P-spline based clustering on Covid-19 mortality trend

Lorenzo Pratesi Mariti

Università degli Studi di Firenze

July 05, 2021

Overview

Index

1. Datasets used.
2. Data Cleaning and Preprocessing.
3. Modelling with P-splines.
4. Clustering.

Datasets used

First dataset: Department of Civil Protection

- Provided by the Department of Civil Protection
  - file name: *dpc-covid19-ita-regioni.json*
  - lists, for each region, the main provincial and regional information on the total number of cases and the collection period.
- Attributes used for the analysis:
  - *data*
  - *denominazione_regione*
  - *deceduti*

Datasets used

Second dataset: Istat

- Provided by the Italian National Institute of Statistics (Istat)
  - file name: *popolazione2019.csv*
  - lists, for each region, the number of exposures.
- Attributes used for the analysis:
  - *Territorio*
  - *Pop2019*

Data Cleaning and Preprocessing

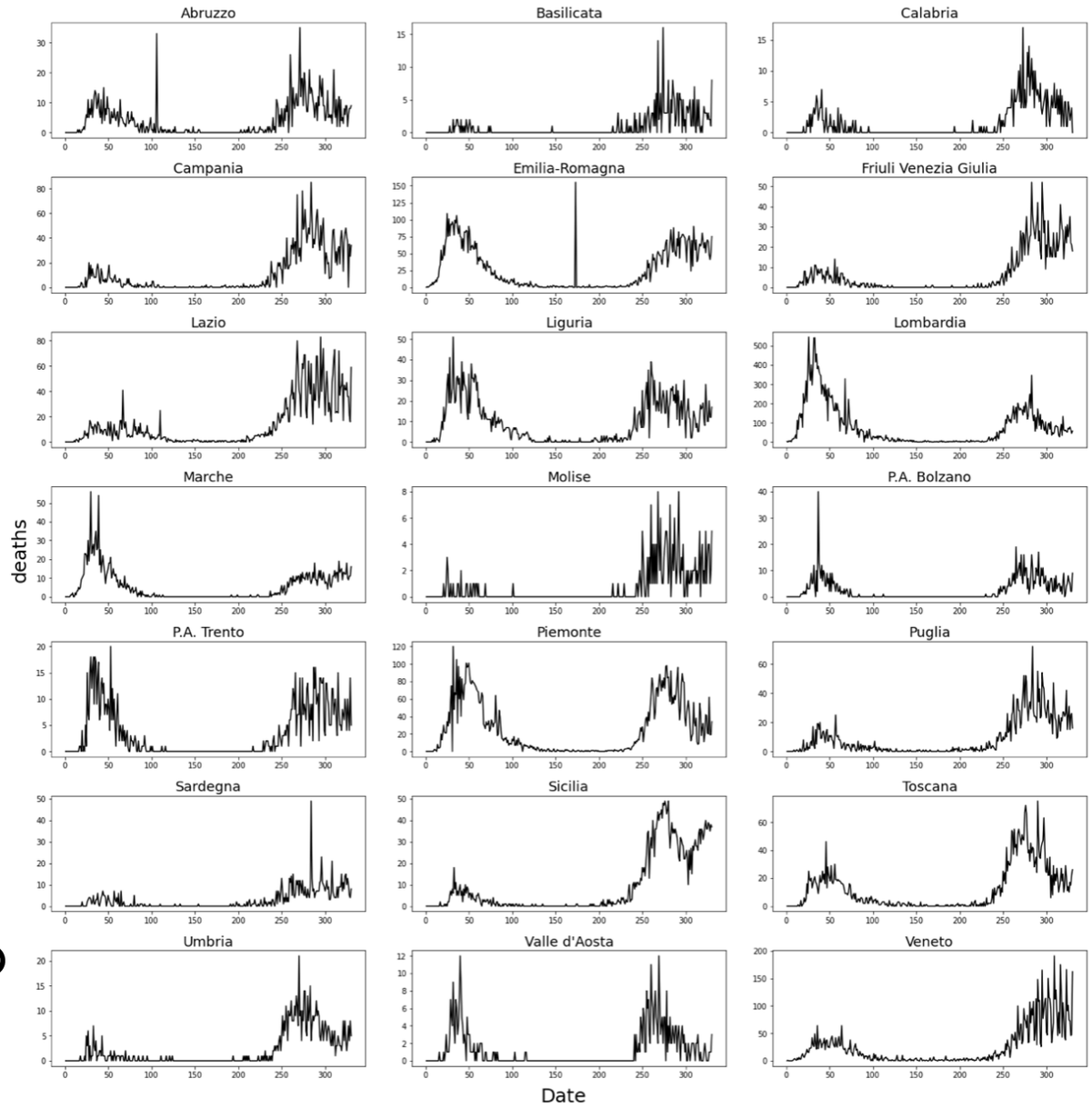
- The main operations carried out on the dataset are:
  - transform the attribute relating to the observation date into YYYY-MM-DD format;
  - remove outlier observations;
  - filter records to get 2020 observations only;
  - transform the deaths attribute from cumulative data to daily data;
  - concatenate the two datasets:
  - join on *denominazione_regione* and *Territorio* attributes.

region	date	population	deaths
Valle d'Aosta	2020-02-25	125666	0.0
Valle d'Aosta	2020-02-26	125666	0.0
Valle d'Aosta	2020-02-27	125666	0.0
Valle d'Aosta	2020-02-28	125666	0.0
...	...	...	...
Sardegna	2020-12-28	1639591	21.0
Sardegna	2020-12-29	1639591	7.0
Sardegna	2020-12-30	1639591	5.0
Sardegna	2020-12-31	1639591	4.0

Data Cleaning and Preprocessing

- Plotting *deaths* attribute.
- We get 21 time series.
- Period of analysis:
  - Feb 25 to Dic 31 (2020).
- Different regional mortality trends

- **Clustering goal**:
  - group the regions according to their temporal trend of mortality.



