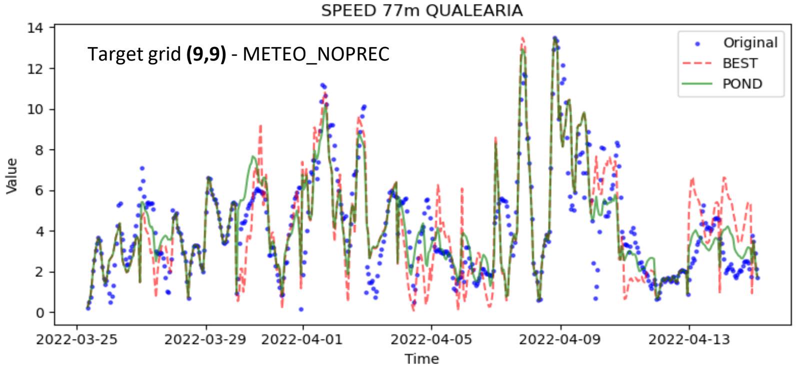


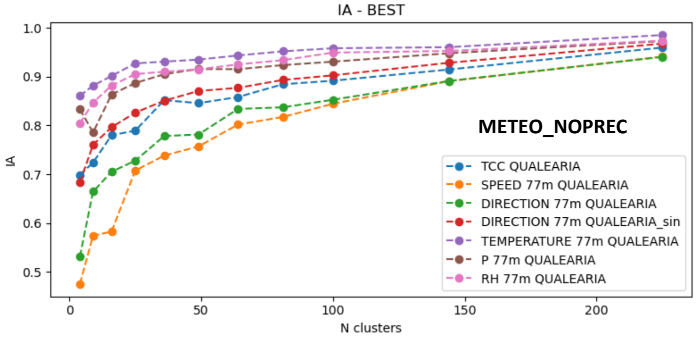
Quantization error keeps on decreasing as clusters number increases, without showing a clear elbow.
Silhouette score has a low value despite the clusters number.
This indicates that clusters have not a well defined border (non-convex)

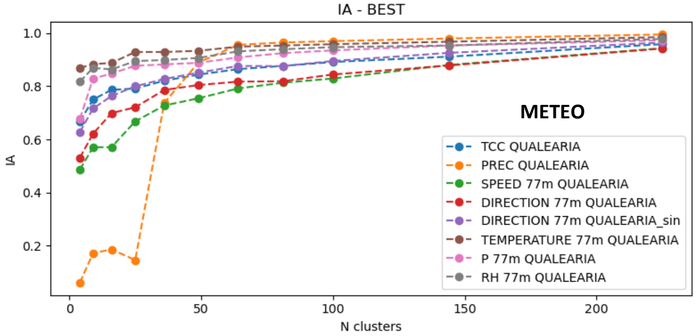
Reconstruction of the feature time series

We can reconstruct the original (hourly) time series in the same way we reconstructed concentration fields





