Clustering meteorological scenarios

Y. Barrera, A. Somacal, A. Umfurer
M. Jonckheere, V. Lefieux, D. Picard, E. Smucler
y.barrera@aristas.com.ar

Abstract

The fluctuations in the temperature have a strong influence in the electric consumption. As a consequence, identifying and finding groups of possible climate scenarios is useful for the analysis of the electric supply system. The scenarios data that we are considering are time series of hourly measured temperatures over a grid of geographical points in France and neighboring areas. Clustering techniques are useful for finding homogeneous groups of times series but the challenge is to find a suitable data transformation and distance metric. In this work, we used several transformations (fourier, wavelets, autoencoders) and distance metrics (MLCC and euclidean) and found consistent groups of climate scenarios using clustering techniques (k-medoids and k-means). We found that k-shape performs the best according a within cluster dispersion index. This is a joint work with RTE (Réseau de Transport d'Électricité), the electricity transmission system operator of France.

Clustering methodology

Clustering can be performed defining the following steps:

1. **The representation of the data.** We consider functional approximations with Fourier and Haar basis, a data based approximations such as PCA and a nonlinear representation learn by an autoencoder.

2. **The distance used to measure dissimilarities between the representations of the data.** We consider the euclidean and the max lagged cross correlation distance(MLCC) distances: The MLCC [1] seeks for an optimal alignment between two signals, with the two series only being allowed to be aligned via shifts in the time axis. The cross-correlation is normalized so that two series can be meaningfully compared.

$$SBD(z,w) = 1 - \max_{|k|\leq k_{max}} \frac{\text{corr}_k(z,w)}{||z||.||w||}$$

   where z, w are times series, corr_k is the lagged cross-correlation at k lags and k_{max} is a user-specified constant.

3. **The method to partition the data into clusters.** We consider the classical algorithms k-means and k-medoids.

4. **Evaluation** In order to evaluate the clustering results we use the within index, that is defined as the within cluster variance over the global variance.

Representation of data	Distance	Clustering method	Name
Embeddings from autoencoders	euclidean	k-medoids	autoencoder
Plain time series	euclidean	k-medoids	ED
PCA 0.95	euclidean	k-medoids	PCA95
Fourier 0.95	euclidean	k-medoids	Fouier95
Haar 0.95	euclidean	k-medoids	Haar95
z-normalized time series	MLCC	k-means	KShape

Introduction

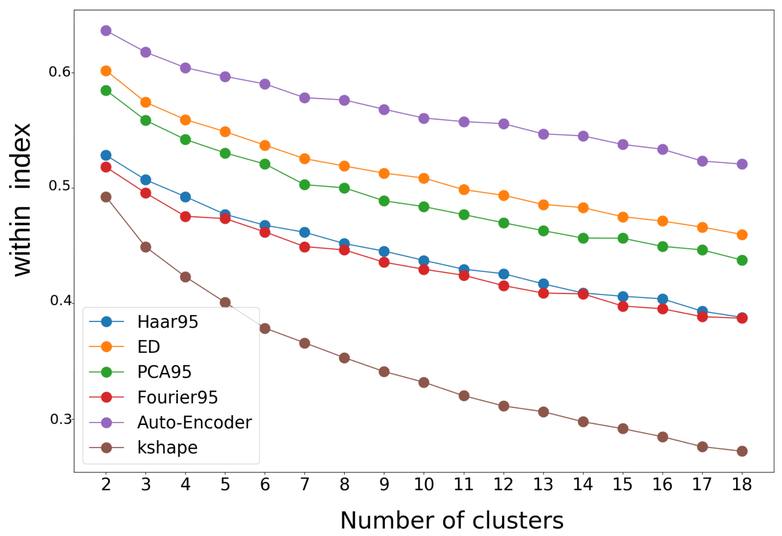
Our task is to find homogeneous groups of time series

Data description

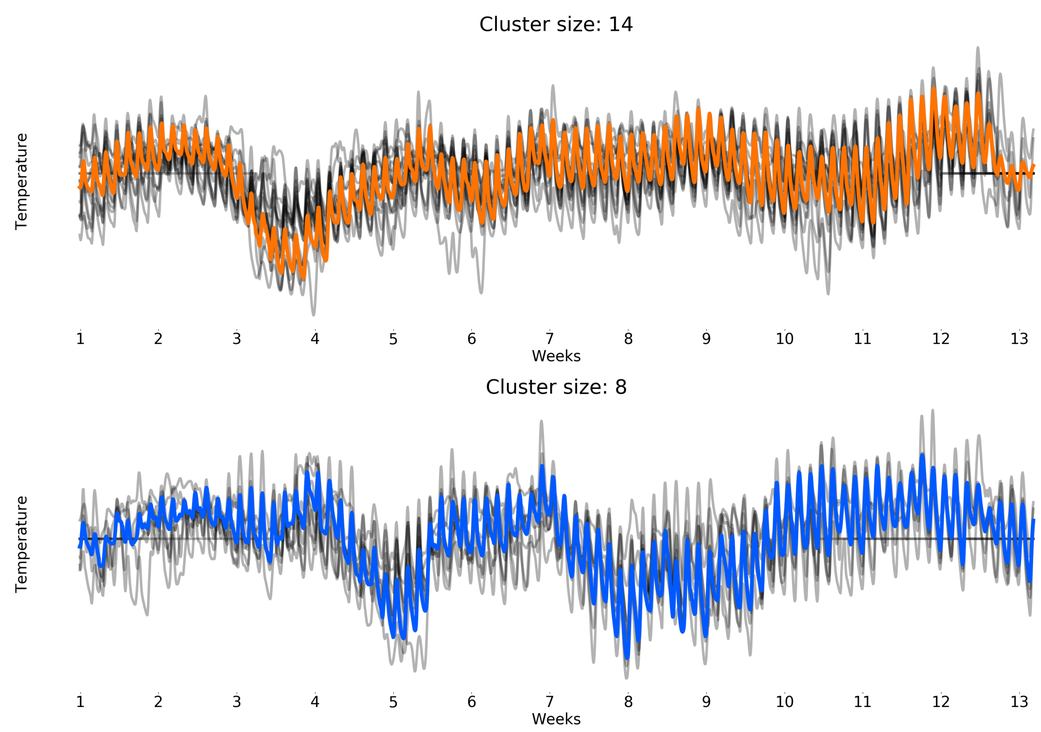
For a fixed geographical point we have 200 climate scenarios consisting each one of hourly temperatures along a winter period (length 24 * 90). These data comes from simulations where the evolution of the atmosphere is simulated. Climate scenarios should be interpreted as sets of possible achievements of 200 years under the same climate. These are neither reanalyzes of past situations nor forecasts. Long simulated climate data series provide a vast sample of meteorological situations.

Results

The following figure has the within index as a function of the number of clusters for each of the models. The lower the index, the better the cluster segmentation. As the number of clusters increases, the within index decreases because the variability inside each cluster decreases.



Two clusters resulting from k-shape algorithm (k=15). In gray all the times series in the cluster and in color the barycenter according to the MLCC distance.



Conclusions and future work

We have presented clustering results on real meteorological data using k-shape algorithm. We considered 200 climate scenarios (temperature times series) and showed the resulting clusters providing a cluster representative that is the barycenter of the cluster series when using the Max Lagged Cross Correlation distance.

We noticed that when the dimension reduction is drastically, the clustering task is easier and thus the clustering indexes improve. We are now working in developing indexes that take into account the clustering quality but also how well the data transformation represents the data.

References

(1) Paparrizos, J. and Gravano, L. *k-shape: Efficient and accurate clustering of time series*. In Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, 2015.