Mortality trend of each region

Problems to deal with similarity on time series

- The similarity measures considered in clustering for independent data doesn't work well with data that has a time dependency;
  - they ignore the interdependence of the observations.

- Metrics to be used in clustering must cope with:
  - *noise*, *temporal drift*, *longitudinal scaling*, *offset translation*, *linear drift*, *discontinuities*, and *amplitude scaling*.

- Various methods have been developed:
  - Linear Transformation.
  - Shape Based.
  - **Temporal Structure based**.

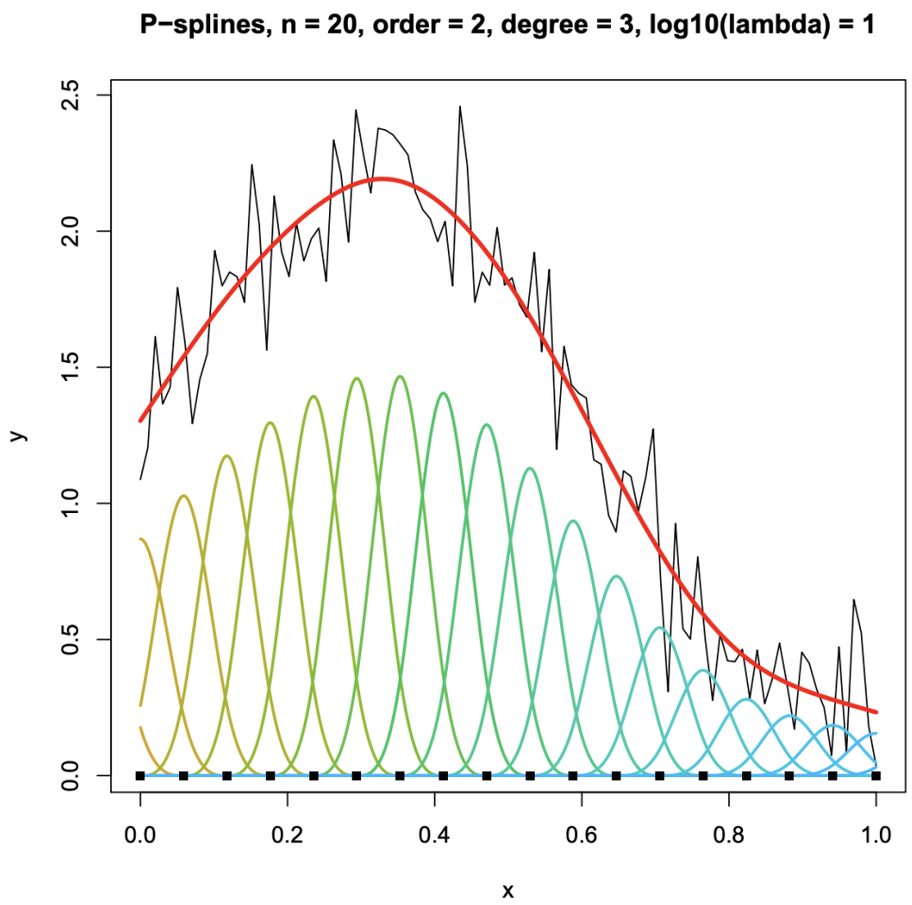
Temporal structure based

- There is an assumption of a statistical model;
- time series can be characterized as a random process whose stochastic parameters can be estimated precisely.
  - Normal Mixture.
  - Autoregressive.
  - Hidden Markov.
  - **Splines**.

P-splines

Eilers and Marx: Statistical Science. [1]

- A flexible tool for smoothing, based on regression.
- Local basis functions: B-splines.
- P-splines = B-spline + Penalization



- Computation:
  - Do regression on (cubic) B-splines
  - Use equally spaced knots
  - Put a difference penalty (order 2 or 3) on the coefficients
  - Tune smoothness with λ (penalty weight)

P-splines

What kind of penalty?

- Given: data series y_i, $i = 1, \ldots, m$
- Wanted: a smooth series z
- Two (conflicting) goals: fidelity to y and smoothness
- Fidelity, sum of squares: $S = \sum_i (y_i - z_i)^2$
- How to quantify smoothness?
  - Use roughness instead: $R = \sum_i (z_i - z_{i-1})^2$

P-splines

Penalized least squares

- Combine fidelity and roughness

$$Q = S + \lambda R = \sum_i (y_i - z_i)^2 + \lambda \sum_i (z_i - z_{i-1})^2$$

- Parameter λ sets the balance

- Operator notation: $\Delta z_i = z_i - z_{i-1}$

$$Q = S + \lambda R = \sum_i (y_i - z_i)^2 + \lambda \sum_i (\Delta z_i)^2$$

P-splines

Matrix-vector notation

- Penalized least squares objective function

$$Q = \|\boldsymbol{y} - \boldsymbol{z}\|^2 + \lambda \|\boldsymbol{D}\boldsymbol{z}\|^2$$

- Differencing matrix $\boldsymbol{D}$, such that $\boldsymbol{D}\boldsymbol{z} = \Delta\boldsymbol{z}$
- Explicit solution: $\hat{\boldsymbol{z}} = (\boldsymbol{I} + \lambda \boldsymbol{D}'\boldsymbol{D})^{-1}\boldsymbol{y}$
- Easy to implement in R and Python

```
m ← length(y)
E ← diag(m) # Identity matrix
D ← diff(E) # Difference operator
G ← E + lambda * t(D) %*% D
z ← solve(G, y) # Solve the equations
```

P-splines

Non-normal data, smoothing of counts

- Fidelity measured by the sum of squares
- How will we handle counts?
  - Generalized Linear Model, use penalized (log-)likelihood

- Given: a series y of counts, we model a smooth linear predictor η
- Assumption: $y_i \sim Pois(\mu_i)$, with $\eta_i = \log(\mu_i)$
- The roughness penalty is the same, but fidelity measured by deviance:

$$Q = 2\sum_i(\mu_i - y_i\eta_i) + \lambda\sum_i(\Delta^d\eta_i)^2$$

- Let's go back to our problem.