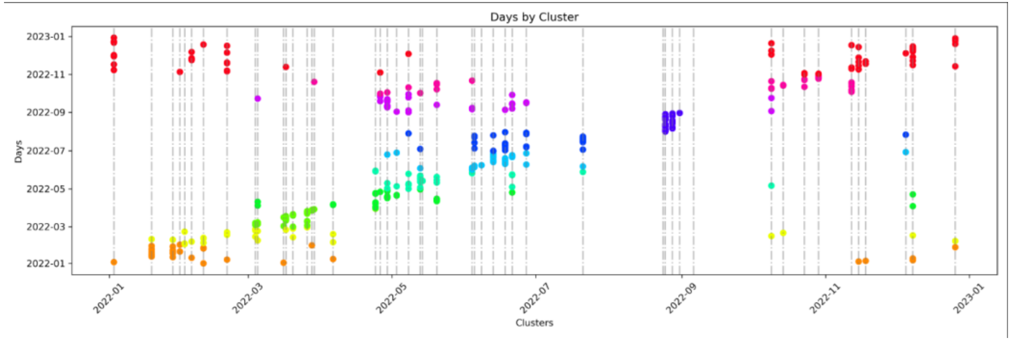


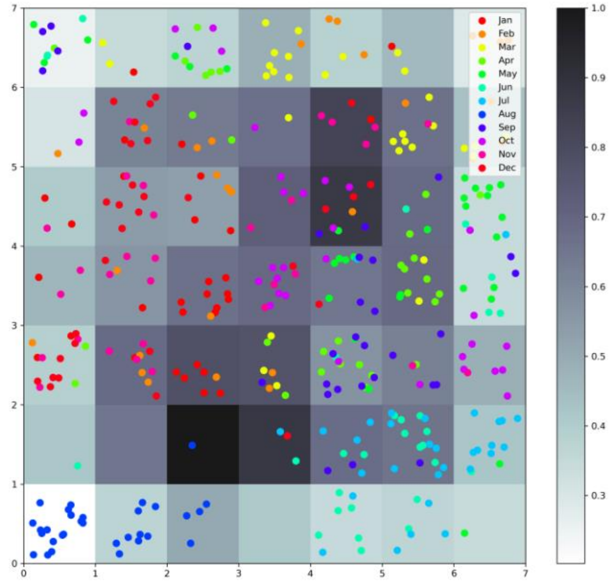


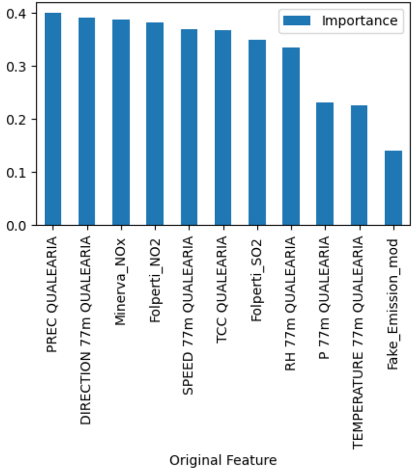
- The IA of reconstructed time series shows elbows at a given cluster number
- **WARNING**: Using high-variable features like PRECIPITATION, leads to abrupt IA improvement at a certain cluster number. But this is not useful for decision making, since PRECIPITATION is not an important variable to reconstruct SPRAY concentrations.

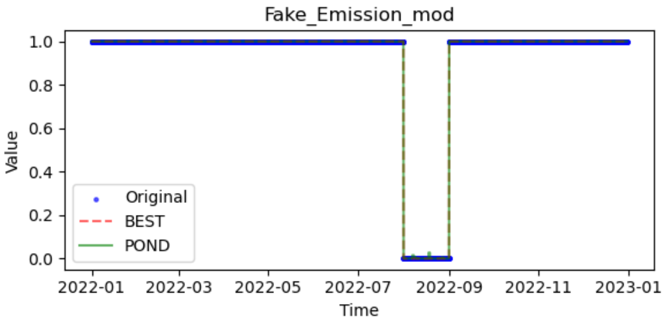
What about the emissions? Stretch experiment

In order to check the procedure against emission modulations, we added as variable the switch-off of the source for the month of August









- The algorithm places august days in the same clusters and the fake modulation is reconstructed
- The PC loadings of the Fake Emission modulation is low, because it is a variable with low variability. Still, such feature must be included because deeply connected to concentration.

The curse of dimensionality

An increase in the number of dimensions of a dataset means there are more entries in the vector of features that represents each observation in the corresponding Euclidean space

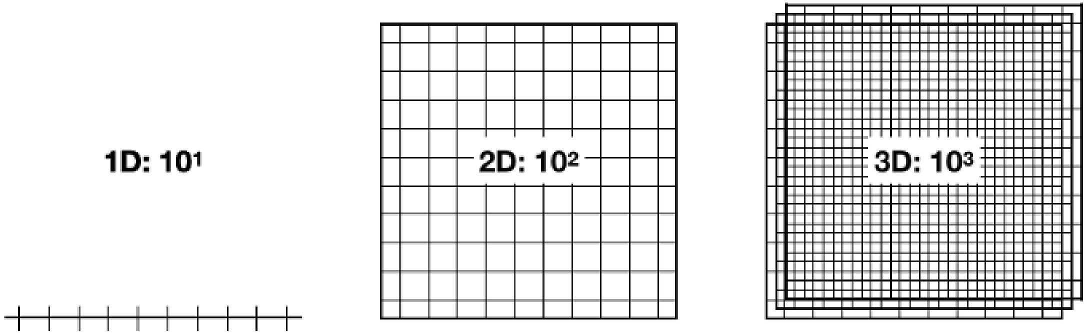
Adding a dimension implies adding a (positive) term in the sum inside the definition of Euclidean distance

$$d(p,q) = \sqrt{\sum_{j=1}^{d} (p_j - q_j)^2}$$

As a consequence, distances take larger values on average and the distance space becomes sparser.

A large average distance implies that the "difference" between different couple of data-points is more vague, making harder for clustering algorithms to perform well.

In a large-dimension space, more data points are needed to keep the average distance constant.

