Time Series Clustering

Andrés M. Alonso

[1]Department of Statistics, UC3M

[2]Institute Flores de Lemus

ASDM - C02
June 25 – 29, 2018, Boadilla del Monte

Outline

What is the meaning of clustering?

Definition

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters).

Wikipedia

Key elements of the definition

- Objects
- Group (that can be hard or soft).
- Similarity.

Algorithms for clustering

- Connectivity-based clustering (hierarchical clustering)
  - Scotto, M., Alonso, A.M. and Barbosa, S. (2010) Clustering time series of sea levels: an extreme value approach, *Journal of Waterway, Port, Coastal, and Ocean Engineering*, 136, 215–225.
- Centroid-based clustering
  - Maharaj, E.A., Alonso, A.M. and D'Urso, P. (2015) Clustering Seasonal Time Series Using Extreme Value Analysis: An Application to Spanish Temperature Time Series, *Communications in Statistics - Case Studies and Data Analysis*, 1, 175–191.
- (Model) Distribution-based clustering
  - Alonso, A.M., Berrendero, J.R., Hernández, A. and Justel, A. (2006) Time series clustering based on forecast densities, *Computational Statistics and Data Analysis*, 51, 762–766.
- Density-based clustering

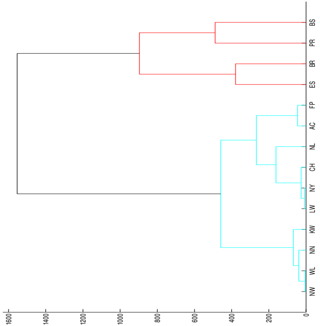
Examples of clustering algorithms

Connectivity-based clustering

These algorithms connect “objects” to form "clusters" based on their distance/similarity.

A cluster can be described by the maximum distance needed to connect parts of the cluster.

At different distances, different clusters will form, which can be represented using a dendrogram.

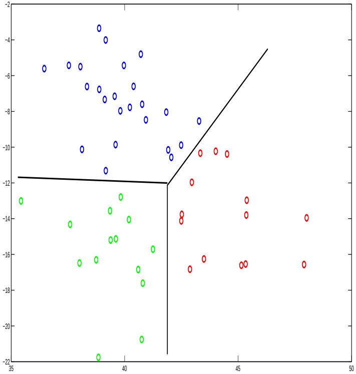


Examples of clustering algorithms

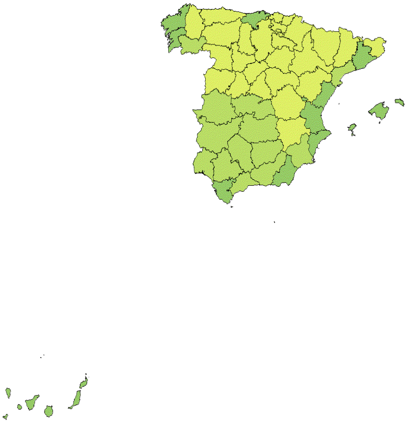
Centroid-based clustering

Clusters are represented by a central "object", which may not necessarily be a member of the data set.

- k-means
- k-mediods or PAM



Examples of clustering algorithms



Centroid-based clustering

Clusters are represented by a central “object”, which may not necessarily be a member of the data set.

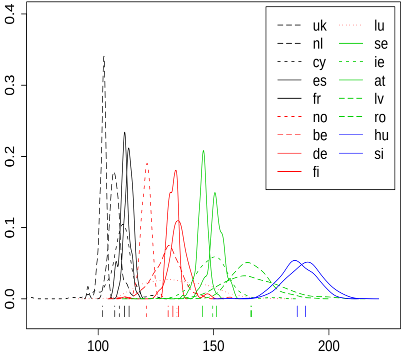
- k-means
- k-mediods or PAM

Examples of clustering algorithms

(Model) Distribution-based clustering

The clustering model most closely related to statistics is based on distribution models.

Clusters can then easily be defined as objects belonging most likely to the same distribution/model.



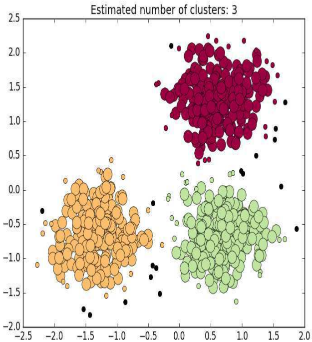
(b)

Examples of clustering algorithms

Density-based clustering

Clusters are defined as areas of higher density than the remainder of the data set.

Objects in sparse areas are usually considered to be noise and border points.



The problem

Time series clustering problems arise when we observe a sample of time series and we want to group them into different categories or clusters.

This a central problem in many application fields and hence time series clustering is nowadays an active research area in different disciplines including finance and economics, medicine, engineering, seismology and meteorology, among others.

Approaches for time series clustering

- Time series clustering by features.
- Model based time series clustering.
- Time series clustering by dependence.

Liao, T.W. (2005) Clustering of time series data-a survey, *Pattern Recognition*, 38, 1857–1874.

Aghabozorgi, S., Shirkhorshidi, A.S. and Wah, T.Y. (2015) Time-series clustering – A decade review. *Information Systems* 53 16–38.

Approaches for time series clustering

- Time series clustering by features.

Kakizawa, Y., Shumway, R.H. and Taniguchi, M. (1998) Discrimination and clustering for multivariate time series, *J. Am. Stat. Assoc.*, 93, 328–340.

Vilar, J.A. and Pértega, S. (2004) Discriminant and cluster analysis for Gaussian stationary processes: Local linear fitting approach, *J. Nonparametr. Stat.*, 16, 443–462.

Caiado, J., Crato, N. and Peña, D. (2006) A periodogram-based metric for time series classification, Comput. Statist. Data Anal. 50, 2668-2684.

Scotto, M., Alonso, A.M. and Barbosa, S. (2010) Clustering time series of sea levels: an extreme value approach, *Journal of Waterway, Port, Coastal, and Ocean Engineering*, 136, 215–225.

D'Urso, P., Maharaj, E.A. and Alonso, A.M. (2017) Fuzzy Clustering of Time Series using Extremes, *Fuzzy Sets and Systems*, 318, 56–79.

Approaches for time series clustering

- Model based time series clustering.
  - Alonso, A.M., Berrendero, J.R., Hernández, A. and Justel, A. (2006) Time series clustering based on forecast densities, *Computational Statistics and Data Analysis*, 51, 762–766.
  - Corduas, M., Piccolo, D. (2008) Time series clustering and classification by the autoregressive metric, *Comput. Statist. Data Anal.*, 52, 1860–1872.
  - Scotto, M.; Barbosa, S. and Alonso, A.M. (2009) Model-based clustering of Baltic sea-level, *Applied Ocean Research*, 31, 4–11.
  - Vilar-Fernández, J.A., Alonso, A.M. and Vilar-Fernández, J.M. (2010) Nonlinear time series clustering based on nonparametric forecast densities, *Computational Statistics and Data Analysis*, 54, 2850–2865.
- Time series clustering by dependence.
  - Alonso, A.M. and Peña, S. (2017) Clustering time series by dependency. *Preprint*.

Packages for time series clustering

- `TSclust`: Package for Time Series Clustering.
  - Montero, P and Vilar, J.A. (2014) TSclust: An R Package for Time Series Clustering. Journal of Statistical Software, 62(1), 1-43.
- `dtwclust`: Time Series Clustering Along with Optimizations for the Dynamic Time Warping (DTW) Distance.
  - https:github.comasardaesdtwclust

Time series clustering by features

We have a set of univariate time series, $\boldsymbol{X} = \{\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n\}$, where $\boldsymbol{X}_i = (X_{i,1}, X_{i,2}, \ldots, X_{i,T})$ and we want to cluster them.

Starting point

To choose a metric to assess the dissimilarity between two time series.

This metric plays a crucial role in time series clustering.

Time series clustering by features

Conceptually most of the dissimilarity criteria proposed for time series clustering lead to a notion of similarity relying on:

- Proximity between raw series data.
- Proximity between features of the time series.
- Proximity between underlying generating processes.

Raw series data can be considered as naïve features of the time series.

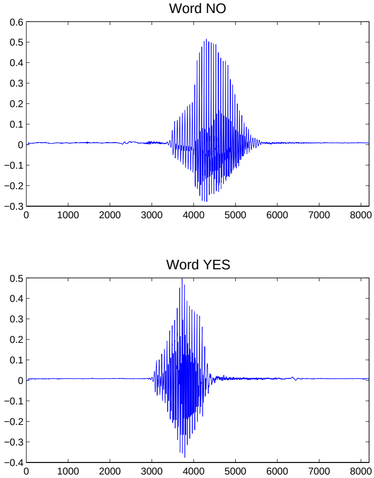
Raw data clustering

It consists on measure the distance, D, between two time series using an element-wise approach:

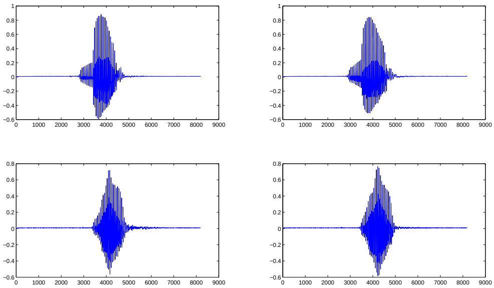
$$D(\boldsymbol{X}_i, \boldsymbol{X}_j) = d(\boldsymbol{X}_i - \boldsymbol{X}_j),$$

where d is a distance on $\mathbb{R}^T$.

This approach has a drawback since it requires that the series to be aligned.



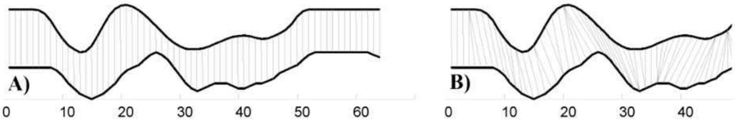
Raw data clustering



Euclidean distance matrix:

$$\begin{pmatrix} 0 & 14.1777 & 12.3613 & 12.7610 \\ 14.1777 & 0 & 10.5822 & 11.3088 \\ 12.3613 & 10.5822 & 0 & 8.0949 \\ 12.7610 & 11.3088 & 8.0949 & 0 \end{pmatrix}$$

Raw data clustering



Dynamic time warping distance matrix:

$$\begin{pmatrix} 0 & 43.4941 & 70.2141 & 70.1087 \\ 43.4941 & 0 & 75.1402 & 78.3575 \\ 70.2141 & 75.1402 & 0 & 36.7705 \\ 70.1087 & 78.3575 & 36.7705 & 0 \end{pmatrix}$$

Datafile `<yesnot.xls>`

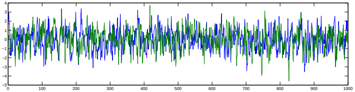
Time series clustering by features

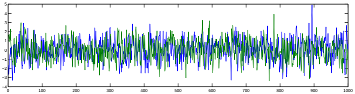
Raw data clustering it is an interesting approach when we expect the differences in the level of the series.