Model

Input

- Let m be the number of days available in the dataset, and n be the number of regions.

- The proposed model requires two datasets as input data:
  - the matrix of deaths $\boldsymbol{Y} = (y_{t,r})$;
  - the vector $\boldsymbol{e}_r$, regional population exposures to the risk of death.

$$\boldsymbol{Y} = \begin{pmatrix} y_{1,1} & \cdots & y_{1,n} \\ y_{2,1} & \cdots & y_{2,n} \\ \vdots & \ddots & \vdots \\ y_{m,1} & \cdots & y_{m,n} \end{pmatrix}; \quad \boldsymbol{E} = \mathbf{1}_m \boldsymbol{e}_r' = \begin{pmatrix} e_{1,1} & \cdots & e_{1,n} \\ e_{2,1} & \cdots & e_{2,n} \\ \vdots & \ddots & \vdots \\ e_{m,1} & \cdots & e_{m,n} \end{pmatrix}.$$

Model

Assumptions

- Assuming deaths to be Poisson distributed.

$$y_{t,r} \sim \mathcal{P}(e_{t,r}\mu_{t,r})$$

- Model the Poisson death counts via a log-link function:

$$\eta_{t,r} = \ln(\mu_{t,r}) = \boldsymbol{B}\boldsymbol{\alpha}_r, \qquad \boldsymbol{\mu} = \exp(\boldsymbol{\eta})$$

  - where $\boldsymbol{B}$ is a matrix of B-splines, which are common for all regions.
  - Regional specific coefficients $\boldsymbol{\alpha}_r$ are estimated in a P-spline setting.

- The penalty is subtracted from the log-likelihood $(\boldsymbol{y}, \boldsymbol{\alpha})$ to form the penalized likelihood function

$$L = l(\boldsymbol{y}, \boldsymbol{\alpha}) - \frac{1}{2}\boldsymbol{\alpha}'\boldsymbol{P}\boldsymbol{\alpha} \quad \text{where} \quad \boldsymbol{P} = \lambda \boldsymbol{D}'\boldsymbol{D}$$

- The optimization of L leads to the following system of equations:

$$\boldsymbol{B}'(\boldsymbol{y} - \boldsymbol{\mu}) = \boldsymbol{P}\boldsymbol{\alpha}$$

- These are solved as usual with iterative weighted linear regressions with the system

$$\hat{\boldsymbol{\alpha}}_{(t+1)} = (\boldsymbol{B}'\hat{\boldsymbol{W}}_{(t)}\boldsymbol{B} + \boldsymbol{P})^{-1}\left[\boldsymbol{B}'\hat{\boldsymbol{W}}_{(t)}\boldsymbol{B}\boldsymbol{\alpha} + \boldsymbol{B}'(\boldsymbol{y} - \hat{\boldsymbol{\mu}}_{(t)})\right]$$

Model

Penalized-GLM

- The estimated coefficients $\boldsymbol{\mathcal{A}} = [\hat{\boldsymbol{\alpha}}_1, ..., \hat{\boldsymbol{\alpha}}_r, ..., \hat{\boldsymbol{\alpha}}_n]$ contain all relevant features of the observed regional mortality trends
- our aim is to classify them, and consequently classify $\mu_{t,r}$.

Select the roughness parameter

How we tune λ to get the best fit for each region?