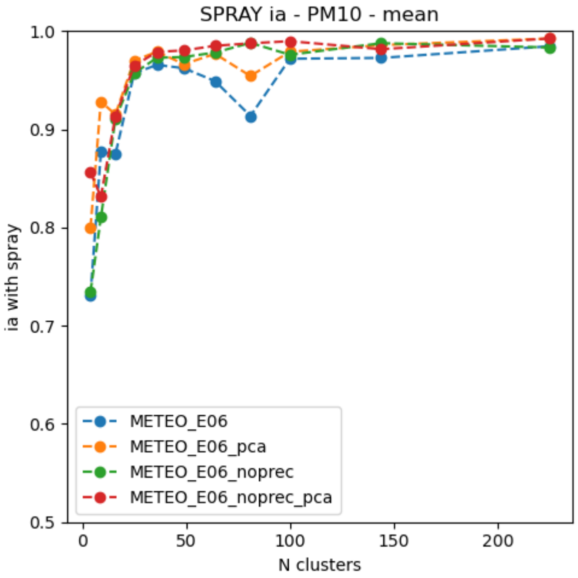


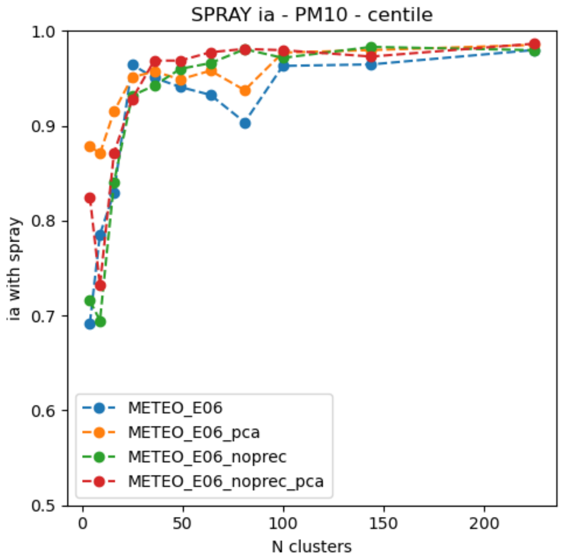
Unsupervised Learning for days clustering

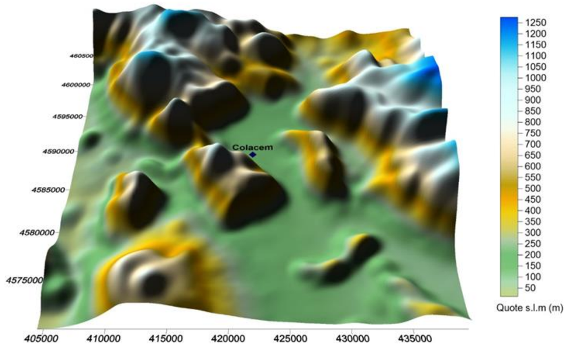
Another case: Colacem simulation (Sesto Campano)

For comparison, we applied the SOM clustering to another case. The main differences with Vellezzo Bellini are:

- ***Different orography***: Pianura Padana vs hill territory (the industrial site is in a valley)
- ***Different emissions***: Some sources are not modulated, some others are modulated







E06 is a not modulated source.

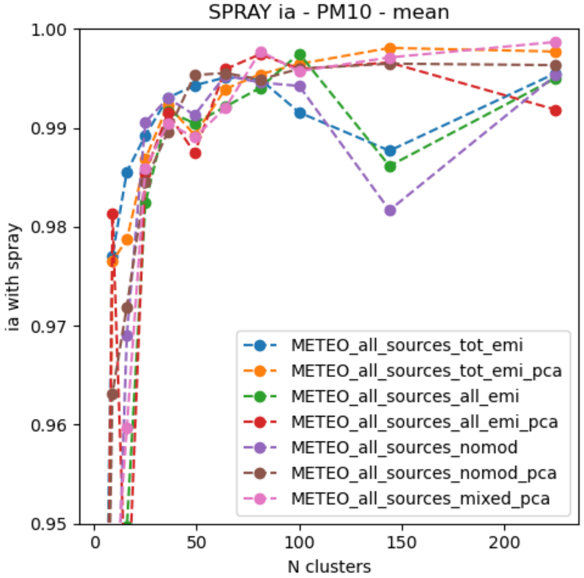
- The saturation to high performance takes place at an inferior number of cluster compared to Vellezzo Bellini
- The percentile is also better resolved
- Removing precipitation still improves the performance

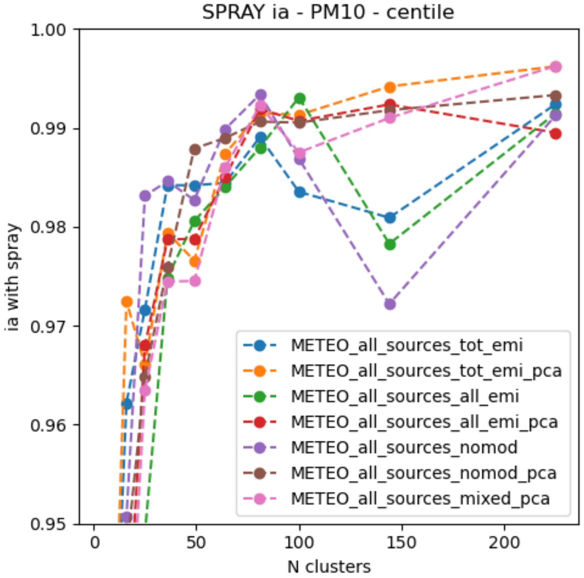
Unsupervised Learning for days clustering