Euclidean distance matrix:

$$\begin{pmatrix} 0 & 51.206 & 48.735 & 51.184 \\ 51.206 & 0 & 51.472 & 50.709 \\ 48.735 & 51.472 & 0 & 51.669 \\ 51.184 & 50.709 & 51.669 & 0 \end{pmatrix}$$

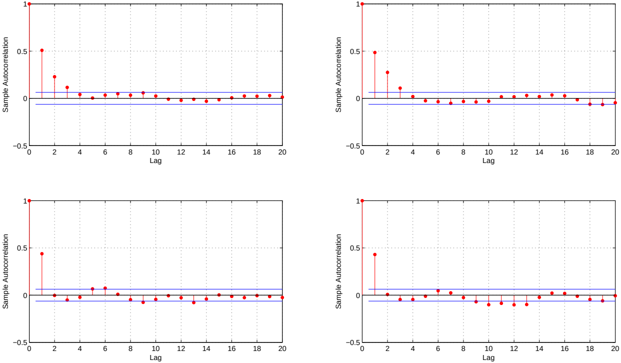
Two AR(1) and two MA(1) time series:





Autocorrelation clustering

But, in this case, autocorrelation functions are a “good” clustering criteria:



Autocorrelation clustering

Assume that we have two stationary series, $\boldsymbol{X}$ and $\boldsymbol{Y}$:

$$\begin{cases} H_0: & \boldsymbol{\rho}_X = (\rho_{X,1}, \rho_{X,2}, \dots, \rho_{X,m})' = \boldsymbol{\rho}_Y = (\rho_{Y,1}, \rho_{Y,2}, \dots, \rho_{Y,m})' \\ H_1: & \boldsymbol{\rho}_X = (\rho_{X,1}, \rho_{X,2}, \dots, \rho_{X,m})' \neq \boldsymbol{\rho}_Y = (\rho_{Y,1}, \rho_{Y,2}, \dots, \rho_{Y,m})' \end{cases}$$

where $\rho_{X,k}$ and $\rho_{Y,k}$ are the corresponding autocorrelations.

We can use the following test statistics:

$$T_{n,m} = n \sum_{k=1}^{m} (r_{X,k} - r_{Y,k})^2,$$

where $r_{X,k}$ and $r_{Y,k}$ are the estimated autocorrelations.

Autocorrelation clustering

- $T_{n,m}$ can be used as a distance measure.
- It is valid/correct when the series are independent.
- But its distribution changes if the series $\boldsymbol{X}$ and $\boldsymbol{Y}$ are cross-dependent.

So, we need a procedure to derive the distribution of $T_{n,m}$ in order to be able to evaluate if a given value $t_{n,m}$ is significantly different from zero.

Autocorrelation clustering

Subsampling algorithm to obtain the distribution of $T_{n,m}$:

1. Let $\boldsymbol{X}_j = (X_j, X_{j+1}, \ldots, X_{j+l-1})$ and $\boldsymbol{Y}_j = (Y_j, Y_{j+1}, \ldots, Y_{j+l-1})$ with $j = 1, 2, \ldots, n-l+1$ be the subsamples of l consecutive observations from $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively. We calculate the j-th subsampling statistic, $T_{l,m}^{(j)}$, by:

$$T_{l,m}^{(j)} = l \sum_{k=1}^{m} (\widehat{\rho}_{X_j,k} - r_{\widehat{\rho}_j,k})^2,$$

   where $\widehat{\rho}_{X_j,k}$ and $\widehat{\rho}_{Y_j,k}$ are the k-th estimated autocorrelations using the subsamples $\boldsymbol{X}_j$ and $\boldsymbol{Y}_j$, respectively.

2. We calculate $g_{n,l}(1-\alpha)$ the $1-\alpha$ quantile of $\widehat{G}_{n,l}(\cdot)$.
3. We reject H_0 if and only if $T_{n,m} > g_{n,l}(1-\alpha)$.

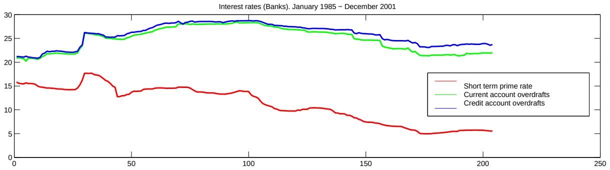
Autocorrelation clustering - Example

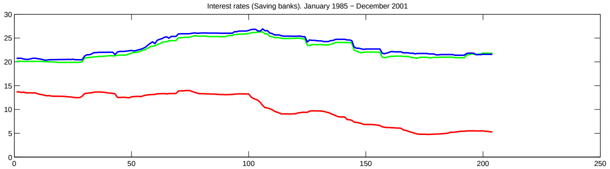
Code and description of interest rate series:

Code	Description
BME08040203001	Reference rates/Banks/Short term prime rate
BME08040203002	Banks lending rates/Current account overdrafts/Effective rate
BME08040203003	Banks lending rates/Exceed in credit card/Effective rate
BME08040203005	Reference rates/Saving banks/Short term prime rate
BME08040203006	Savings banks lending rates/Current account overdrafts/Effective rate
BME08040203007	Savings banks lending rates/Credit account overdrafts/Effective rate

Autocorrelation clustering - Example

It is clear that series are dependent.





Autocorrelation clustering - Example

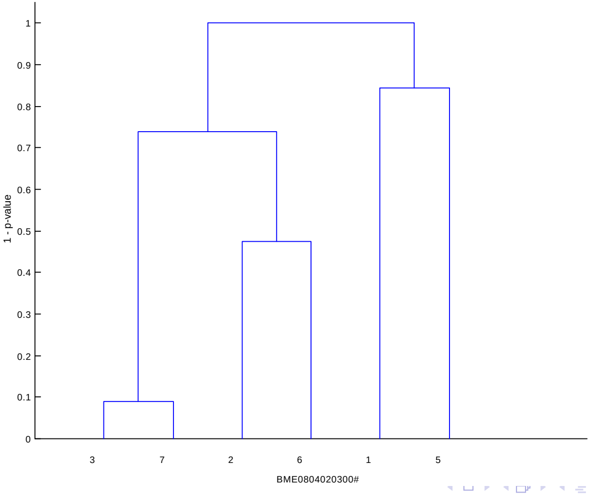
Associated p-value for each pair stationary series:

BME0804020300#	1	2	3	5	6	7
1	-	0.000	0.000	0.155	0.000	0.000
2		-	0.442	0.139	0.524	0.598
3			-	0.065	0.623	0.909
5				-	0.008	0.000
6					-	0.262
7						-

Alonso, A.M. and Maharaj, E.A. (2006) Comparison of time series using subsampling, *Computational Statistics and Data Analysis*, 50, 2589–2599.

Datafile <BME.xls>

Autocorrelation clustering - Example



Spectral domain clustering

Assume that we have two stationary series, X and Y with spectral densities

$$\lambda_X = \sum_{k=-\infty}^{\infty} \gamma_{X,k} \exp(-ik\omega)$$

and

$$\lambda_Y = \sum_{k=-\infty}^{\infty} \gamma_{Y,k} \exp(-ik\omega)$$

As before, we are interested on testing:

$$\begin{cases} H_0: & \lambda_X(\omega) = \lambda_Y(\omega) \quad (0 \leq \omega \leq \pi) \\ H_1: & \lambda_X(\omega) \neq \lambda_Y(\omega) \end{cases}.$$

Spectral domain clustering

Diggle y Fisher (1991) propose to compare the integrated periodograms:

$$F_X(\omega_j) = \sum_{i=1}^{j} I_X(\omega_i) / \sum_{i=1}^{m} I_X(\omega_i),$$

and

$$F_Y(\omega_j) = \sum_{i=1}^{j} I_Y(\omega_i) / \sum_{i=1}^{m} I_Y(\omega_i),$$

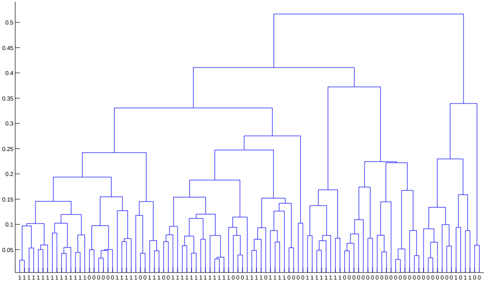
where $\omega_i = 2\pi i/n$, $I_X(\cdot)$ is the periodogram, and $m = \lceil (n-1)/2 \rceil$.

We can use the following test statistics:

$$D_m = \sup |F_X(\omega) - F_Y(\omega)| \quad \text{or} \quad W_m = \int_0^{\pi} (F_X(\omega) - F_Y(\omega))^2 \, d\bar{F}(\omega).$$

Spectral domain clustering

We retake the word classification problem (boat versus goat):



Alonso, A.M., Casado, D., Lopez-Pintado, S. and Romo, J. (2014) Robust Functional Classification for Time Series, *Journal of Classification*, 31, 325–350.

Extreme value clustering

In some applications, the main interest is the highest (lowest) level that we can observe in a time series in a given period.

- To build dykes you need to know the maximum level of the sea in the area that you want to protect.
  - Rising sea levels are of great concern to coastal communities around the world.
- To prevent the effect of temperatures in health, you need information about the highest temperature.
- In finance/insurance, the lowest values correspond to capital losses.

Extreme value clustering

The Generalized Extreme Value distribution, as the following form:

$$G(x) = \exp\left\{ - \left[1 + \xi \left(\frac{X-\mu}{\sigma} \right) \right]^{-\frac{1}{\xi}} \right\} \tag{1}$$

defined on $\{x : 1 + \xi(\frac{X-\mu}{\sigma}) > 0\}$ where $-\infty < \mu < \infty$, $\sigma > 0$, and $-\infty < \xi < \infty$,

- The three parameters μ, σ and ξ are the location, scale and shape parameters, respectively where ξ determines the three extreme value types.
- When $\xi < 0$, $\xi > 0$ or $\xi = 0$, the GEV distribution is the negative Weibull, the Fréchet or the Gumbel distribution, respectively.

Extreme value clustering

GEV distribution fitting

- The GEV log-likelihood function presents a difficulty if the number of extreme events is small.
- It is particularly severe when the method of maxima over fixed intervals is used.
- A possible solution is to consider the r-largest values over fixed intervals (Coles 2001).

Extreme value clustering

GEV distribution fitting

The number of largest values per year, r, should be chosen carefully.

