Clustering climate scenarios

Matthieu Jonckheere

People and institutions involved

- Yamila Barrera, Aristas
- Ezequiel Smucler, Aristas, University Di Tella
- Agustin Somacal, Aristas
- Alfredo Umfurer, Aristas

- Leonardo Boechi, UBA-CONICET, Aristas,
- Matthieu Jonckheere UBA-CONICET, Aristas

- Dominique Picard, Université Paris Sorbonne

- Vincent Lefieux, RTE

Aristas.

instituto de cálculo
UBA - CONICET

université PARIS DIDEROT PARIS 7
Université de Paris

Introduction

Climate strongly impacts energy consumption.

Identifying different possible climate scenarios can be instrumental to understand how the electric network would respond to variations in the weather.

RTE gets simulated time series of temperatures over a grid of geographical points in France and neighboring areas.

Objectives

1. Cluster climate scenarios
2. Evaluate and interpret the clustering
3. Give representatives and define the notion of quantiles
4. Get insights on the dynamics of the scenarios

Objectives

1. **Cluster climate scenarios**

2. **Evaluate and interpret the clustering**

3. **Give representatives and define the notion of quantiles**

4. Get insights on the dynamics of the scenarios

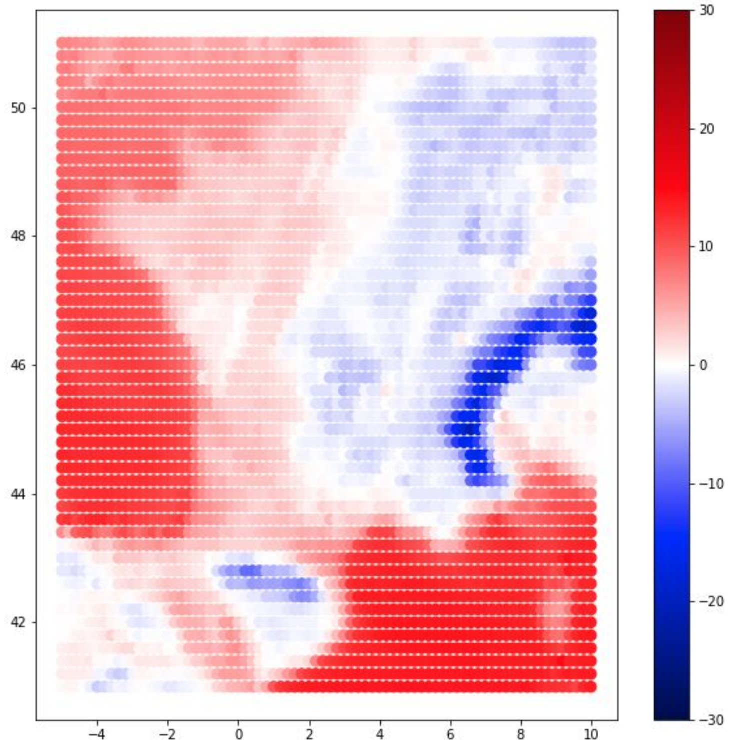
Data

Dimension = 71 x 51 x 8760

For points in a grid of 71 x 51, we have the temperature of every point every for 200 years every hour.

The simulation of 200 years are not forecasts.

They are built to represent the climate of the 1984-2013 period, based on the model Arpege Climat 6.0 and Hirlam reanalysis



Clustering ingredients

1. Choosing a transformation on the data space possibly reducing the dimension, and defining the feature space,

2. Choosing a distance on the feature space,

3. Choosing a clustering algorithm on the feature space.

4. Choosing a distance in the original space and associated criteria to possibly choose between different algorithms using the performances of these criteria.

Clustering tradeoffs

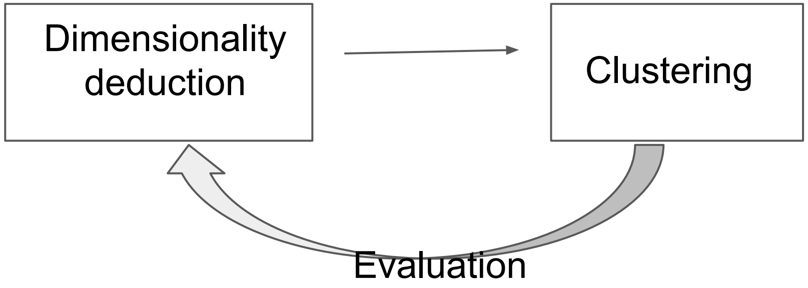
1. Choosing a transformation ⟶ fight dimensionality and concentration of distances

2. Choosing a distance on the feature space
   ⟶ Right representation for scenarios differences
   ⟶ Clustering efficiency

   For the clustering to make sense, a non-trivial tradeoff must be found between distance information, dimensionality reduction and clustering efficiency

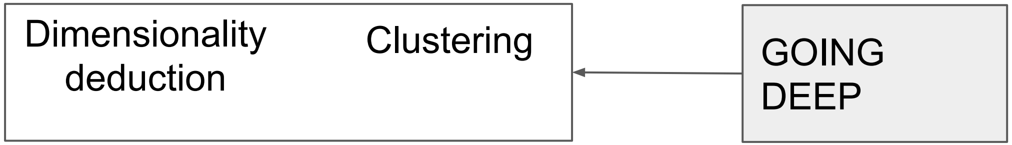
Towards Deep Clustering

Sequentially



vs

Optimize the two task jointly



Find representations “clustering friendly”

Meaningfulness and Interpretability?