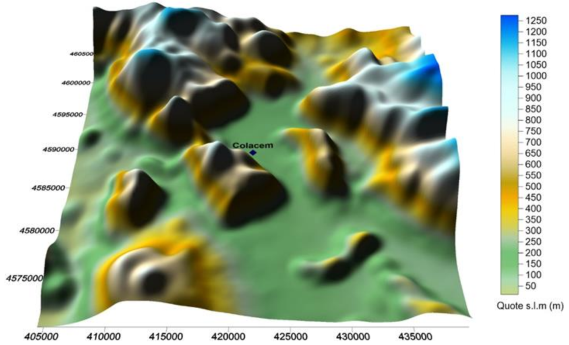
Another case: Colacem simulation (Sesto Campano)

For comparison, we applied the SOM clustering to another case. The main differences with Vellezzo Bellini are:

- ***Different orography***: Pianura Padana vs hill territory (the industrial site is in a valley)
- ***Different emissions***: Some sources are not modulated, some others are modulated







The case with modulated emissions is considered:

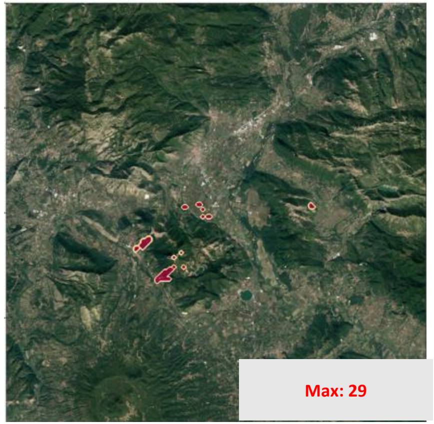
- Considering modulations as features does not improve the performance
- Still, the overall performance borders on perfection

Unsupervised Learning for days clustering

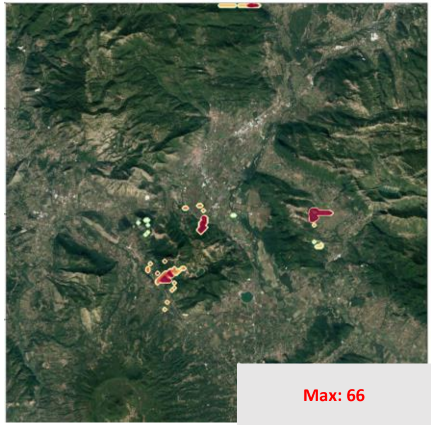
METEO_NOPREC

Performance evaluation: Overcomings 200 mcg/m3 [NOx]