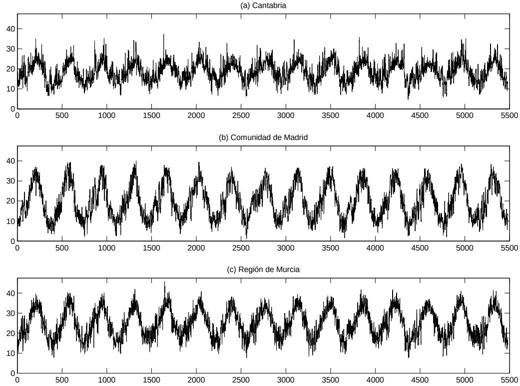
- Small values will produce likelihood estimators with high variance, whereas large values will produce biased estimates.
- In practice, r is selected as large as possible subject to adequate model diagnostics.
- The validity of the models can be checked through the application of graphical methods (Reiss and Thomas, 2000).

Extreme value clustering

- The implications of a fitted extreme value model are usually made with reference to extreme quantiles.
- By inversion of the GEV distribution function, the quantile, x_p for a specified exceedance probability p is
    - for $\xi \neq 0$, we have $x_p = \mu - \frac{\sigma}{\xi}\left[1 - (-\log(1-p)^{-\xi})\right]$.
    - for $\xi = 0$, we have $x_p = \mu - \sigma \log[-\log(1-p)]$.
- x_p is referred to the return level associate with a return period $1/p$.
- x_p is expected to be exceeded by the annual maximum in any particular year with probability p.

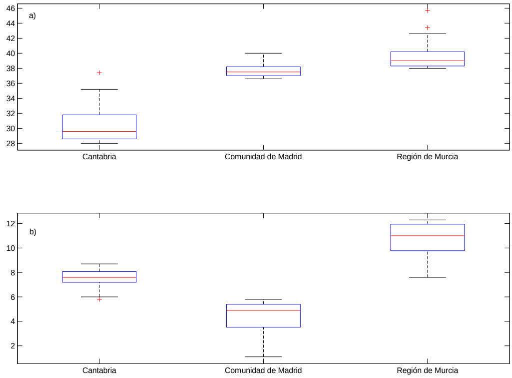
Extreme value clustering - Example

We consider 52 time series of daily maximum temperatures (in degrees Celsius, ^{o}C) observed in Spain from 1990 to 2004.



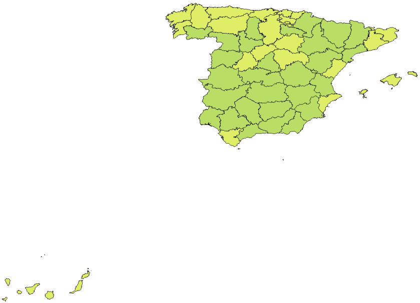
Extreme value clustering - Example

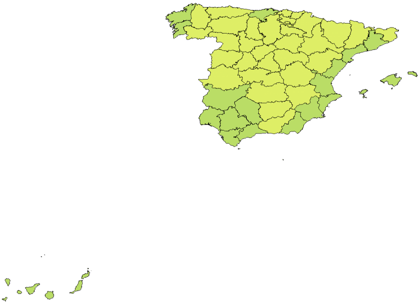
Box-plot of the exceedances (1990-2004) above (below) the 95% (5%) percentile during summer (winter) period.



Extreme value clustering - Example

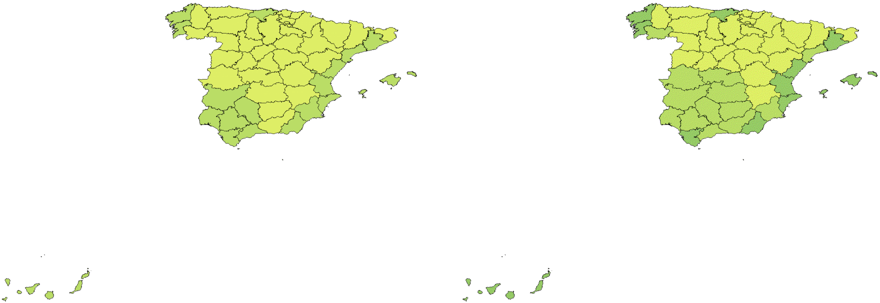
(a) Two clusters based on GEV estimates for highest temperatures
(b) Two clusters based on GEV estimates for lowest temperatures





Extreme value clustering - Example

(a) Two clusters based on two sets of GEV estimates
(b) Three clusters based on two sets of GEV estimates



Extreme value clustering - Example

Means of the 25 and 100 years returns levels with 95% confidence intervals for the three clusters based on GEV estimates:

Cluster		25 yr	95% CI		100 yr	95% CI	
1	sum	**39.12**	38.33	39.91	**39.61**	38.63	40.60
	win	**-0.63**	-1.41	0.15	**-1.31**	-2.40	-0.23
2	sum	**43.08**	42.33	43.83	**43.67**	42.68	44.66
	win	**4.87**	4.15	5.59	**4.25**	3.29	5.20
3	sum	**38.37**	37.30	39.44	**39.63**	37.75	41.51
	win	**8.76**	8.03	9.48	**8.10**	7.13	9.07

Datafile `<SpainTemperature.xls>`

GEV estimates `<SpainTemperatureEstimates.xls>`

Model based time series clustering

We need to define a distance in the space of the parameters of the models:

- Lets assume that $\{X_t\}_{t\in\mathbb{Z}}$ and $\{Y_t\}_{t\in\mathbb{Z}}$ follow an ARIMA(p,d,q) model with $\Phi_X(B)(1-B)^d X_t = \Theta_X(B)\varepsilon_{X,t}$ and $\Phi_Y(B)(1-B)^d Y_t = \Theta_Y(B)\varepsilon_{Y,t}$. Then, we can use:

$$d(X,Y) = (\Xi_X - \Xi_Y)'\boldsymbol{\Sigma}_{\Xi}^{-1}(\Xi_X - \Xi_Y),$$

where $\Xi_X = (\phi_{X,1}, \phi_{X,2}, \ldots, \phi_{X,p}, \theta_{X,1}, \theta_{X,2}, \ldots, \theta_{X,q})'$ and $\Xi_Y = (\phi_{Y,1}, \phi_{Y,2}, \ldots, \phi_{Y,p}, \theta_{Y,1}, \theta_{Y,2}, \ldots, \theta_{Y,q})'$.

Model based time series clustering

- If the ARIMA(p, d, q) model is invertible, then we can write it as AR models: $\Pi_X(B)X_t = \varepsilon_{X,t}$ and $\Pi_Y(B)Y_t = \varepsilon_{Y,t}$. Then the following distance can be used (Piccolo, 1990):

$$d(X, Y) = \left\{ \sum_{j=1}^{\infty} (\pi_{X,j} - \pi_{Y,j})^2 \right\}^{1/2}.$$

- For stationary $ARMA(p, q)$ models, we can define a similar measure using the moving average representation: $X_t = \Psi_X(B)\varepsilon_{X,t}$ and $X_t = \Psi_Y(B)\varepsilon_{Y,t}$ (Galeano y Peña, 2000):

$$d(X, Y) = \left\{ \sum_{j=1}^{\infty} (\psi_{X,j} - \psi_{Y,j})^2 \right\}^{1/2}.$$

Model based time series clustering

- For stationary and invertible $ARMA(p,q)$ models, Maharaj (1996) propose a test that can be used as a distance among models.

$$\boldsymbol{Z} = \begin{bmatrix} \boldsymbol{X} \\ \boldsymbol{Y} \end{bmatrix} = \boldsymbol{W}\boldsymbol{\pi} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{W} = \begin{bmatrix} \boldsymbol{W}_X & \mathbf{0} \\ \mathbf{0} & \boldsymbol{W}_Y \end{bmatrix}$, $\boldsymbol{W}_X$ and $\boldsymbol{W}_Y$ are $T-k \times k$ matrices of lagged observations observaciones retardadas, $\boldsymbol{\pi} = [\boldsymbol{\pi}'_X \boldsymbol{\pi}'_Y]'$, and $\boldsymbol{\varepsilon} = [\boldsymbol{\varepsilon}'_X \boldsymbol{\varepsilon}'_Y]'$.

$$\mathrm{E}[\varepsilon] = \mathbf{0},\ \mathrm{E}[\varepsilon\varepsilon'] = V = \Sigma \otimes \boldsymbol{I}_{n-k},\ \mathrm{y}\ \Sigma = \begin{bmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{yx} & \sigma_y^2 \end{bmatrix}$$

Model based time series clustering

Under $H_0 : \boldsymbol{\pi}_X = \boldsymbol{\pi}_Y$, the following statistics is distributed as χ^2_k (Maharaj, 2000):

$$D = (\boldsymbol{R}\widehat{\boldsymbol{\pi}})' \left[\boldsymbol{R}(\boldsymbol{W}\widehat{\boldsymbol{V}}\boldsymbol{W})^{-1}\boldsymbol{R}'\right]^{-1}(\boldsymbol{R}\widehat{\boldsymbol{\pi}}),$$

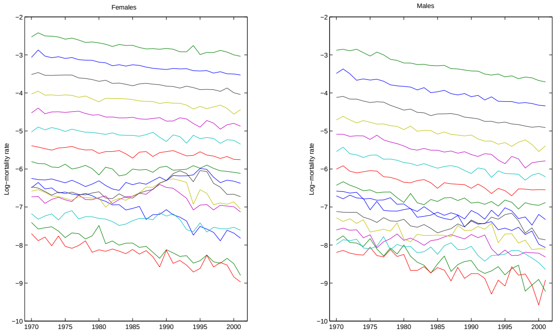
where $\widehat{\boldsymbol{V}}$ is the least squared estimator of $\boldsymbol{V}$, $\widehat{\boldsymbol{\pi}}$ is the least squared estimator of $\boldsymbol{\pi}$, and $\boldsymbol{R} = [\boldsymbol{I}_p \vdots -\boldsymbol{I}_p]$.

The statistics, D, can be used as a distance measure between $\boldsymbol{X}$ and $\boldsymbol{Y}$.

Model based time series clustering - Example

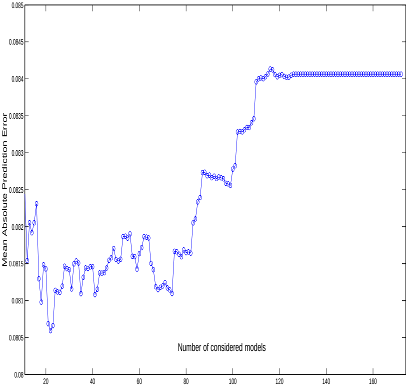
We use the Maharaj's approach for demographical data in

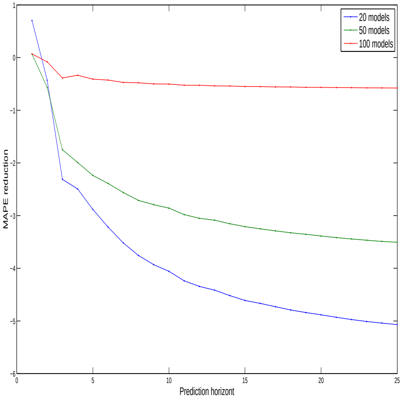
Alonso, A.M., Peña, D. and Rodríguez, J. (2013) Predicción de clusters de series temporales demográficas, MedULA, 22 (1), 25-28.