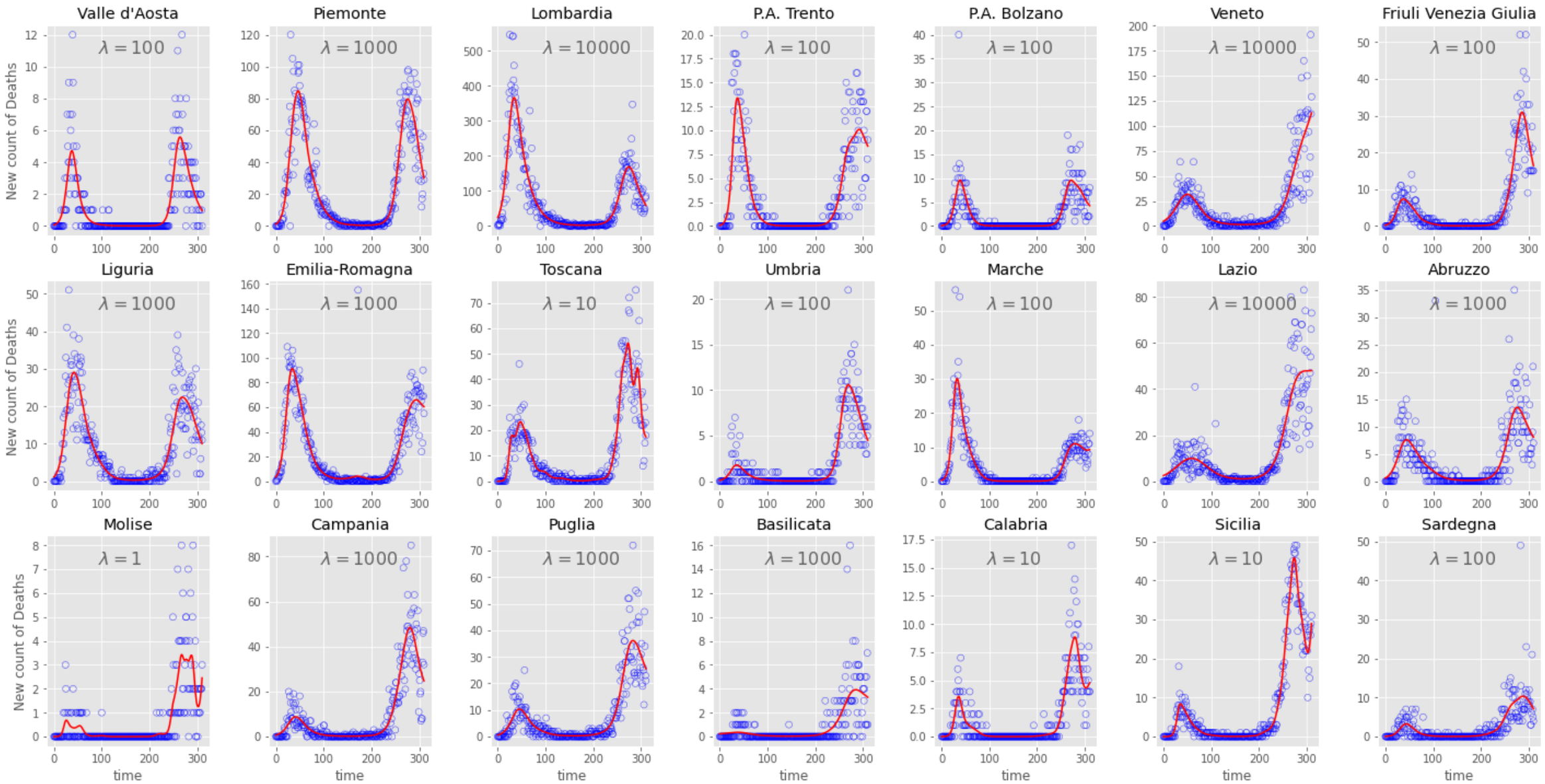
- *out-of-sample forecast accuracy* procedure:
  - out-of-sample window: $y_t = y_{m-n}, \ldots, y_m$;
  - evaluate the corresponding forecasts: $\hat{y}_t = \hat{y}_{m-n}, \ldots, \hat{y}_m$
  - compute the root-mean-square error for a given λ

$$\text{RMSE}(\lambda) = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(\hat{y}_t(\lambda) - y_t)^2}.$$

- Repeat for each $\lambda \in \{10^i\}_{i=0}^{5}$ and select the one that minimizes the RMSE.

Model

P-spline fitting for each region



Clustering

- **Goal**: grouping the n vectors of $\boldsymbol{\mathcal{A}}$ into K sets.
- **Idea**: perform the classification task on the reduced space spanned by optimal spline coefficients.
  - Any clustering algorithms and distance measure can be used to cluster the coefficients of P-spline.
- **Methods**: *K-means* and *Hierarchical*

Clustering

K-means

- **Dissimilarity measure**: Euclidean distance between the estimated coefficients of the B-splines

$$d(x_i, x_{i'}) = \sum_{j=1}^{n} (x_{ij} - x_{i'j})^2 = \|x_i - x_{i'}\|^2$$

- The problem is to choose $z = \{c_1, \ldots, c_k\}$ which minimizes

$$\frac{1}{n} \sum_{i=1}^{n} \min_{c \in z} d(\hat{\boldsymbol{\alpha}}_i, c)$$

- that is equivalent to looking for a partition $\{C_1, \ldots, C_k\}$ of $\boldsymbol{\mathcal{A}}$ in k classes such that

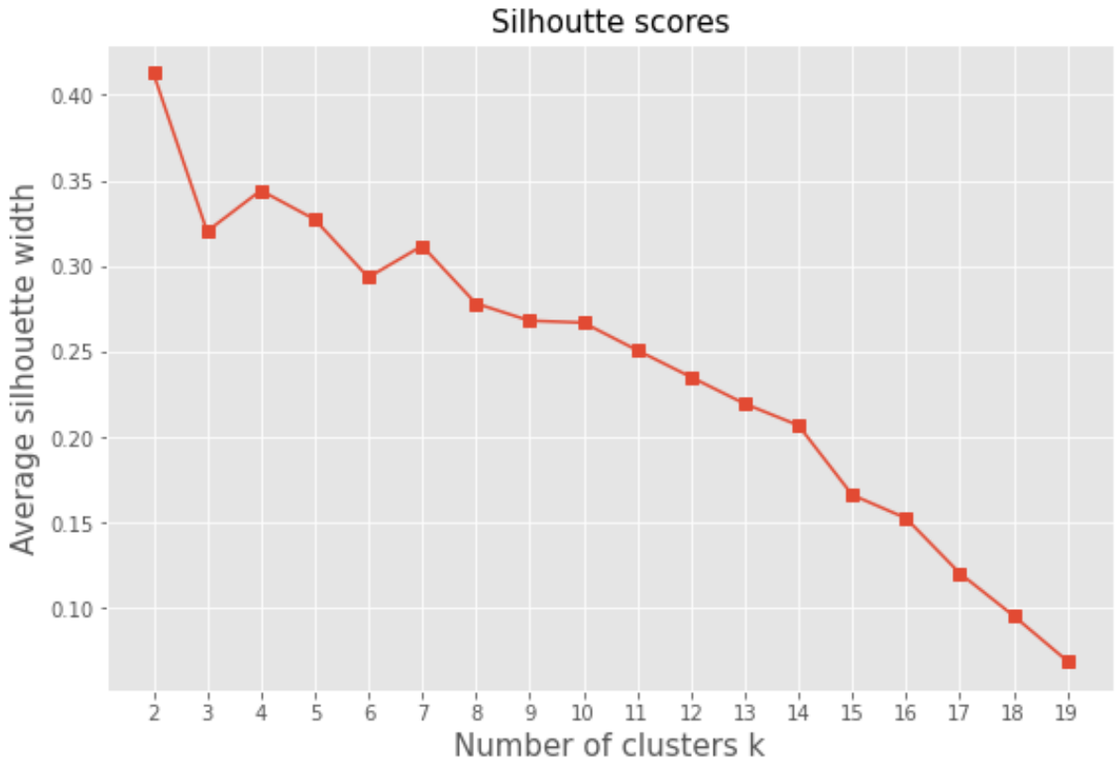
$$\frac{1}{n} \sum_{j=1}^{k} \sum_{\hat{\boldsymbol{\alpha}}_i \in C_j} d(\hat{\boldsymbol{\alpha}}_i, c_j)$$