Highlights

1. Choice of the distance/transformation

2. Time vs space

3. Definition of a clustering index

Distances/transformations

Dimension reduction

Fourier
Wavelets
PCA
kernel PCA

Embeddings (autoencoder)

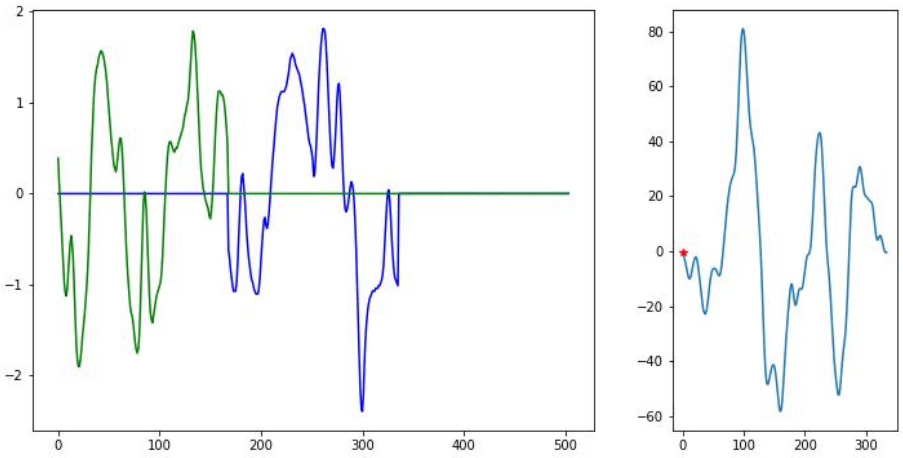
Distance

L2

MLCC

DTW

Max Lagged Cross Correlation

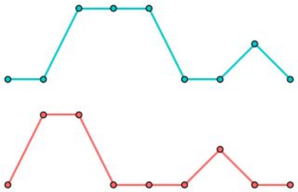


The Max Lagged Cross Correlation (Max CC) distance looks for an optimal alignment between two signals, with the two series only being allowed to be aligned via shifts in the time axis.

[Paparrizos and Gravano 2016]

Max CC takes into account the dynamic nature of the data, the fact that we are dealing with time series, which by their very definition can have **lagged** relations.

Dynamic Time Warping

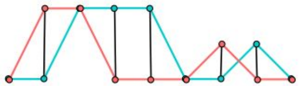


Series 1

Series 2

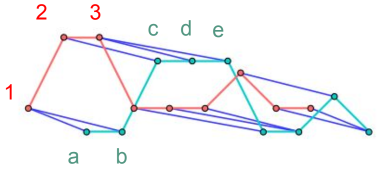
Dynamic Time Warping (DTW) looks for the optimal alignment between two time series. Like the Max Lagged CC Distance, DTW takes into account the dynamic nature of the data. Unlike Max Lagged CC Distance, DTW allows for non-linear alignments of the series and is more computationally demanding.

Euclidean



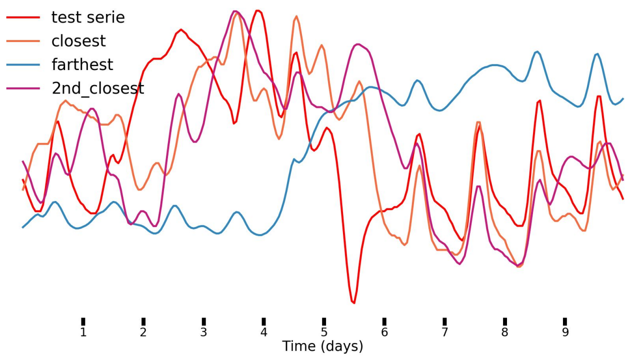
[Berndt and Clifford 1994]

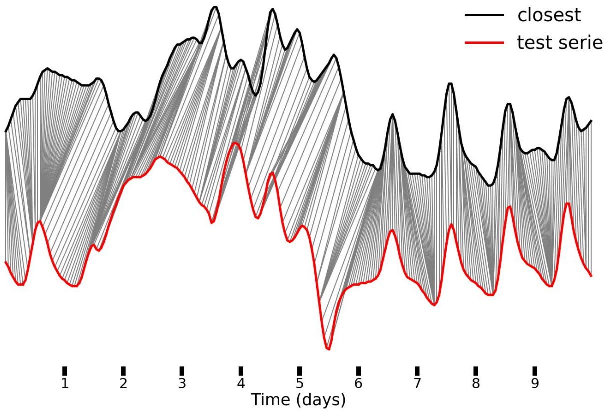
DTW



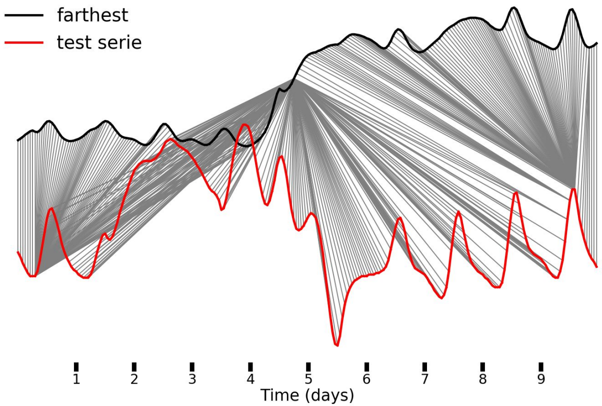
Interactive visualization:

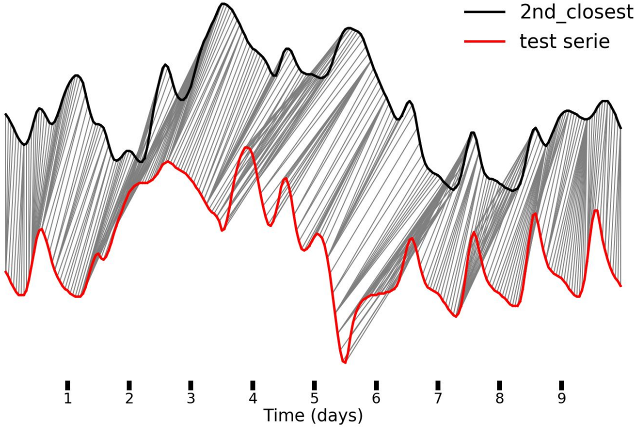
https://plot.ly/~aumfurer/2/closest/#/





https://plot.ly/~aumfurer/4/farthest/#/





Examples of clustering pipeline

Example 1:

Transformation: wavelet basis + thresholding
Distance in feature space: L2 between the selected coefficients
Cost of dimension reduction: cost of reconstruction
Clustering: k-medoids

Example 2:

Transformation: none
Distance: DTW
Cost of dimension reduction: none
Clustering: k-medoids

Meaning of the distances

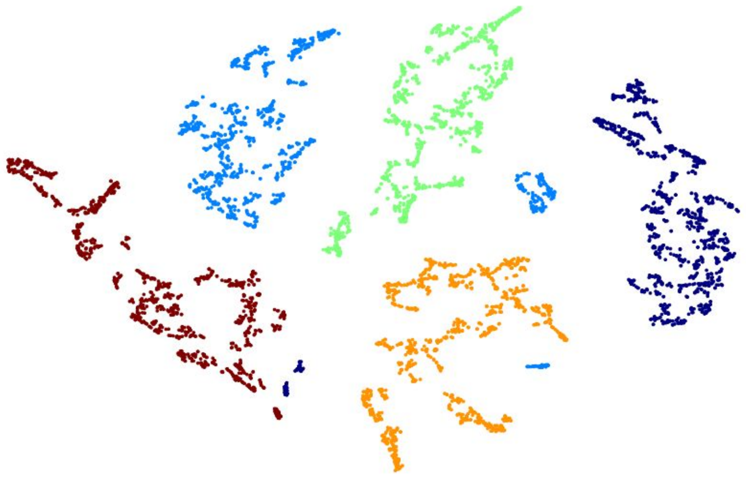
Depending on the distance (filtering / dimension reduction) chosen, the **variability** of the data might be dominated either:

- **by spatial characteristics (close points in space look alike)**

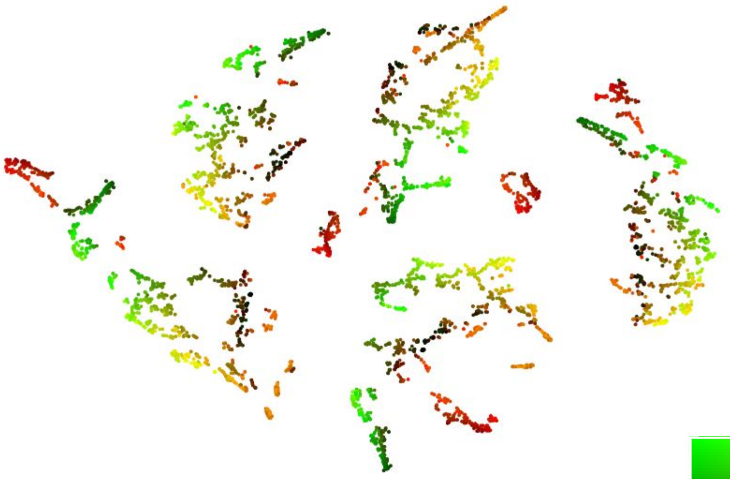
- **by time characteristics (series of the same year look alike)**

Autoencoders

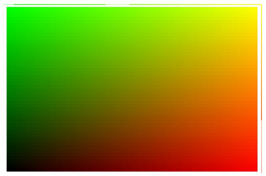
Embeddings of series of 100 geographical points for 5 scenarios



TSNE projection (colored by years)



TSNE projection (colored by geographics)



color code

Variability among years is stronger than among geographical points.