Model based time series clustering - Example

Why we cluster models?





Forecast density clustering

Most of distances or dissimilarity criteria proposed up to this point rely on the proximity between raw (features) series data, or proximity between underlying generating processes

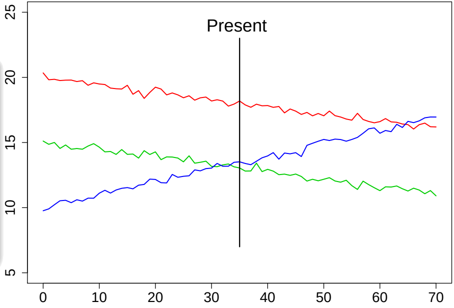
In both cases, the classification task becomes inherently static since similarity searching is governed only by the behavior of the series over their periods of observation.

In some practical situations, the real interest of clustering is the future behavior and, in particular, on how the forecasts at a specific horizon can be grouped.

Forecast density clustering

The clusters will be different if we consider:

- the models;
- the last available observation;
- the future values.



Forecast density clustering

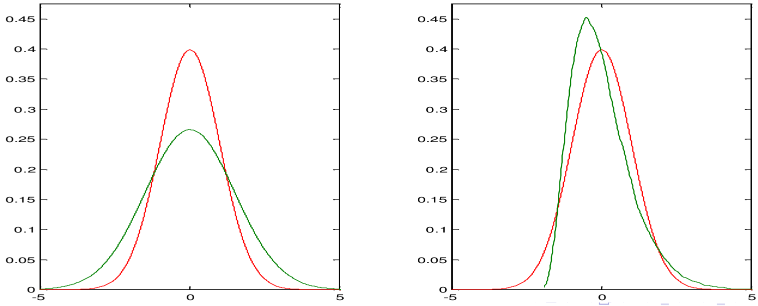
Why prediction clustering?

- It reduce the high dimensionality of the problem.
- The predictions include information about the past observations and about the data generating model.
- In some problems, the interest is on the future behaviour or if the series converge or not to some level:
  - Sustainable development.
  - (European) convergence of macroeconomic indicators.
    - Convergence of β-type (see, Barro and Sala-i-Martin, 1995).
    - Carvalho and Harvey (2005) analyze the short- and long-term convergence of the per capita income in the Euro zone.
  - Emissions of CO_2 (Kyoto Protocol).

Forecast density clustering

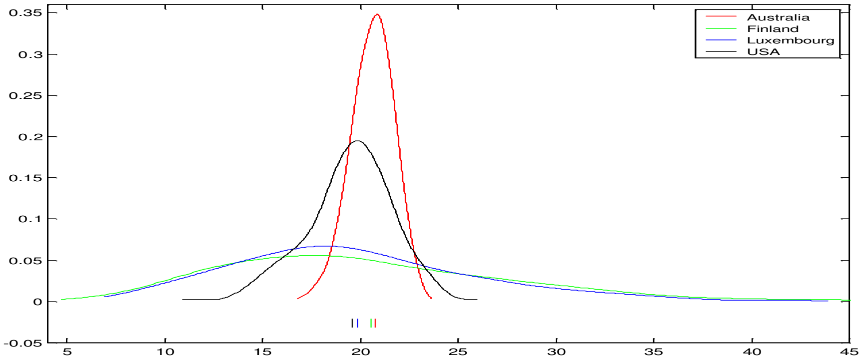
Punctual predictions or prediction densities?

- Suppose that we have series where the punctual prediction are similar (or equals).
  - Example: Prediction of financial asset returns is $E[r_t] = 0$.
- We want to distinguish among the following situations:



Forecast density clustering

Punctual predictions or prediction densities?



Real data example - Kyoto protocol

Forecast density clustering

Steps for clustering procedure

1. Prediction calculation by bootstrap.
2. Dissimilarity matrix calculation by non-parametric kernel estimators.
   - For each pair of series, $\boldsymbol{X}$ and $\boldsymbol{Y}$, we calculate the L_2 (L_1) distance among the prediction densities:

$$D_{ij} = \int \left| f_{X_{T+h}}(x) - f_{Y_{T+h}}(x) \right|^p dx,$$

     where $p = 1, 2$.
3. Finally, we use classical clustering procedures that allows distances as inputs.

Forecast density clustering

Prediction step

A general class of autoregressive processes

Let $\{X_t\}_{t\in\mathbb{Z}}$ a real valued stationary processes such that

$$X_t = m(\boldsymbol{X}_{t-1}) + \varepsilon_t,$$

where

- $\{\varepsilon_t\}$ is an i.i.d. sequence
- $\boldsymbol{X}_{t-1}$ is a d-dimensional vector of known lagged variables
- $m(\cdot)$ is assumed to be a smooth function but it is not restricted to any pre-specified parametric model.

Of course, other models can be considered.

Forecast density clustering

Prediction step

1. Estimate m using a Nadaraya-Watson estimator $\hat{m}_{g_1}$.
2. Compute the nonparametric residuals, $\hat{\varepsilon}_t = X_t - \hat{m}_{g_1}(\boldsymbol{X}_{t-1})$.
3. Construct a kernel estimate, $\hat{f}_{\tilde{\varepsilon},h}$, of the density function associated to the centered residual.
4. Draw a bootstrap-resample ε_t^* of i.i.d. data from $\hat{f}_{\tilde{\varepsilon},h}$.
5. Define the bootstrap series X_t^*, by $X_t^* = \hat{m}_{g_1}(\boldsymbol{X}_{t-1}^*) + \varepsilon_t^*$.
6. Obtain the bootstrap autoregressive function, $\hat{m}_{g_2}^*$, using the bootstrap sample $(X_1^*, \ldots, X_T^*)$.
7. Compute bootstrap prediction-paths by $X_t^* = \hat{m}_{g_2}^*(\boldsymbol{X}_{t-1}^*) + \varepsilon_t^*$, for $t = T+1, \ldots, T+H$, and $X_t^* = X_t$, for $t \leq T$.
8. Repeat Steps (1)-(7) a large number B of times.

Forecast density clustering

Dissimilarity calculation step

In practice, distances $D_{p,XY}$ are consistently approximated by replacing the unknown $f_{X_{T+b}}$ by kernel-type density estimates $\hat{f}_{X_{T+b}}$ constructed on the basis of bootstrap predictions, that is

$$\hat{D}^*_{p,XY} = \int \left| \hat{f}_{X^*_{T+b}}(x) - \hat{f}_{Y^*_{T+b}}(x) \right|^p dx, \quad i,j = 1,\ldots,s,$$

for $p = 1, 2$.

Forecast density clustering

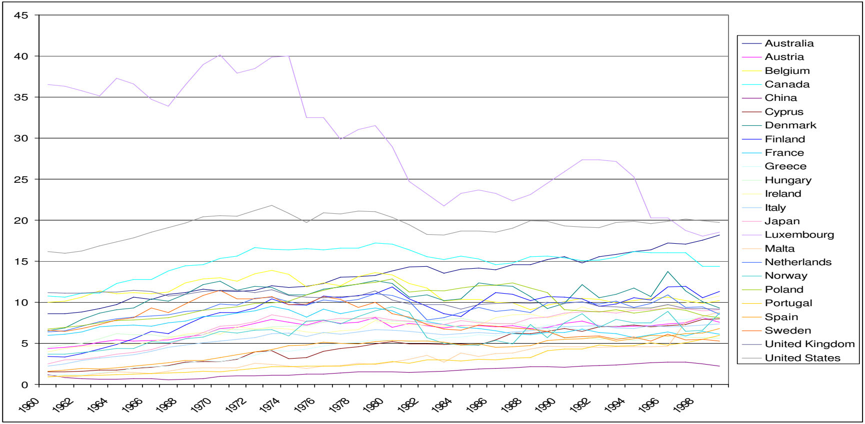
Clustering step

Application of a agglomerative hierarchical cluster algorithm

Once the pairwise dissimilarity matrix $\hat{D}_p^* = \left(\hat{D}_{p,XY}^*\right)$ is obtained, a standard agglomerative hierarchical clustering algorithm based on $\hat{D}_p^*$ is carried out.

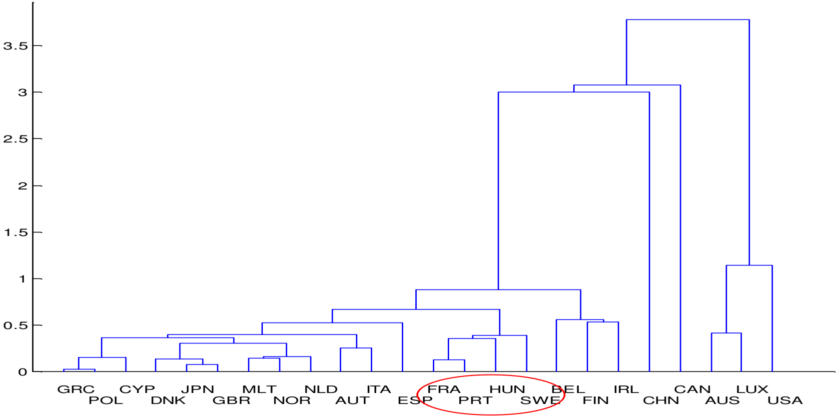
Forecast density clustering - Example

Dataset: Emissions of CO_2 in 24 industrialized countries.