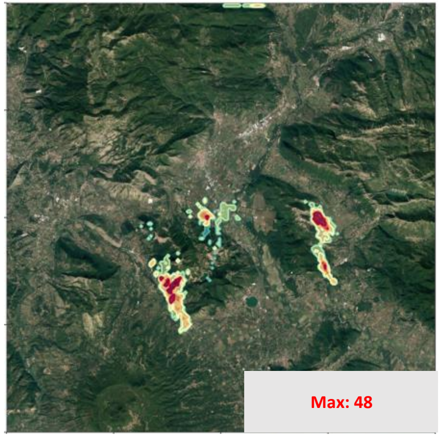
SOM 4x4

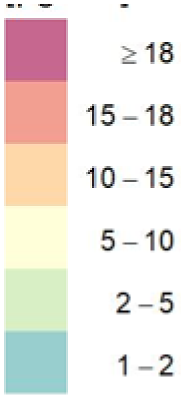


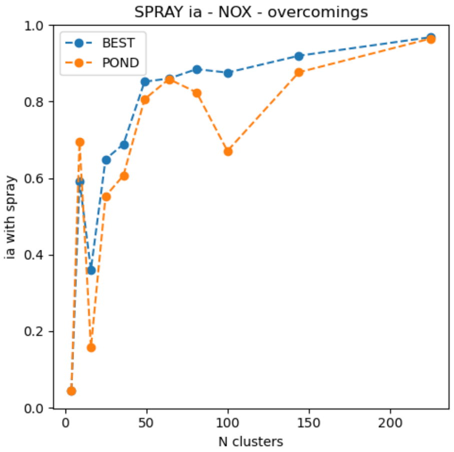
SOM 5x5



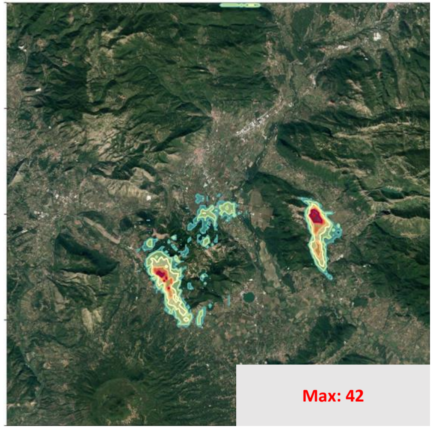
SOM 9x9



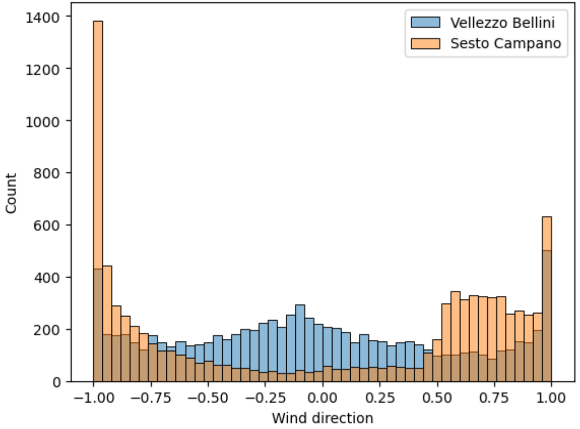


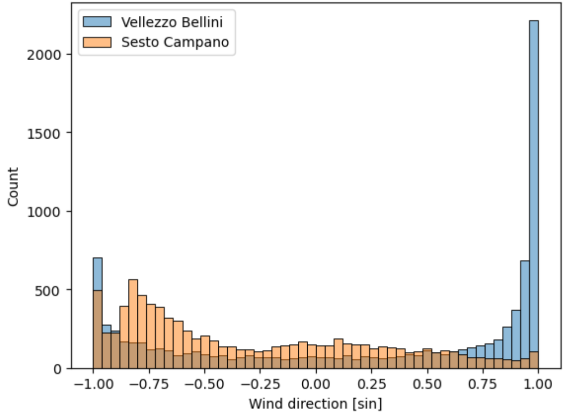


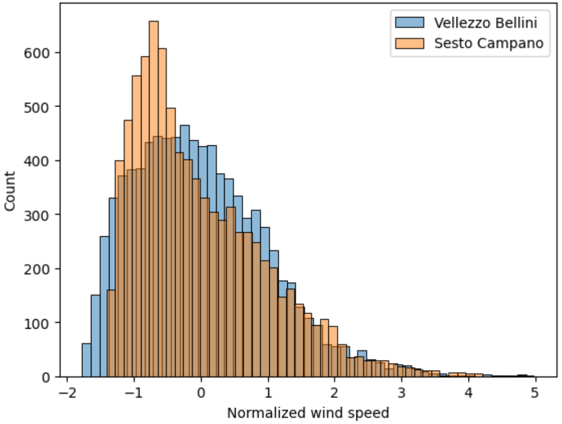
True Field →



Feature distribution comparison

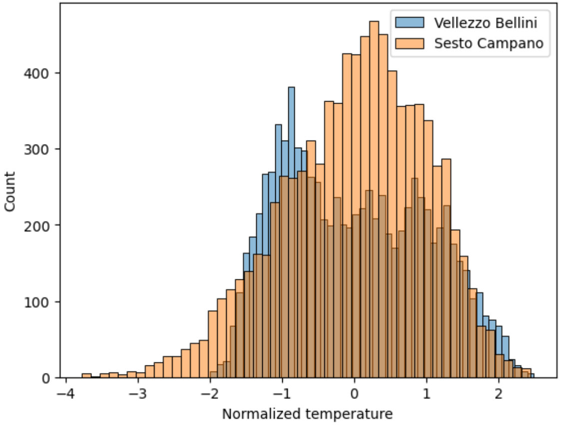






To grasp the origin of the different performance, we compare the feature distributions of the two cases:

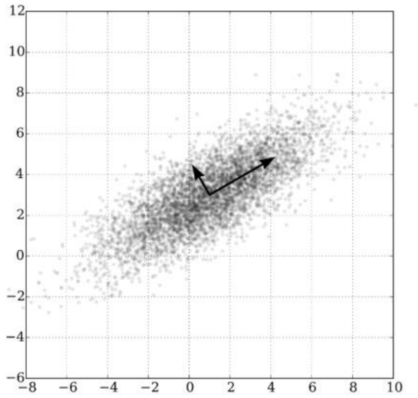
- Although all the variables show some differences in the distributions, the most evident is the **wind direction**.
- Due to its geographical location, the wind at the Colacem site in Sesto Campano is polarized along the valley
- For this reason, less days are necessary to represent the dataset distribution