Experiments

Data

Select the temperature series of a geographical point and a year (reference series) and consider:

(a) The temperature series corresponding **to the same geographical point in all the others scenarios** (198 series)

(b) The temperature series corresponding **to the same year but other geographical points** (we set the number of geo points to be the same as scenarios: 198, although the number of geo points is around 3500)

Goal

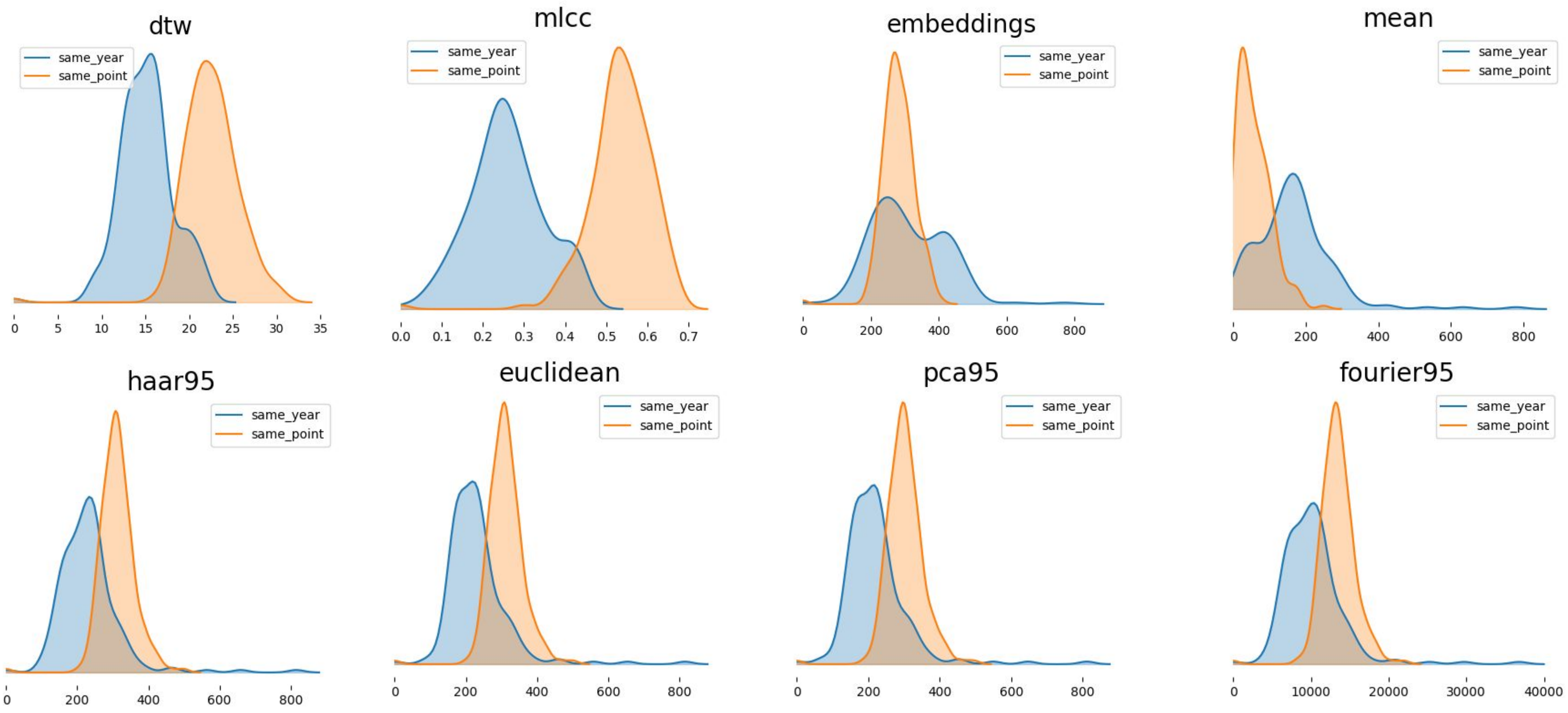
We want to compare:

the distribution of distances between the reference series and the series in group (a)

versus

the distribution of distances between the reference series and the series in group (b)