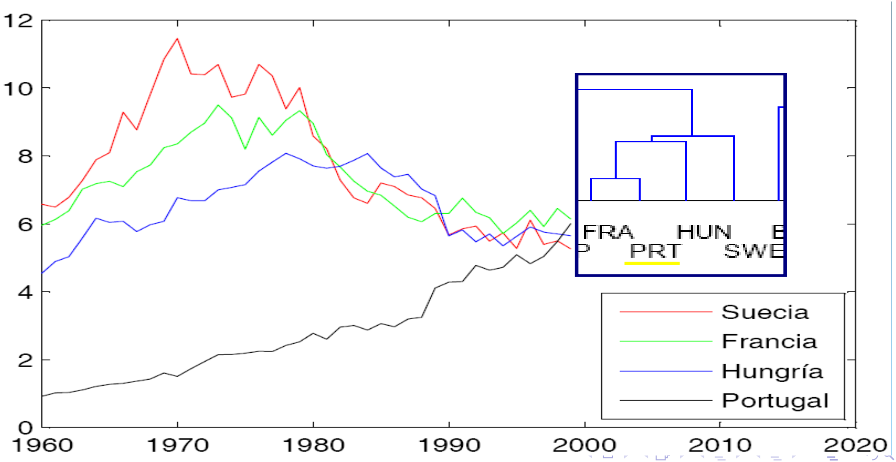
Forecast density clustering - Example

Dendrogram based on the last available observation



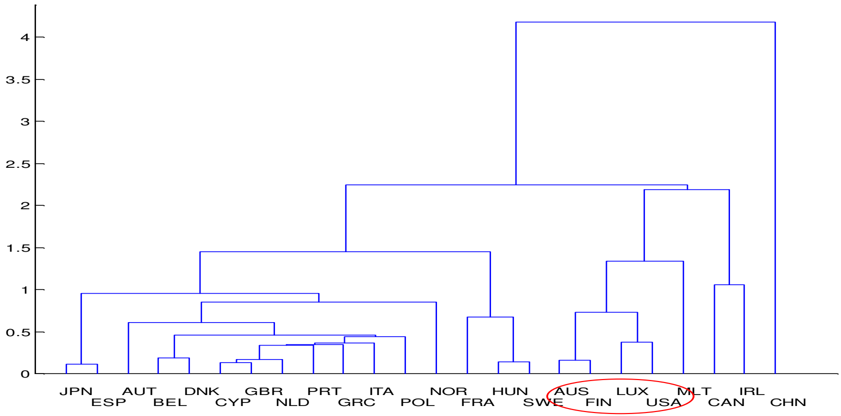
Forecast density clustering - Example

Dendrogram based on the last available observation



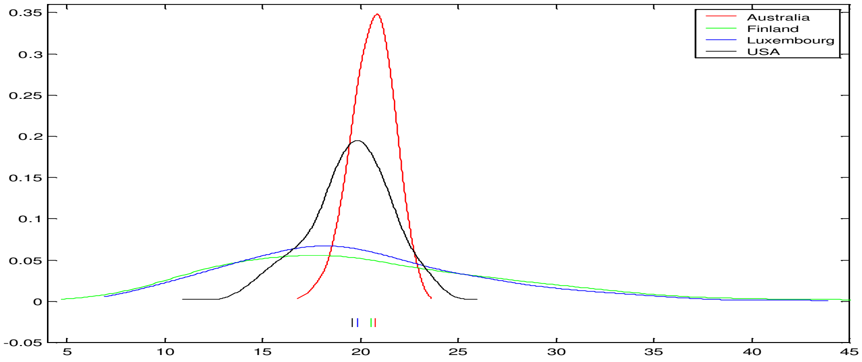
Forecast density clustering - Example

Dendrogram based on the punctual prediction for 2012



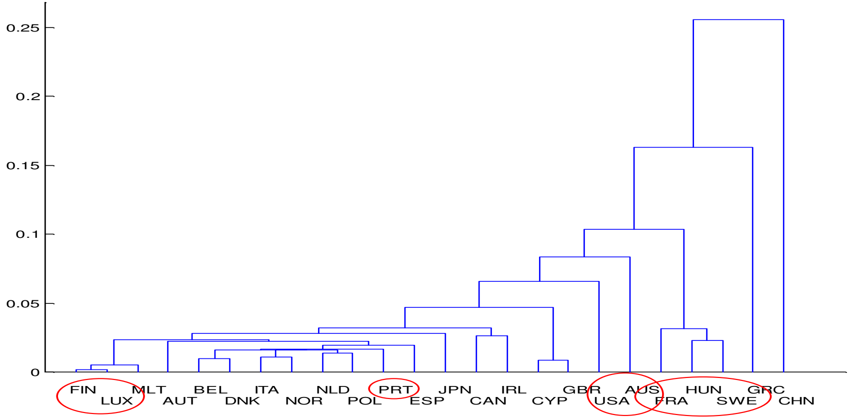
Forecast density clustering - Example

Dendrogram based on the punctual prediction for 2012



Forecast density clustering - Example

Dendrogram based on the prediction densities for 2012



Multivariate models with cluster structure

Dynamic factor models:

- When the number of series is large, VARMA models are hard to build or even unfeasible.
- Dynamic Factor Models can deal with large sets of time series.
  - Engle and Watson (1981), Peña and Box (1987), Forni et al (2000), Bai and Ng (2002), Peña and Poncela (2006), Hallin and Liska (2007), Alonso et al (2011), Lam and Yao (2012), Forni et al (2015, 2016,2017).
- For large panels of time series we often found group structure and different factors affecting to different groups.
  - Hallin and Liska (2011), Su et al (2014) and Ando and Bai (2016, 2017).

Multivariate models with cluster structure

Dynamic factor models with cluster structure:

Let $\mathbf{x}_t = (x_{1t}, \ldots, x_{mt})'$ be an m-dimensional vector time series.

$$\mathbf{x}_t = \mathbf{P}_0 \mathbf{f}_{0t} + \sum_{i=1}^{k} \mathbf{P}_i \mathbf{f}_{it} + \mathbf{n}_t,$$

where

- $\mathbf{f}_{0t} = (f_{01t}, \ldots, f_{0r_0t})'$ is a r_0-dimensional vector of common factors, $\mathbf{P}_0$ is a $m \times r_0$ factor loading matrix and k is the number of clusters.
- $\mathbf{f}_{it} = (f_{i1t}, \ldots, f_{ir_it})'$ be a r_i-dimensional vector of group-specific factors corresponding to the *ith* cluster and $\mathbf{P}_i$ is the $m \times r_i$ factor loading of these specific factors. The columns of the matrix $\mathbf{P}_i$ are of the form $(0, \ldots, 0, p_{j1}, \ldots, p_{jm_i}, 0, \ldots, 0)$, for $j = 1, \ldots, r_i$.

Multivariate models with cluster structure

- Ando, T. and Bai J. (2016) Panel data models with grouped factor structure under unknown group membership, *Journal of Applied Econometrics*, 31, 163–191.
- Ando, T. and Bai J. (2017) Clustering huge number of financial time series: A panel data approach with high-dimensional predictor and factor structures, *Journal of the American Statistical Association*, in press.

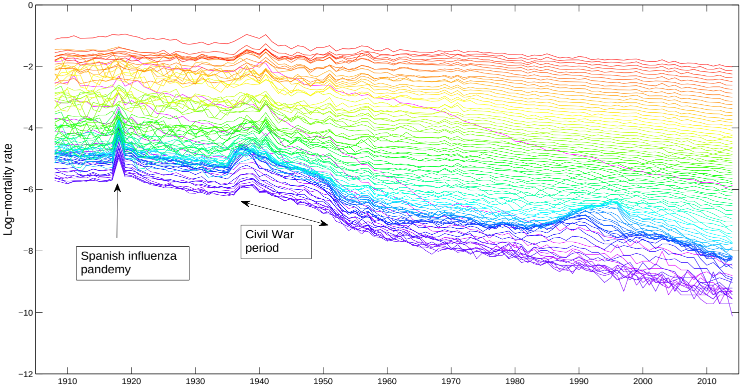
Implemented in `JAE1.R`, `JAE2.R` and `JASA.R`

Multivariate models with cluster structure

- We should to provide the number of clusters, k, the number of common factors, r, and the number of group-specific factors, r_i.
- Ando, T. and Bai J. (2017) provides a procedure for selecting, k, r and r_i but it is computationally intensive.
  - $k = 1, 2, \ldots, K$.
  - $r = 0, 1, \ldots, R$.
  - $r_i = 1, \ldots, R$.
- An information criteria is used to select those parameters.

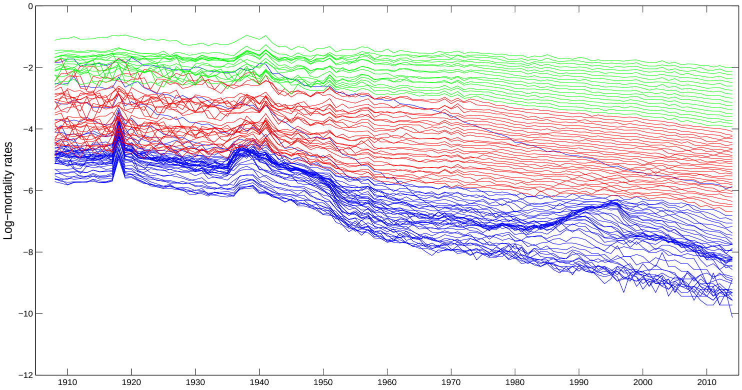
Multivariate models with cluster structure - Example

Dataset: Mortality rates by single age, Spain 1908 - 2015.



Multivariate models with cluster structure - Example

We use $k = 3$, $r = 1$ and $r_i = 1$:



Time series clustering by dependence

Up to this point, the classification task becomes inherently univariate since similarity searching is governed only by the behavior of each series but doesn't take into account the cross-dependency among the series.

Suppose that we have stationary (standardized) time series. Define $r_{xx}(h) = E(x_{it}x_{i,t-h})$ and $r_{xy}(h) = E(x_{it}y_{j,t-h})$.

We can build a measure of the dependency as follows:

- Let $\mathbf{B}(h) = \begin{bmatrix} r_{xx}(h) & r_{xy}(h) \\ r_{yx}(h) & r_{yy}(h) \end{bmatrix}$.

Time series clustering by dependence

- Then the matrix

$$\mathbf{B}_k = \begin{bmatrix} \mathbf{B}(0) & \mathbf{B}(1) & \cdots & \mathbf{B}(k) \\ \mathbf{B}(-1) & \mathbf{B}(0) & \cdots & \mathbf{B}(k-1) \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{B}(-k) & \mathbf{B}(-k+1) & \cdots & \mathbf{B}(0) \end{bmatrix}$$

is the covariance matrix of the vector stationary process $\mathbf{Z}_t = (x_t, y_t, x_{t-1}, y_{t-1}, ..., x_{t-k}, y_{t-k})^T$.

A dissimilarity measure based on mutual dependency

A convenient measure of dissimilarity based on their joint dependency is

$$D(X, Y) = |\mathbf{B}_k|^{1/2(k+1)}$$

- Notice that $0 \leq |\mathbf{B}_k| \leq 1$ with equality to one when $\mathbf{B}_k$ is diagonal.
- This measure will be non-negative, symmetric and will be zero if $x = y$.
- The dissimilarity will reach the largest value, one, when the two series are independent, and will be zero if they are identical.

A dissimilarity measure based on mutual dependency

Note that

$$|\mathbf{B}_k| = \left|\mathbf{R}(x)_k\right| \left|\mathbf{R}(y)_k - \mathbf{R}(y,x)_k \mathbf{R}^{-1}(x)_k \mathbf{R}(x,y)_k\right|$$

It should be noticed that if x is integrated then $|\mathbf{R}(x)_k|$ will be close to zero and the product will be small whatever the second term is.

This suggest the *alternative measure*

$$RD(X,Y) = |\mathbf{B}_k|^{1/2(k+1)} / (|\mathbf{R}(x)_k| \cdot |\mathbf{R}(y)_k|)^{1/2(k+1)},$$

which has not this limitation.