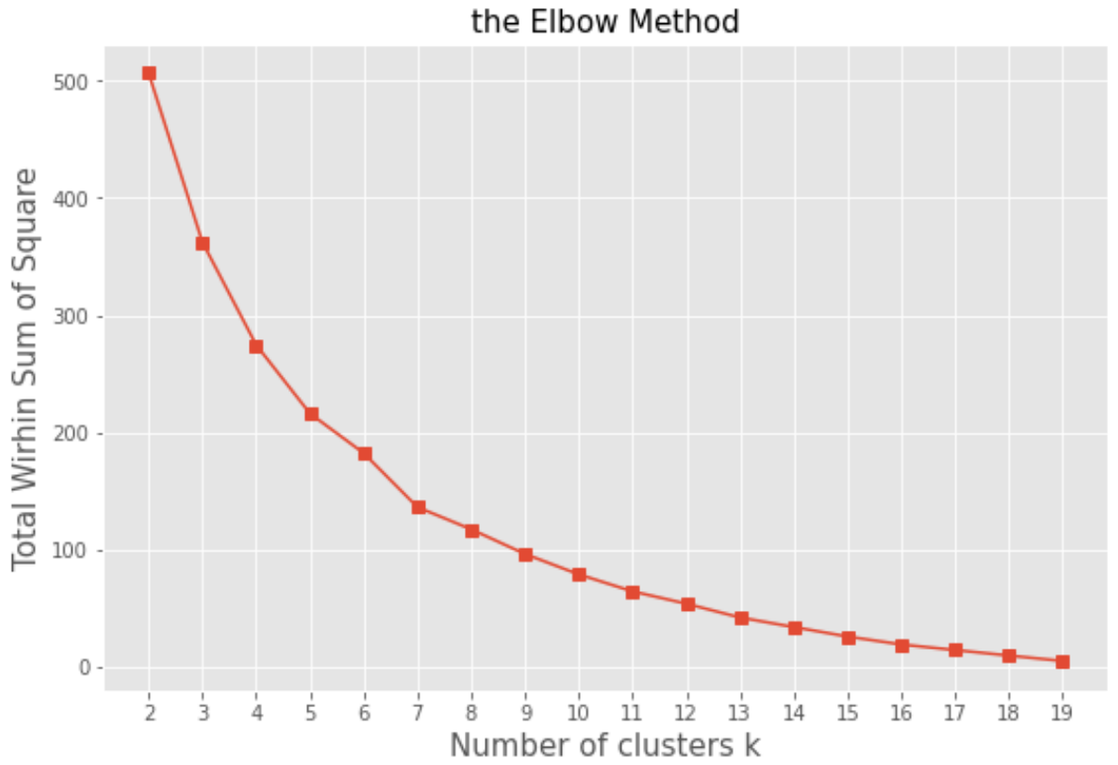
attains its minimum.

Clustering

K-means

- Choosing the optimal parameter K
  - *Silhouette*, *SSE* (Elbow Method) and *Pearson's correlation index*