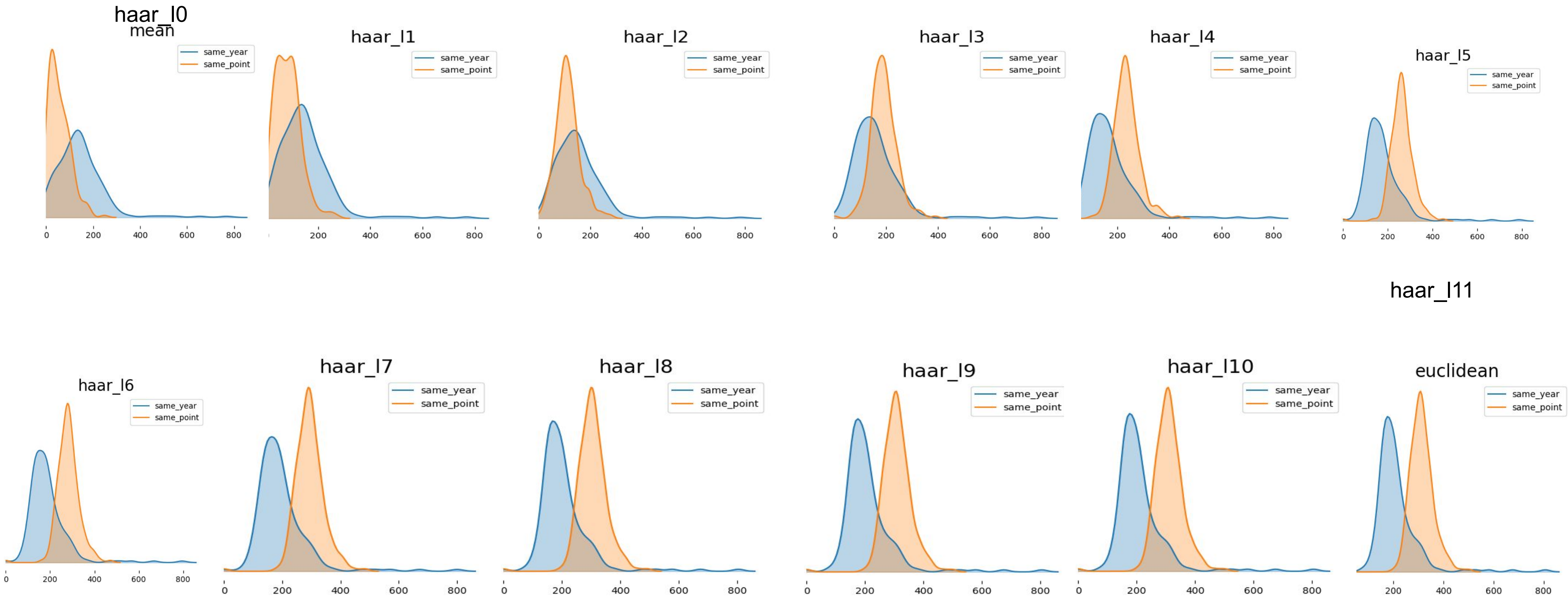
Results



Distances from reference series to group (b): **Same Year, faraway points**

Distances from reference series to group (a): **One Point, different years**

Transition from mean to euclidean by Haar decomposition (series without z-normalization)
(Point near Alpes- year 2127)



Phase transition for distances

haar
fourier

mean

Euclidean

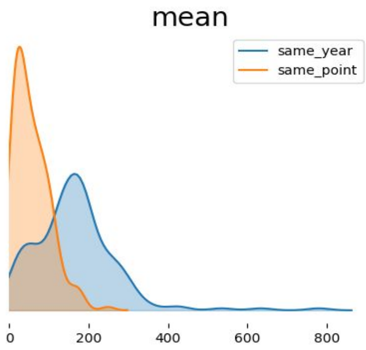
DTW

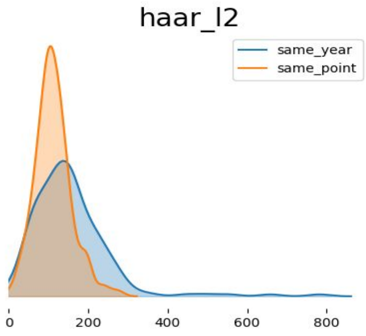
z-scored
DTW

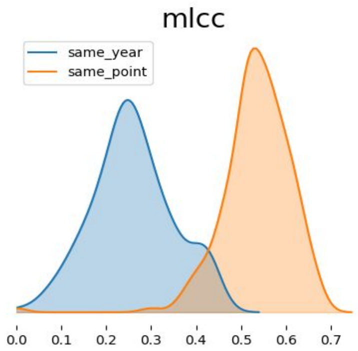
MLCC

SPACE

TIME







How to find the dimensional tradeoff ?

Suppose we can choose among a certain family of well chosen distances.

Given a subset of scenarios (for instance winters of given point), how to optimize the clustering efficiency without loosing meaningfulness?

Clustering index

We propose a clustering index that helps to select a metric/dimension reduction

For d a given “metric”:

Index(d) = F(d) * Q(d)

F stands for fidelity and Q for cluster quality

Q(d): clustering quality index

We decide to use one minus within index as a measure of the quality of the clusters. That is :

Q(d) = 1 - within index (d)

Notice that 0<= within index <=1 and therefore 0<=Q(d) <=1. The best within index is 0 and so the best Q(d) is 1.

Within Index: Ratio between the average distance from each point to its center and the average distance between points.

$$\text{Within index (d)} = \frac{\sum_i d(x_i, c_{x_i})^2 / n}{\sum_{i<j} d(x_i, x_j)^2 / (n \times (n+1)/2)}$$

where $\{x_i\}_i$ are the data points, n is the number of data points and c_{x_i} is the center of the cluster associated to x_i

! This index benefits k-means over k-medoids.

F(d): fidelity index

How to construct a measure of deformation from the original space with distance D to the feature space with distance d?

We want an index that gives 1 to 0 deformation and 0 to huge deformations.
(We do not have necessarily a notion of reconstruction or decoder)

F(d): fidelity index

In order to be able to evaluate the fidelity of the representation of the data, a reference metric D is set. We want to evaluate how well the distances d preserve D.