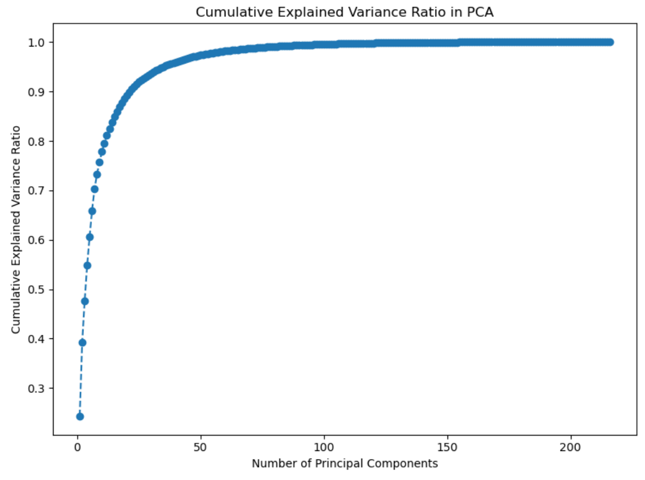


Dimensionality reduction: PCA

Principal Component Analysis is a dimensionality reduction technique, that allows to project linearly the features onto the *directions of decreasing variance in the dataset*. Each PC will explain a percentage of the dataset variance.

More technically, PCA is a linear decomposition: principal components are the eigenvectors of the ***covariance matrix*** and the corresponding eigenvalues are the explained variance by each principal component.





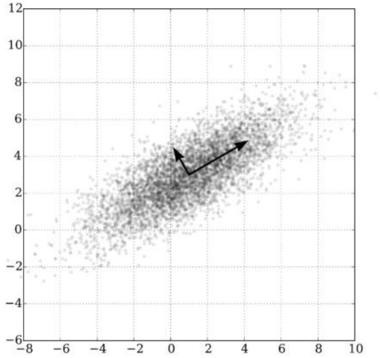
NB: In order to reduce the dimensionality of the dataset, the SOM clustering can be applied directly on the Principal Components, that cumulatively explain a large fraction (90%) of the dataset variance. We observed that this tends to stabilize the functional form of the performance across cluster numbers, but without tremendous effects

Estimation of feature importances: PCA

Principal Components can be used to estimate the representativeness of the features in the dataset.

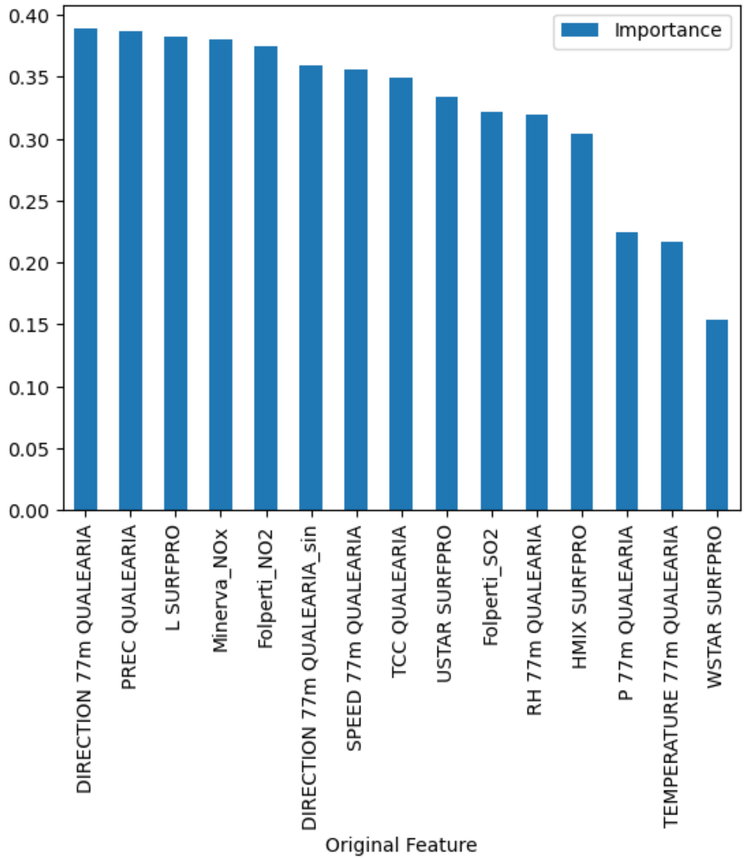
The ***PCA loadings*** are the contribution of the Principal Component on each original feature

$$PC_{\text{load}}^{j} = \sum_{i=1}^{D} e_{ij} \sqrt{\text{ExplainedVar}_i}$$



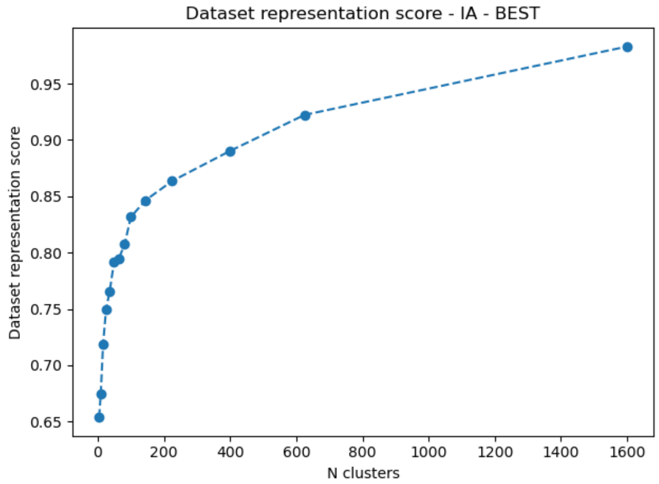
These numbers cannot be used as a black box.