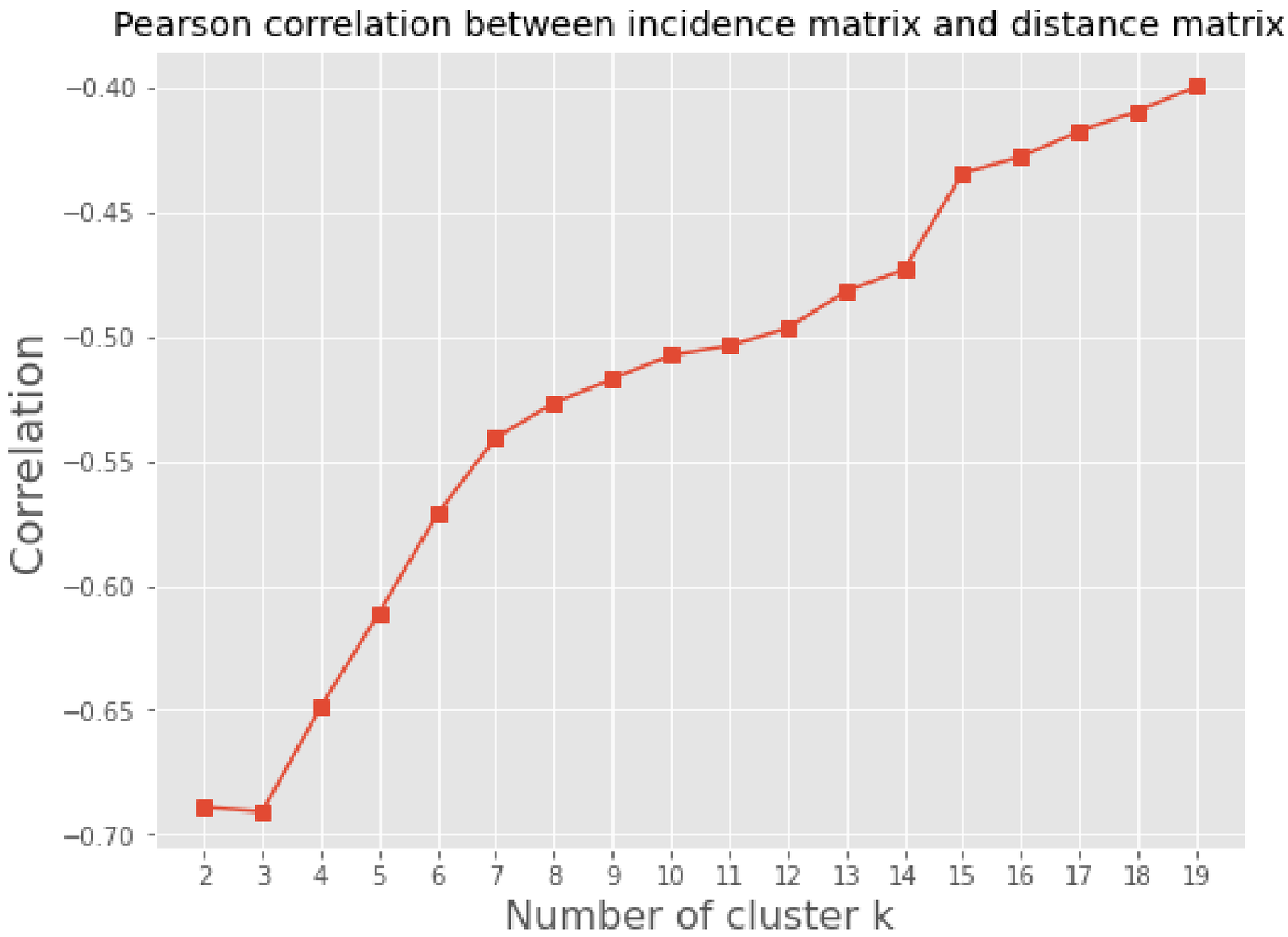




Clustering

K-means

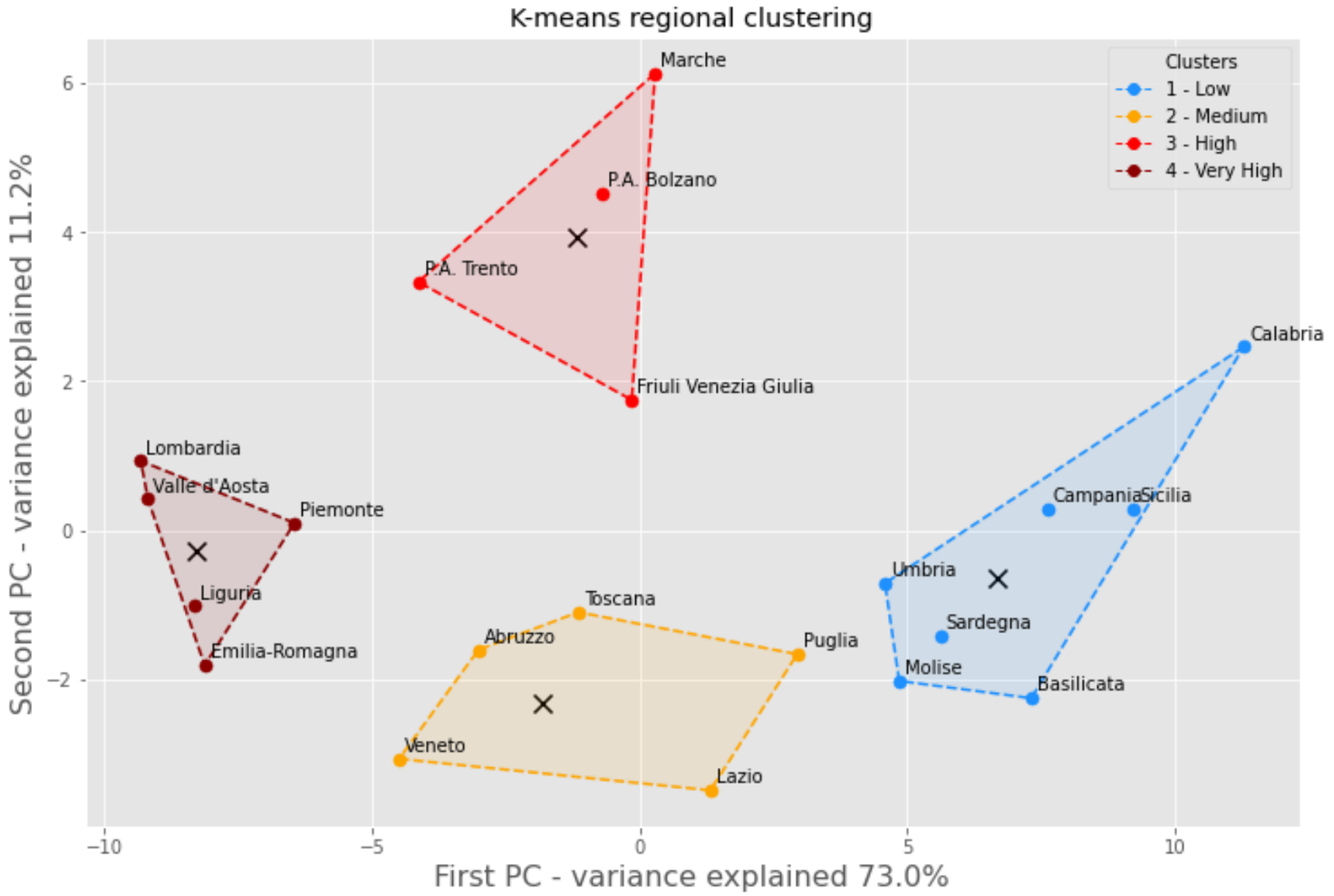
- Choosing the optimal parameter K
  - *Silhouette*, *SSE* (Elbow Method) and *Pearson's correlation index*