We use the T-SNE Stochastic embedding for the original space and feature space:

$$p_{j|i} = \frac{\exp(-D(x_i, x_j)^2/2\sigma_i^2)}{\sum_{k \neq l} \exp(-D(x_i, h_k)^2/2\sigma_i^2)}$$

$$q_{ij} = \frac{(1 + d(y_i, y_j)^2)^{-1}}{\sum_{k \neq l}(1 + d(y_k, y_l)^2)^{-1}}$$

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n}$$

We define the distance between our evaluation distribution P and our reference distribution Q, as the symmetrized Kullback-Leibler (called Jensen–Shannon divergence):

$D_{JS}(P, Q) = D_{KL}(P||Q)/2 + D_{KL}(Q||P)/2$

Since this metric gives a value between 0 and infinite, we will use a logistic function to limit it to the interval [0, 1)

then our Fidelity function F(d) is defined as:

$$F(d) = \frac{1}{1 + e^{-(D_{JS} - \beta)}}$$

where

beta = 2 x sigma, where sigma is the standard deviation of the D_{JS} values for all the considered models {d}.

About RTE Data

We considered the times series corresponding to 199 winter scenarios for one geographical point (next to Paris location: lat long 48.8°, 2.3°).

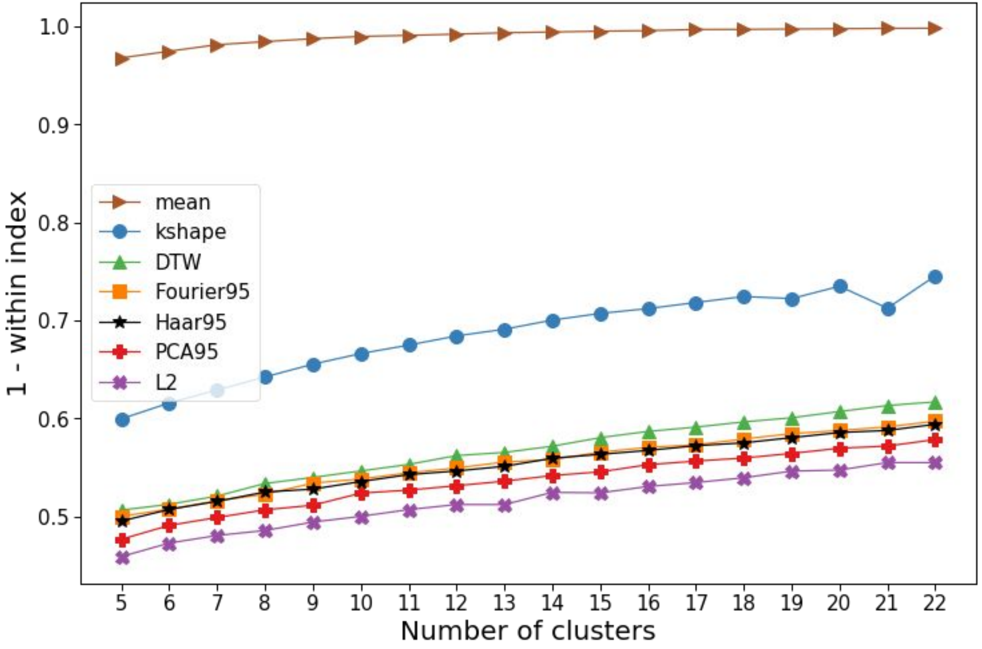
For each of the 199 times series we considered 2048 winter hours (aprox 85 winter days). The length of the series was chosen to be a power of 2, which simplifies the wavelets decomposition.

Models

Representation of the data	**Distance**	**Number of clusters**	**Clustering method**	**Name**
Plain time series	euclidean	range from 5 to 22	k-medoids	L2
PCA 95%	euclidean	range from 5 to 22	k-medoids	PCA95
Fourier 95%	euclidean	range from 5 to 22	k-medoids	Fourier95
Haar 95%	euclidean	range from 5 to 22	k-medoids	Haar95
Plain time series	euclidean	range from 5 to 22	k-medoids	mean
z-normalized time series	DTW	range from 5 to 22	k-medoids	DTW
z-normalized time series	max lag cc	range from 5 to 22	k-means	MLCC

Results on RTE

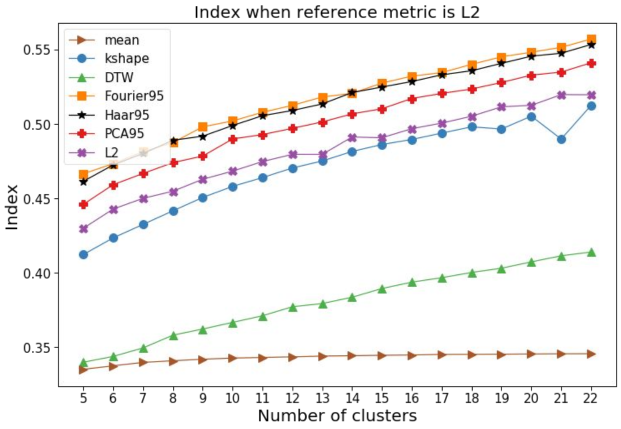
Q(d) = 1 - within index (d)
(best = 1)