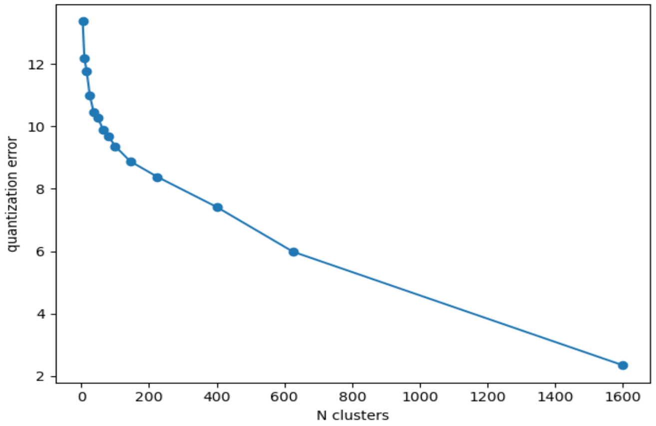
For instance PRECIPITATION has a high importance in the dataset variance, but small correlation with SPRAY concentration (just a small effect via wet deposition)

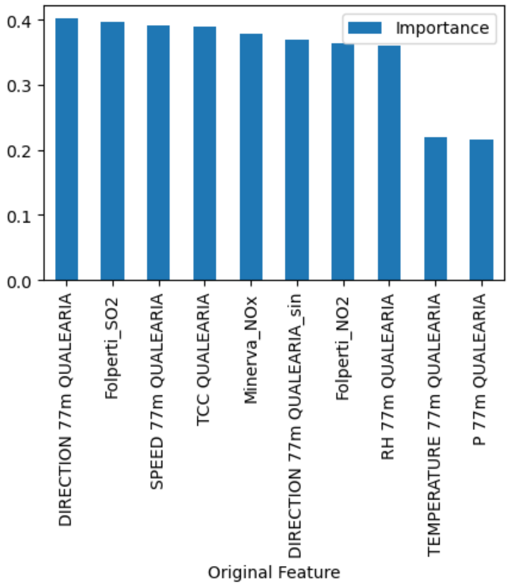


(*) A further average across features of the same variable but at different hours is performed