Clustering

K-means

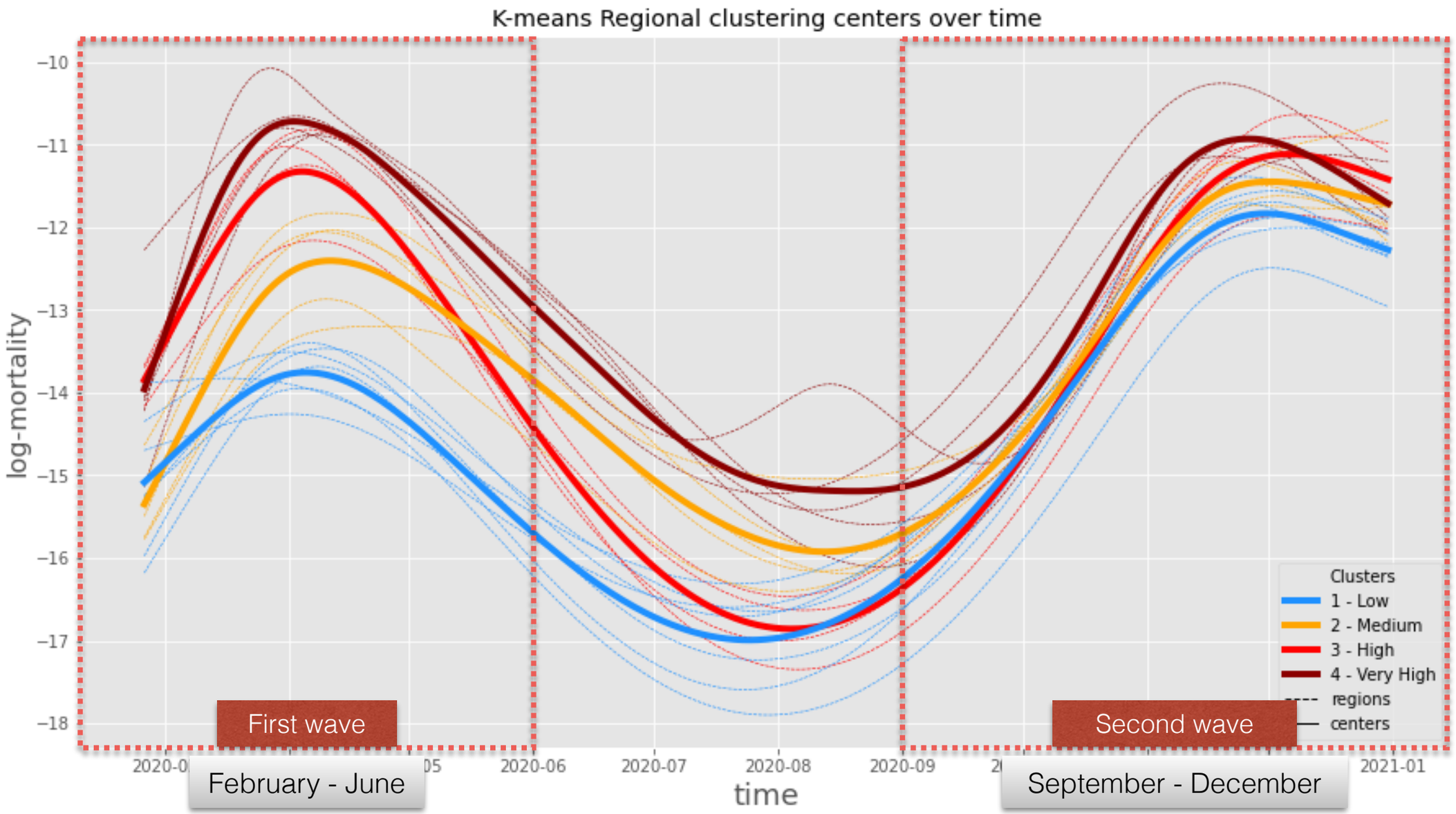
- Clusters are labeled on 4 mortality levels:
  - **Low**, **Medium**, **High**, **Very High**.