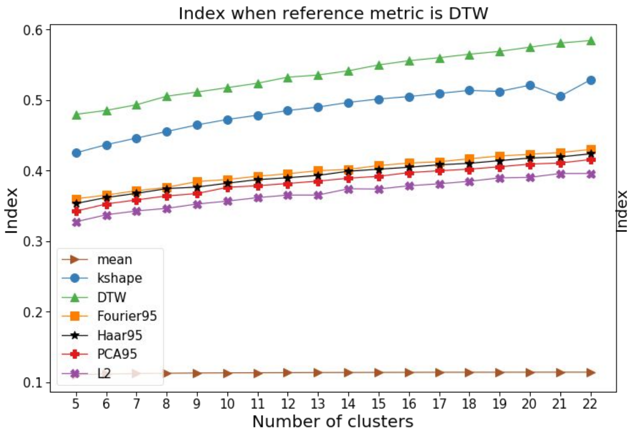


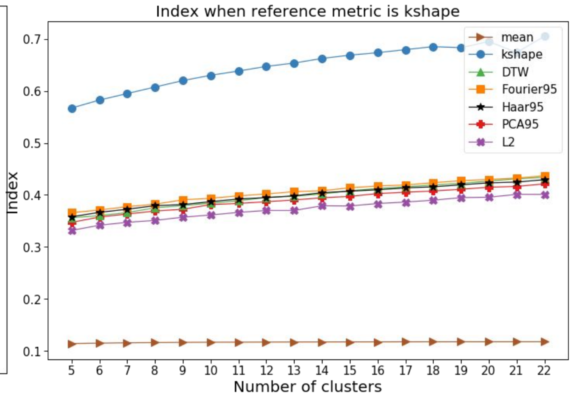
Each point in the plot is the mean value of 25 independent runs of the clustering algorithm (the same with the following plots).

Results on RTE

Index = F(d) * Q(d)







L2 as reference metric

Best: Fourier 0.95
2best: Haar 0.95

DTW as reference metric

Best: DTW
2best: MLCC
3 best: fourier 0.95

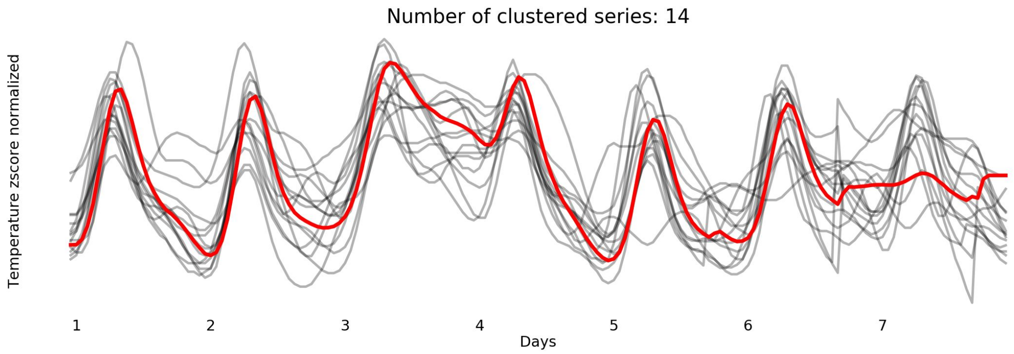
MLCC as reference metric

Best: MLCC
2best: Fourier 0.95

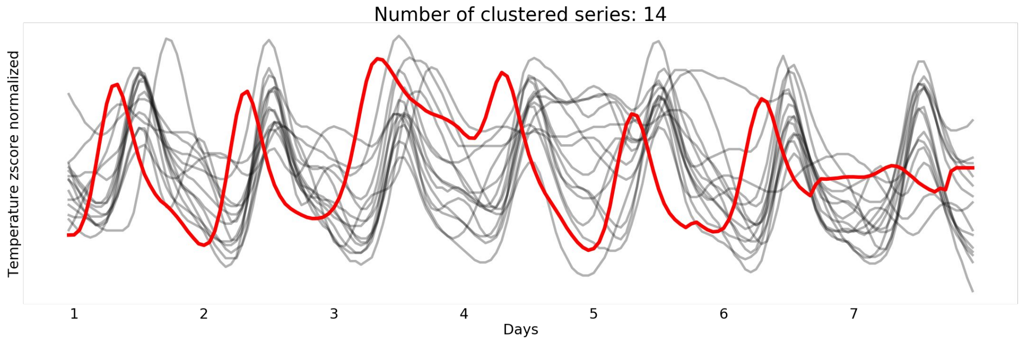
Remember that each point is the mean of 25 independent runs