The clustering procedure:

We use the dissimilarity defined by

$$RD(X,Y) = |\mathbf{B}_k|^{1/2(k+1)} / (|\mathbf{R}(x)_k| \cdot |\mathbf{R}(y)_k|)^{1/2(k+1)}$$

as input of an agglomerative hierarchical clustering.

The clustering procedure

The nonlinear features of some time series, as for instance, volatility and nonlinear behavior are not indicated by the measures such as simple or partial autocorrelation.

We know that these nonlinear features can be shown by the autocorrelation of the absolute values or the squared residuals of a linear fit.

Suppose that we fit an AR(p) model to the series where p is chosen by the AIC or BIC criterion and we obtain:

$$e_t = y_t - \widehat{\pi}_1 y_{t-1} - \ldots - \widehat{\pi}_p y_{t-p}.$$

Time series clustering by dependence
Synthetic example

Dependent series

- The models for the three populations are:
  1. AR(1) $X_t^{(1,i)} = 0.9X_{t-1}^{(1,i)} + \epsilon_t^{(1,i)}$ with $i = 1, 2, ..., 5$.
  2. AR(1) $X_t^{(2,i)} = 0.2X_{t-1}^{(2,i)} + \epsilon_t^{(2,i)}$ with $i = 1, 2, ..., 5$.
  3. AR(1) $X_t^{(3,i)} = 0.2X_{t-1}^{(3,i)} + \epsilon_t^{(3,i)}$ with $i = 1, 2, ..., 5$.

  That is, the second and the third models have the same autocorrelation structure.
- The five scenarios differs in the dependence structure of the innovations. In the following, we present the autocorrelation matrices of $(\epsilon_t^{(1,1)}, \epsilon_t^{(1,2)}, ..., \epsilon_t^{(3,5)})$.

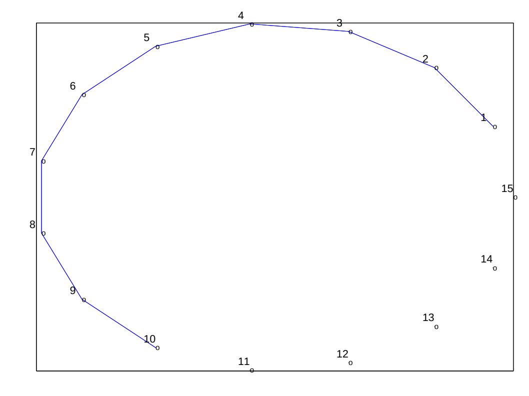
Time series clustering by dependence
Synthetic example

Scenario D.1

$$R_{D.1} = \begin{pmatrix} 1 & .5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ & 1 & .5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ & & 1 & .5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ & & & 1 & .5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ & & & & 1 & .5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ & & & & & 1 & .5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ & & & & & & 1 & .5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ & & & & & & & 1 & .5 & 0 & 0 & 0 & 0 & 0 & 0 \\ & & & & & & & & 1 & .5 & 0 & 0 & 0 & 0 & 0 \\ & & & & & & & & & 1 & 0 & 0 & 0 & 0 & 0 \\ & & & & & & & & & & 1 & 0 & 0 & 0 & 0 \\ & & & & & & & & & & & 1 & 0 & 0 & 0 \\ & & & & & & & & & & & & 1 & 0 & 0 \\ & & & & & & & & & & & & & 1 & 0 \\ & & & & & & & & & & & & & & 1 \end{pmatrix}$$

Time series clustering by dependence
Synthetic example

Scenario D.1



Time series clustering by dependence
Synthetic example

Scenario D.2

$$R_{D.2} = \begin{pmatrix} 1 & .5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ & 1 & .5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ & & 1 & .5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ & & & 1 & .5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ & & & & 1 & .5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ & & & & & 1 & .5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ & & & & & & 1 & .5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ & & & & & & & 1 & .5 & 0 & 0 & 0 & 0 & 0 & 0 \\ & & & & & & & & 1 & .5 & 0 & 0 & 0 & 0 & 0 \\ & & & & & & & & & 1 & 0 & 0 & 0 & 0 & 0 \\ & & & & & & & & & & 1 & .5 & 0 & 0 & 0 \\ & & & & & & & & & & & 1 & .5 & 0 & 0 \\ & & & & & & & & & & & & 1 & .5 & 0 \\ & & & & & & & & & & & & & 1 & .5 \\ & & & & & & & & & & & & & & 1 \end{pmatrix}$$

Time series clustering by dependence
Synthetic example

Scenario D.3

$$R_{D.3} = \begin{pmatrix}
1 & .9 & .9 & .9 & .9 & .9 & .9 & .9 & .9 & .9 & 0 & 0 & 0 & 0 & 0 \\
 & 1 & .9 & .9 & .9 & .9 & .9 & .9 & .9 & .9 & 0 & 0 & 0 & 0 & 0 \\
 & & 1 & .9 & .9 & .9 & .9 & .9 & .9 & .9 & 0 & 0 & 0 & 0 & 0 \\
 & & & 1 & .9 & .9 & .9 & .9 & .9 & .9 & 0 & 0 & 0 & 0 & 0 \\
 & & & & 1 & .9 & .9 & .9 & .9 & .9 & 0 & 0 & 0 & 0 & 0 \\
 & & & & & 1 & .9 & .9 & .9 & .9 & 0 & 0 & 0 & 0 & 0 \\
 & & & & & & 1 & .9 & .9 & .9 & 0 & 0 & 0 & 0 & 0 \\
 & & & & & & & 1 & .9 & .9 & 0 & 0 & 0 & 0 & 0 \\
 & & & & & & & & 1 & .9 & 0 & 0 & 0 & 0 & 0 \\
 & & & & & & & & & 1 & 0 & 0 & 0 & 0 & 0 \\
 & & & & & & & & & & 1 & 0 & 0 & 0 & 0 \\
 & & & & & & & & & & & 1 & 0 & 0 & 0 \\
 & & & & & & & & & & & & 1 & 0 & 0 \\
 & & & & & & & & & & & & & 1 & 0 \\
 & & & & & & & & & & & & & & 1
\end{pmatrix}$$

Time series clustering by dependence
Synthetic example

Scenario D.4

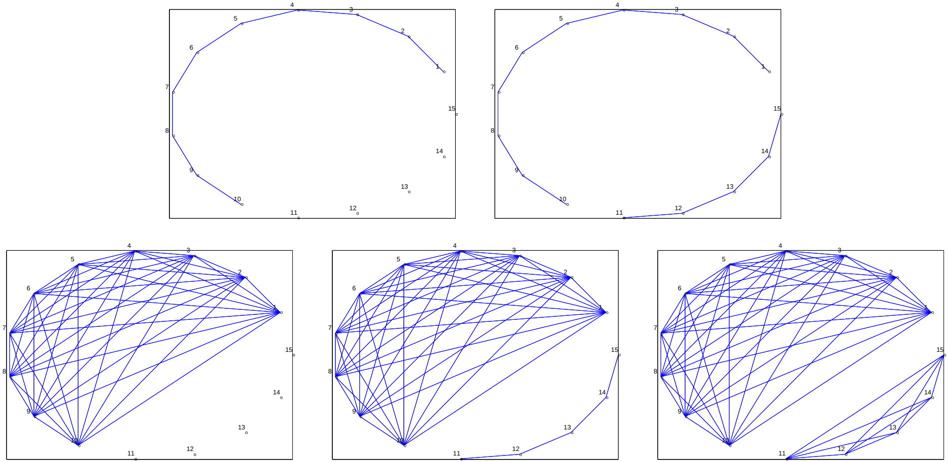
$$R_{D.4} = \begin{pmatrix}
1 & .9 & .9 & .9 & .9 & .9 & .9 & .9 & .9 & .9 & 0 & 0 & 0 & 0 & 0 \\
 & 1 & .9 & .9 & .9 & .9 & .9 & .9 & .9 & .9 & 0 & 0 & 0 & 0 & 0 \\
 & & 1 & .9 & .9 & .9 & .9 & .9 & .9 & .9 & 0 & 0 & 0 & 0 & 0 \\
 & & & 1 & .9 & .9 & .9 & .9 & .9 & .9 & 0 & 0 & 0 & 0 & 0 \\
 & & & & 1 & .9 & .9 & .9 & .9 & .9 & 0 & 0 & 0 & 0 & 0 \\
 & & & & & 1 & .9 & .9 & .9 & .9 & 0 & 0 & 0 & 0 & 0 \\
 & & & & & & 1 & .9 & .9 & .9 & 0 & 0 & 0 & 0 & 0 \\
 & & & & & & & 1 & .9 & .9 & 0 & 0 & 0 & 0 & 0 \\
 & & & & & & & & 1 & .9 & 0 & 0 & 0 & 0 & 0 \\
 & & & & & & & & & 1 & 0 & 0 & 0 & 0 & 0 \\
 & & & & & & & & & & 1 & .5 & 0 & 0 & 0 \\
 & & & & & & & & & & & 1 & .5 & 0 & 0 \\
 & & & & & & & & & & & & 1 & .5 & 0 \\
 & & & & & & & & & & & & & 1 & .5 \\
 & & & & & & & & & & & & & & 1
\end{pmatrix}$$

Time series clustering by dependence
Synthetic example

Scenario D.5

$$R_{D.5} = \begin{pmatrix}
1 & .9 & .9 & .9 & .9 & .9 & .9 & .9 & .9 & .9 & 0 & 0 & 0 & 0 & 0 \\
 & 1 & .9 & .9 & .9 & .9 & .9 & .9 & .9 & .9 & 0 & 0 & 0 & 0 & 0 \\
 & & 1 & .9 & .9 & .9 & .9 & .9 & .9 & .9 & 0 & 0 & 0 & 0 & 0 \\
 & & & 1 & .9 & .9 & .9 & .9 & .9 & .9 & 0 & 0 & 0 & 0 & 0 \\
 & & & & 1 & .9 & .9 & .9 & .9 & .9 & 0 & 0 & 0 & 0 & 0 \\
 & & & & & 1 & .9 & .9 & .9 & .9 & 0 & 0 & 0 & 0 & 0 \\
 & & & & & & 1 & .9 & .9 & .9 & 0 & 0 & 0 & 0 & 0 \\
 & & & & & & & 1 & .9 & .9 & 0 & 0 & 0 & 0 & 0 \\
 & & & & & & & & 1 & .9 & 0 & 0 & 0 & 0 & 0 \\
 & & & & & & & & & 1 & 0 & 0 & 0 & 0 & 0 \\
 & & & & & & & & & & 1 & .9 & .9 & .9 & .9 \\
 & & & & & & & & & & & 1 & .9 & .9 & .9 \\
 & & & & & & & & & & & & 1 & .9 & .9 \\
 & & & & & & & & & & & & & 1 & .9 \\
 & & & & & & & & & & & & & & 1
\end{pmatrix}$$

Synthetic example: Scenarios D.1 - D.5



Synthetic example: Scenarios D.1 - D.5

The following results are the means of the Gravilov index from 10000 replicates for the sets A and B with $T = 100$.