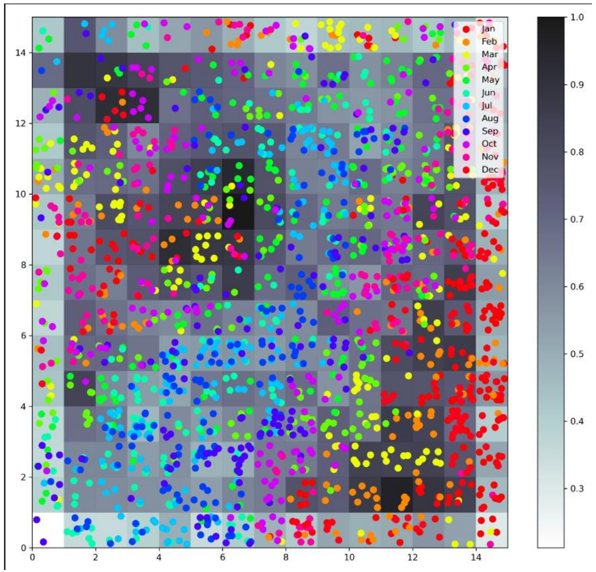
Reconstruction over 5 years







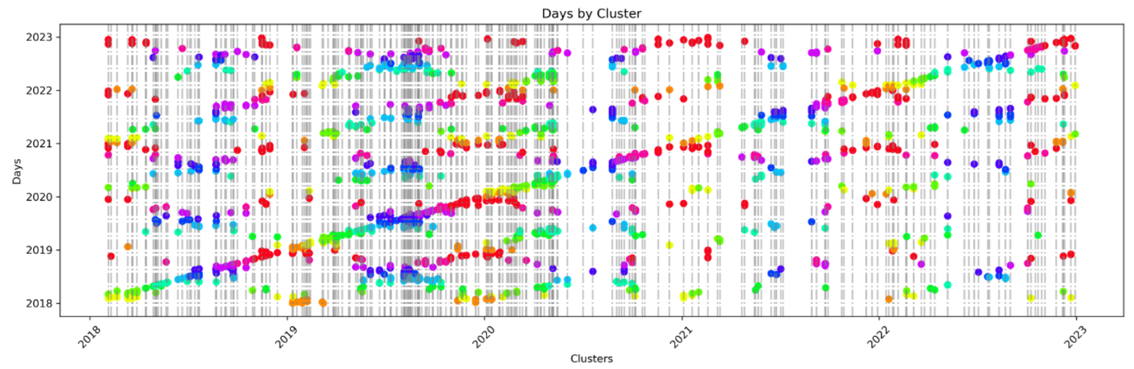
SOM 15x15

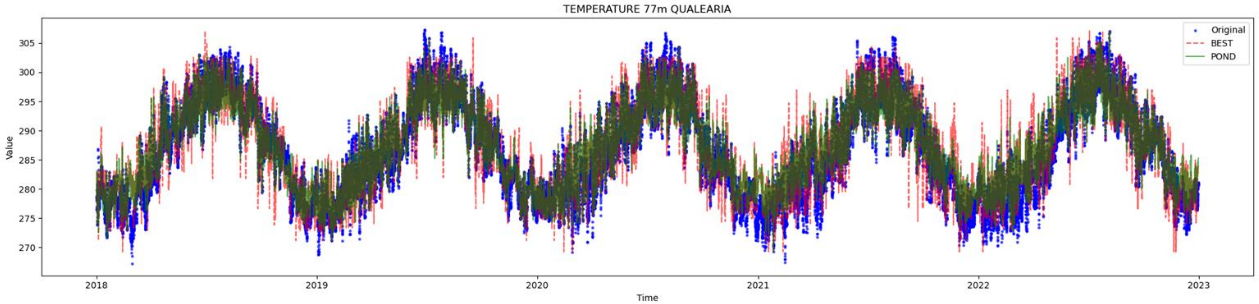


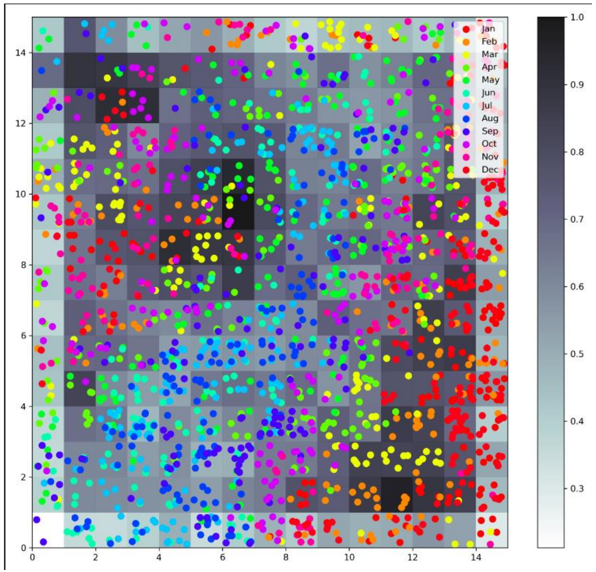
- Extending the clustering to 5 years, we observe a similar dataset representation score: we obtain again an elbow curve, with the elbow point at around **200 days**
- Also the Quantization Error shows a better elbow than the one-year case, indicating clustering improvement.
- The explanation of clustering improvement may lie in the annual periodicity of the original variables. ***Close days of different years fall in the same cluster***

Unsupervised Learning for days clustering

Reconstruction over 5 years



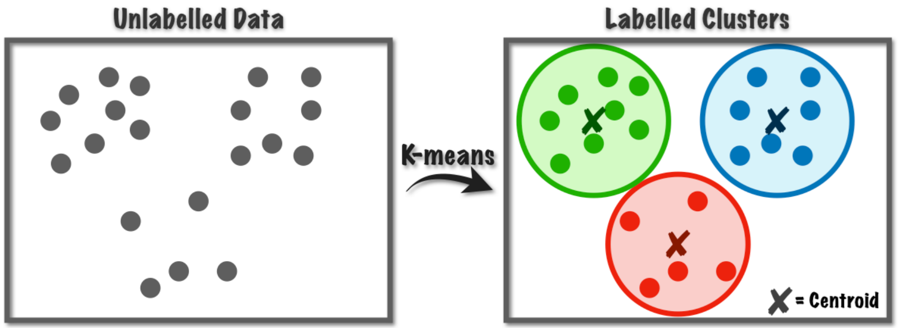




- Extending the clustering to 5 years, we observe a similar dataset representation score: we obtain again an elbow curve, with the elbow point above 200 days
- Also the Quantization Error shows a slight elbow, suggesting that for sure selecting less than 200 days will give bad clustering
- The explanation of clustering improvement may lie in the annual periodicity of the original variables. Close days of different years fall in the same cluster

K-Means clustering

K-Means is one of the simplest clustering algorithm.



Given a dataset and a number of cluster into which to partition it, the algorithm:

1 Initialize randomly the cluster representative units (random D-dimensional vectors)

2 Assign each sample of the dataset to a representative unit, choosing the closest one (according to Euclidean distance)

3 Updates the representative units as centroids of the samples assigned to it

4 Repeat steps 2 and 3 until convergence