MLCC results - 1 week

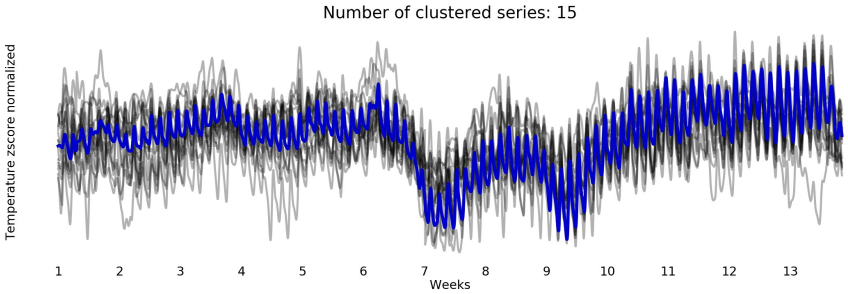
Lagged series

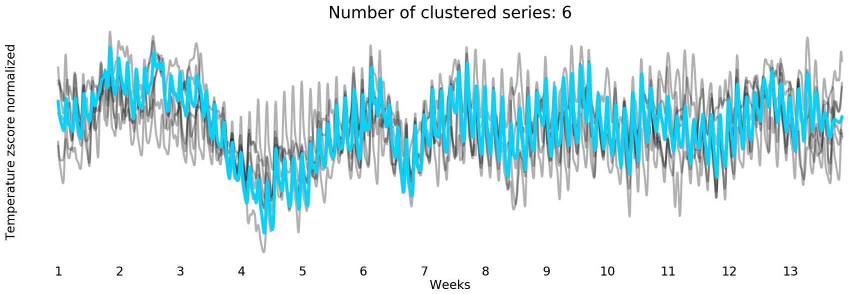


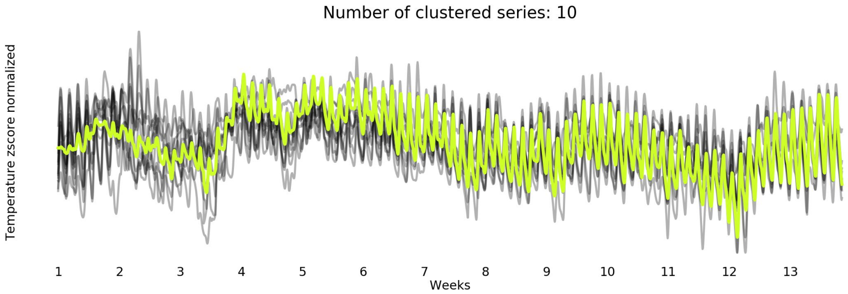
Original series

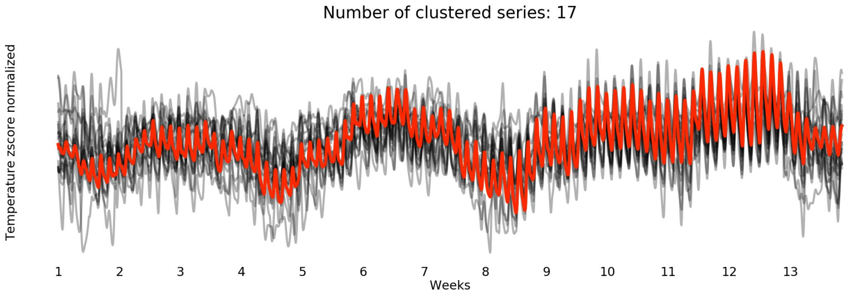


MLCC results









Experiments on UCR data set

The UCR series datasets is a set of labeled temporal series datasets.

In order to evaluate our metric we will compare the distance chosen by our index with the one selected using the accuracy.

Statistics over UCR

Metric with best accuracy	Number of data sets	Reference metric for our index	proportion of datasets where our index selects the metric with best accuracy
MLCC	27	**MLCC**	**23/27 = 0.852**
MLCC	27	DTW	12/27 = 0.444
MLCC	27	l2	17/27 = 0.63
DTW	38	MLCC	3/38 = 0.078
DTW	38	**DTW**	**15/38 = 0.395**
DTW	38	l2	3/38 = 0.079
other (haar95, fourier95, PCA95, l2)	34	MLCC	1/34 = 0.029
other (haar95, fourier95, PCA95, l2)	34	DTW	2/34 = 0.059
other (haar95, fourier95, PCA95, l2)	34	l2	3/34 = 0.088

Notion of quantile

We aim at a definition of quantiles for the whole serie (not for marginals).

Several definitions in the literature:
For instance:

Daniel Peña, Ruey S. Tsay & Ruben Zamar (2019): Empirical
Dynamic Quantiles for Visualization of High-Dimensional Time Series, Technometrics, 2019.

Chernozhukov, V., Galichon, A., Hallin, M., & Henry, M. (2017). Monge–
Kantorovich depth, quantiles, ranks and signs. The Annals of Statistics , 45(1),
223-256.

Gouriéroux, C., & Jasiak, J. (2008). Dynamic quantile models. Journal of
Econometrics , 147(1), 198-205.Hallin, M., Paindaveine, D., Šiman, M., Wei, Y., Serfling, R., Zuo, Y., ... &
Mizera,

I. (2010). Multivariate Quantiles and Multiple-Output Regression Quantiles: From
L1 Optimization to Halfspace Depth.[with Discussion and Rejoinder]. The Annals
of Statistics , 635-703.

Our proposal (work in progress)

Can we define a notion of quantile (or tube around a serie) using wavelet coefficients?

Practical algorithm (for a fixed quantile a):

1. Compute the wavelet coefficients (in your favorite base) of all series
2. Define the estimators for the mean coefficients beta and variance (obvious way)
3. Threshold the coefficients
   First keep 95% variability,
   then adjust to be at “small enough” distance to one of the data curves depending on the variance
4. Using a constant L(a,n,thresholding), define the tube around the function:

$$\sum_{k,i \leq J_0} \hat{\beta}_{i,k} \psi_{i,k}$$

Some theoretical backup

In the case of a (simplistic) model:

$$Y_i(t) = f(t) + \epsilon_i,$$

with f with some regularity, and the noise Gaussian i.i.d.

Then we get results of the type:

Theorem

+ ***Regularity of noise***
+ ***Regularity of f (in functional space)***
+ ***Right scaling of tube and thresholding*** implies

$$P\left(f \not\in \mathrm{Tube}_\alpha\right) \leq \alpha.$$

THANKS