The similarity index used in Gavrilov et al. (2000) compares two different cluster partitions, $C = (C_1, \ldots, C_k)$ and $C' = (C'_1, \ldots, C'_{k'})$ using the following formulas:

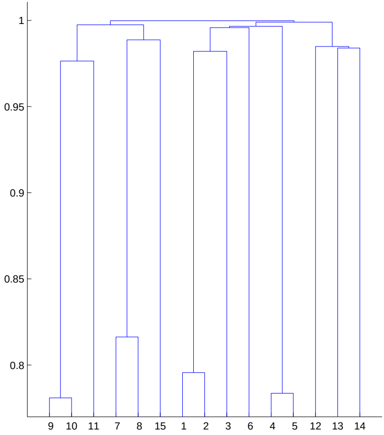
$$Sim(C_i, C'_j) = 2\frac{\#(C_i \bigcap C'_j)}{\#(C_i) + \#(C'_j)},$$

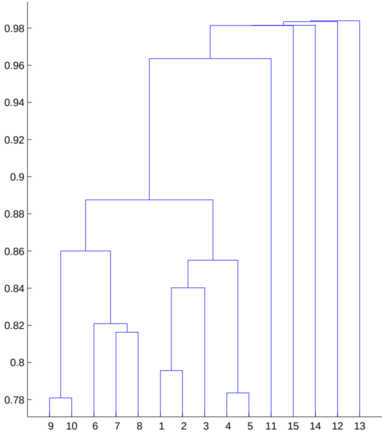
and

$$Sim(C, C') = k^{-1} \sum_{i=1}^{k} \max_{1 \leq j \leq k'} Sim(C_i, C'_j).$$

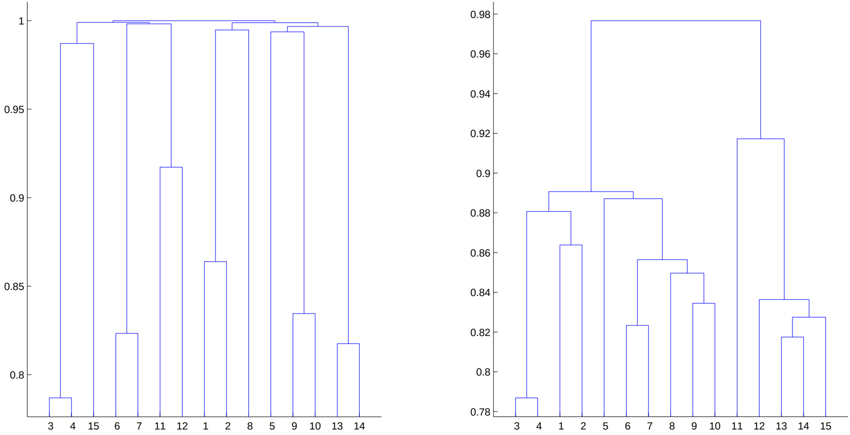
The closer to one the index, the higher is the agreement between the two partitions.

Synthetic example: Scenario D.1

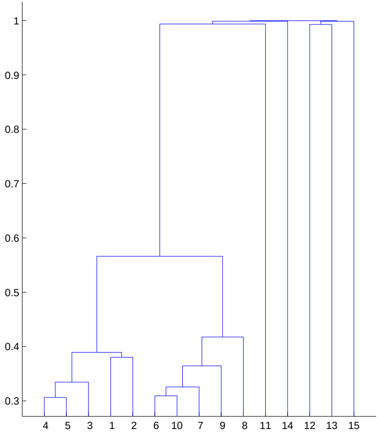


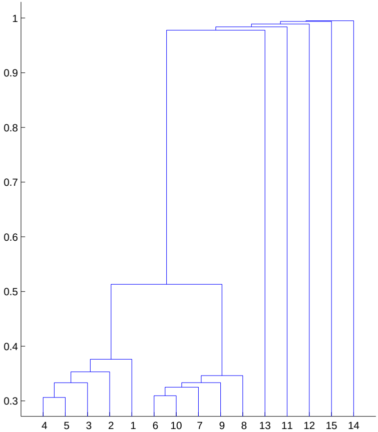


Synthetic example: Scenario D.2

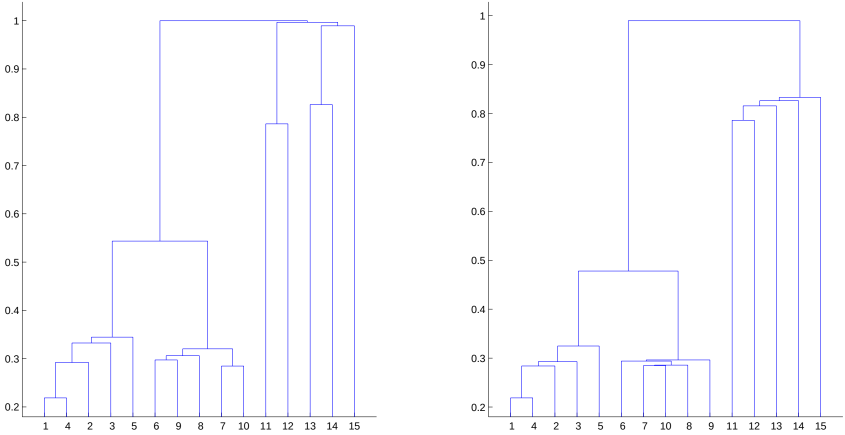


Synthetic example: Scenario D.3

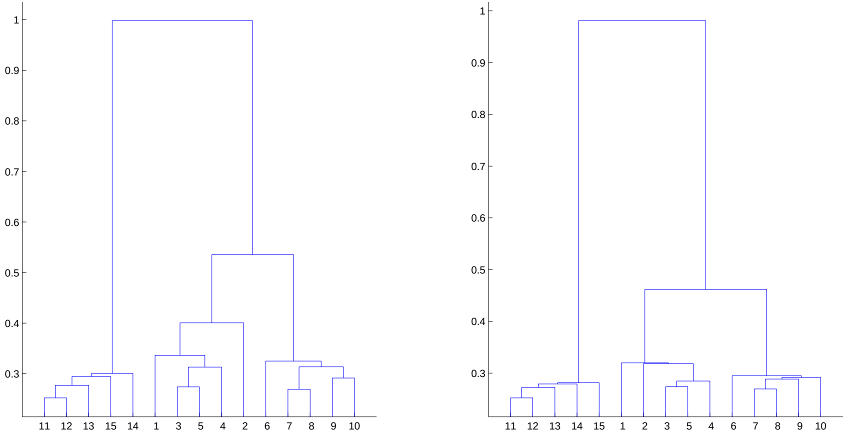




Synthetic example: Scenario D.4



Synthetic example: Scenario D.5



Synthetic example: Scenarios D.1 - D.5

The following results are the means of the Gravilov index from 10000 replicates for the set D using the complete and single linkage

Method	D.1	D.2	D.3	D.4	D.5
SAC	0.443	0.643	0.717	0.665	0.665
PAC	0.491	0.666	0.814	0.678	0.689
D	0.698	0.664	1.000	0.842	1.000
RD	0.527	0.654	1.000	0.865	1.000

Method	D.1	D.2	D.3	D.4	D.5
SAC	0.478	0.666	0.635	0.667	0.667
PAC	0.474	0.666	0.637	0.667	0.667
D	0.923	0.830	1.000	0.988	1.000
RD	0.934	0.843	1.000	0.993	1.000
ABC	-	0.612	-	0.698	0.840

Synthetic example: Scenarios D.1 - D.5

Main conclusions

- The results of the univariate methods are similar and they don't change much across linkage methods.
  - Notice that here a Gravilov index around 0.667 corresponds to approximately separate the first population from the third one in scenarios D.2, D.4 and D.5.
- For scenarios D.3, D.4 and D-5 where there are some "strong" clusters, the complete linkage for both multivariate measures improve the univariate measures.
- For all scenarios, the single linkage and RD is preferable to other considered alternatives.

Case-study with real data- I

Spanish mortality rates

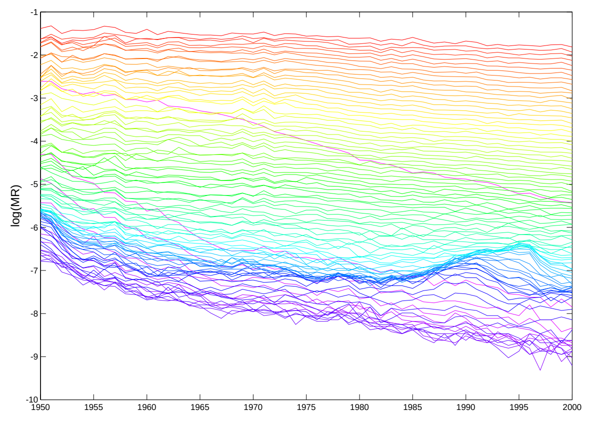
We consider the Spanish mortality rates by age (0 – 90 years) for both genders taken from the Human Mortality Database (http://www.mortality.org).

The data is available from 1908 to 2015. We skip the period 1908 – 1949.

This allows us to use the period 1950 – 2000 as a model adjustment period and 2001 – 2015 as a test period in the forecasting exercise.

Case-study with real data - I

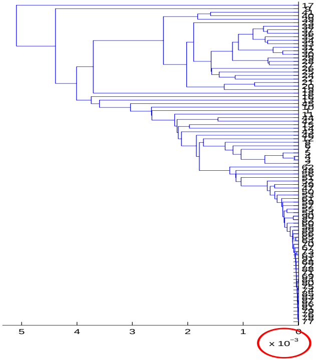
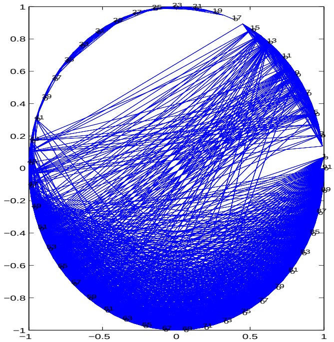
Spanish mortality rates



Case-study with real data: Data description

Spanish mortality rates

It is clear that these series has an strong negative trend. In fact they share a common trend.

Case-study with real data: Data description

Lee-Carter model

It is a well-known model which looks at the dependence between mortality time series. It relates the mortality rates by age with a single unobservable factor:

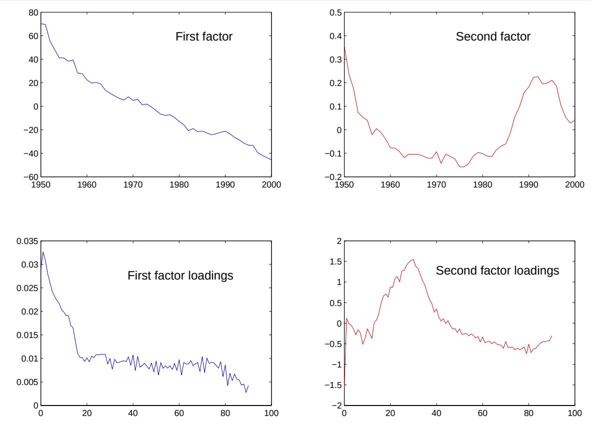
$$\begin{array}{rcl} \ln(MR_{x,t}) & = & a_x + b_x k_t + \varepsilon_{x,t} \\ k_t & = & c + k_{t-1} + \eta_t \end{array},$$

where a_x and b_x are parameters which depend on age, x; k_t is the unobservable factor which picks up the general characteristics of mortality in the year t, and $\varepsilon_{x,t}$ are the age-specific factors.

We will cluster the series of age-specific factors, $\varepsilon_{x,t}$.

Case-study with real data: Factors & Loadings

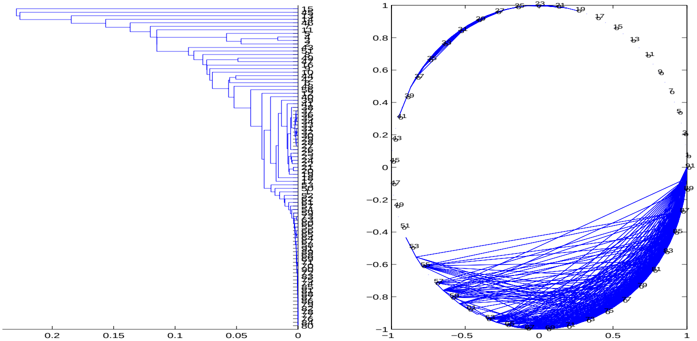
Spanish mortality rates



Spanish mortality rates: Clustering results

Spanish mortality rates

At the age-specific factors, we find two clusters and some "independent" series.



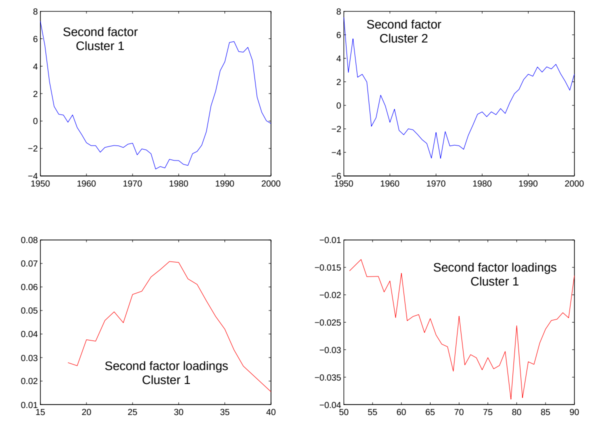
Spanish mortality rates: Clustering results

Here, we will compare the forecasting performance of three models:

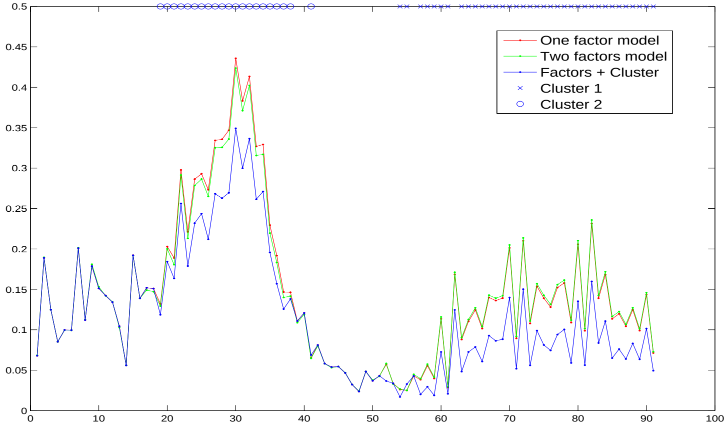
- A factorial model with a single unobservable factor, as in Lee-Carter (1992).
- A factorial model with two unobservable factors, as in Alonso, Peña and Rodríguez (2005).
- A factorial model with two unobservable factors where:
  - the first factor is estimated using all series.
  - the second factor is estimated using the two obtained clusters.

Case-study with real data: Factors & Loadings

Spanish mortality rates

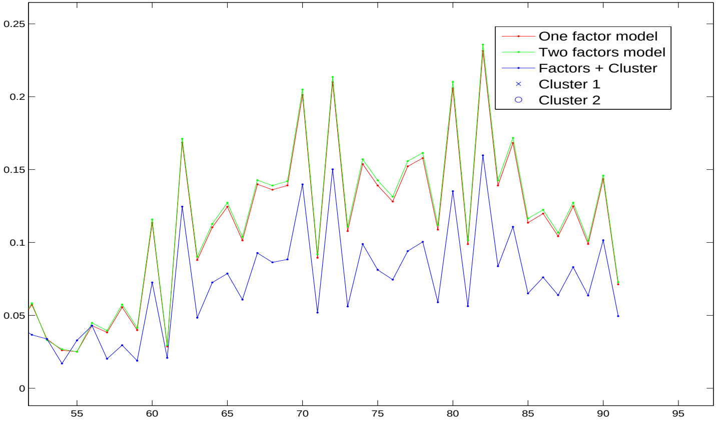


Mean absolute prediction errors



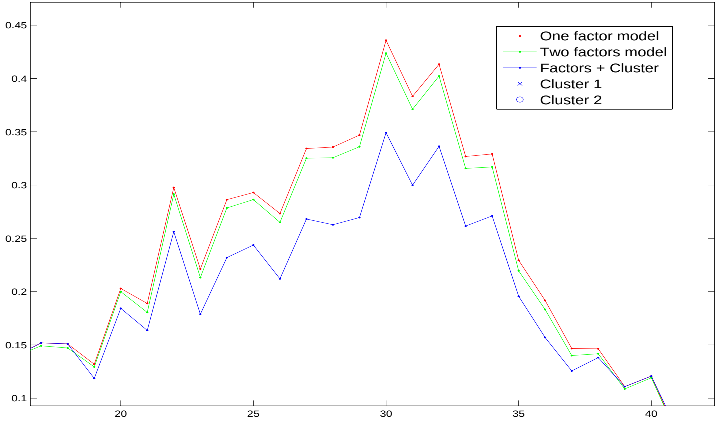
We observe improvements in almost all ages

Mean absolute prediction errors



We observe improvements in ages where two factors is worse than one factor

Mean absolute prediction errors

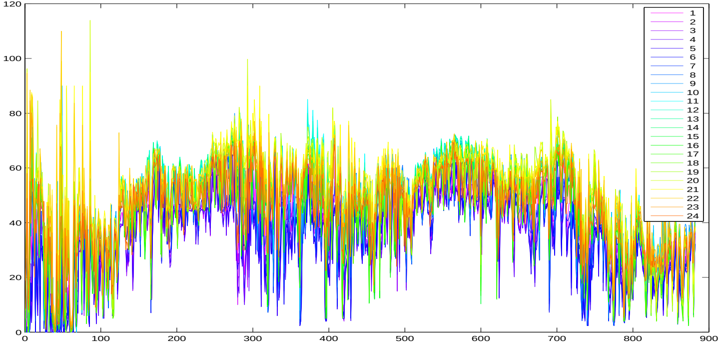


But also in ages where two factors is better than one factor

Case-study with real data- II

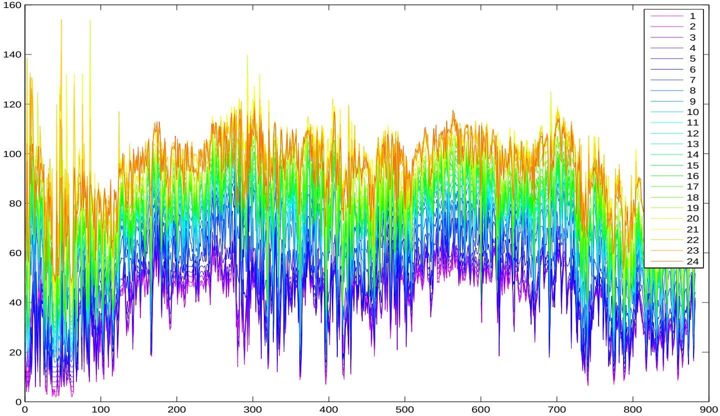
Spanish electricity prices

We study the 24 series of hourly prices for the Iberian electricity market from January 2014 to May 2016.



Case-study with real data- II

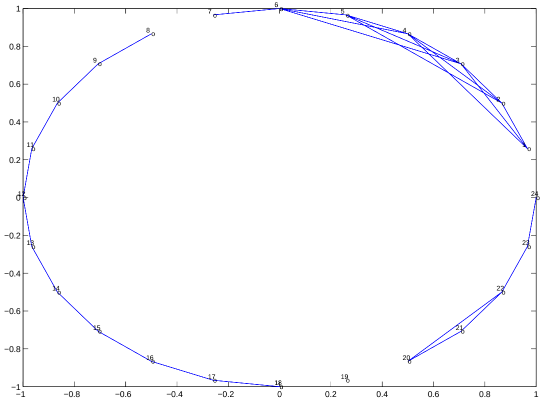
Spanish electricity prices - Translated for better visualization.



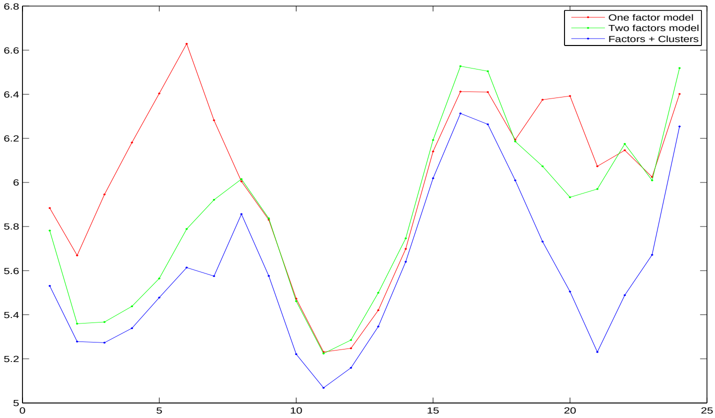
Spanish electricity prices: Clustering results

There are three clusters:

- Sleeping hours
- Working hours
- Arriving & staying at home.



Mean absolute prediction errors



We observe improvements in all hours for one-day-ahead forecast

- Time series clustering by features.
  - Raw data.
  - Autocorrelation.
  - Spectral density.
  - Extreme value behaviour.
- Model based time series clustering.
  - Forecast based clustering.
  - Model with cluster structure.
- Time series clustering by dependence.