Clustering

K-means

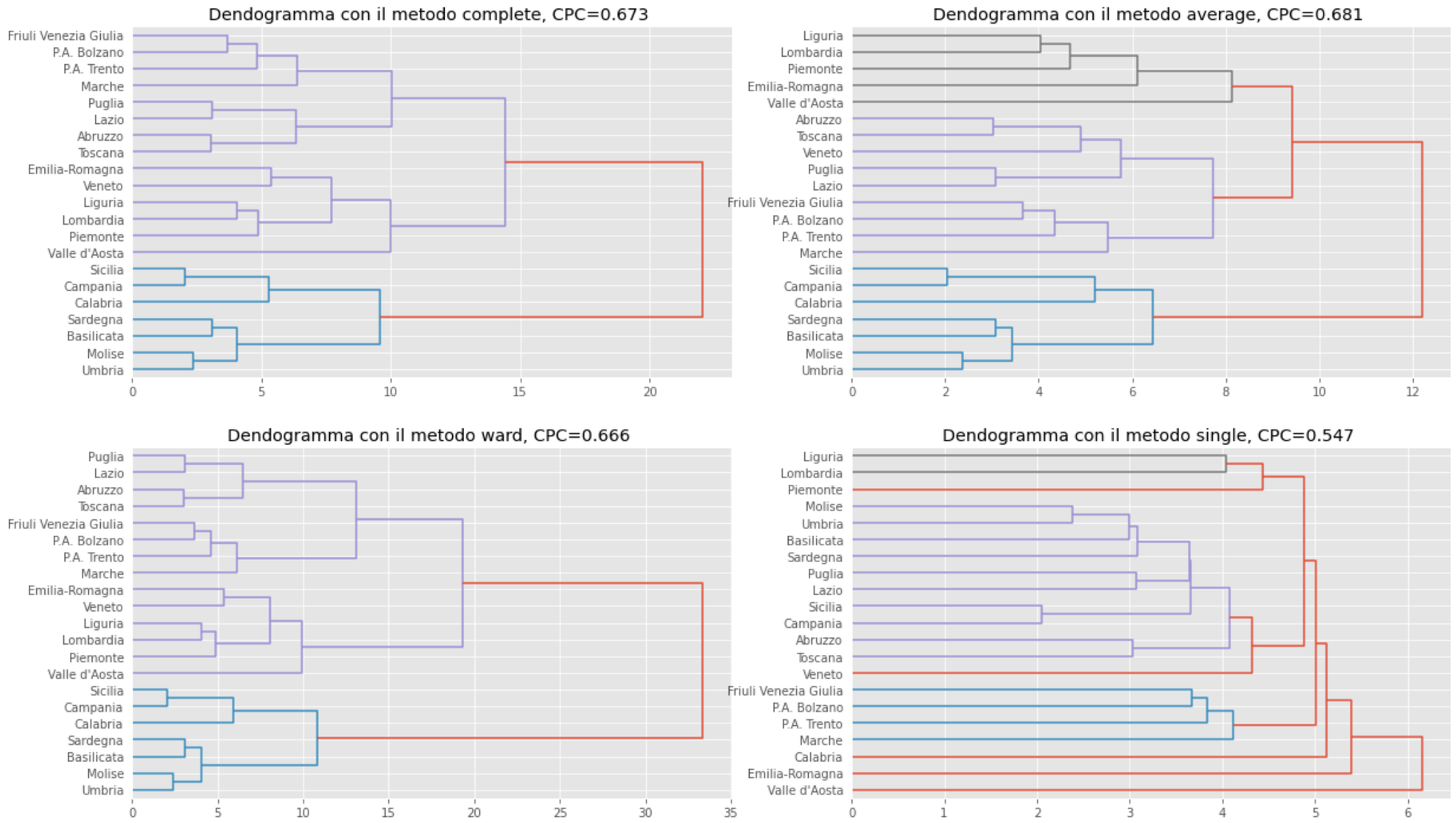
- Curve centers



Clustering

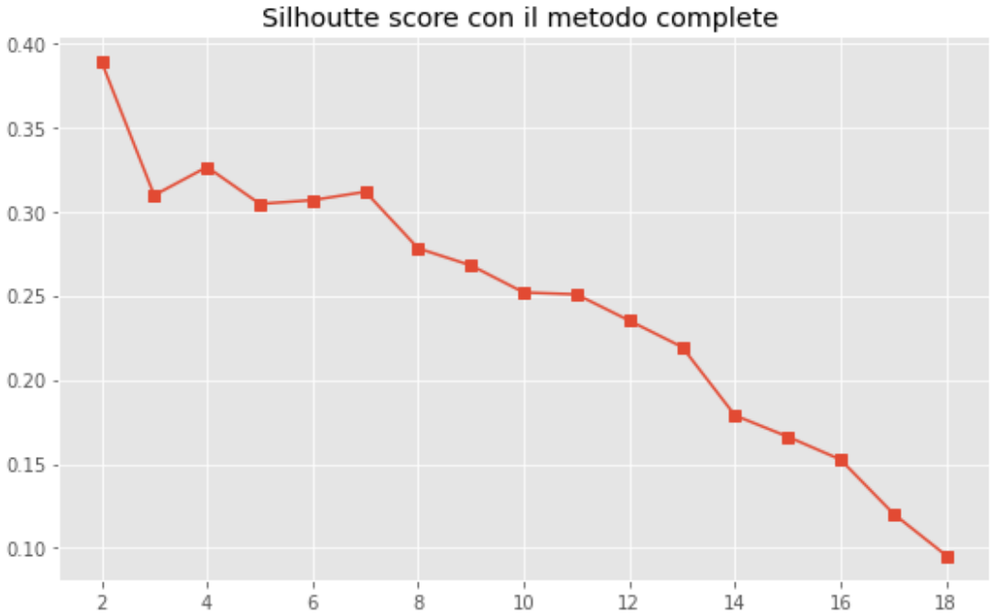
Hierarchical clustering

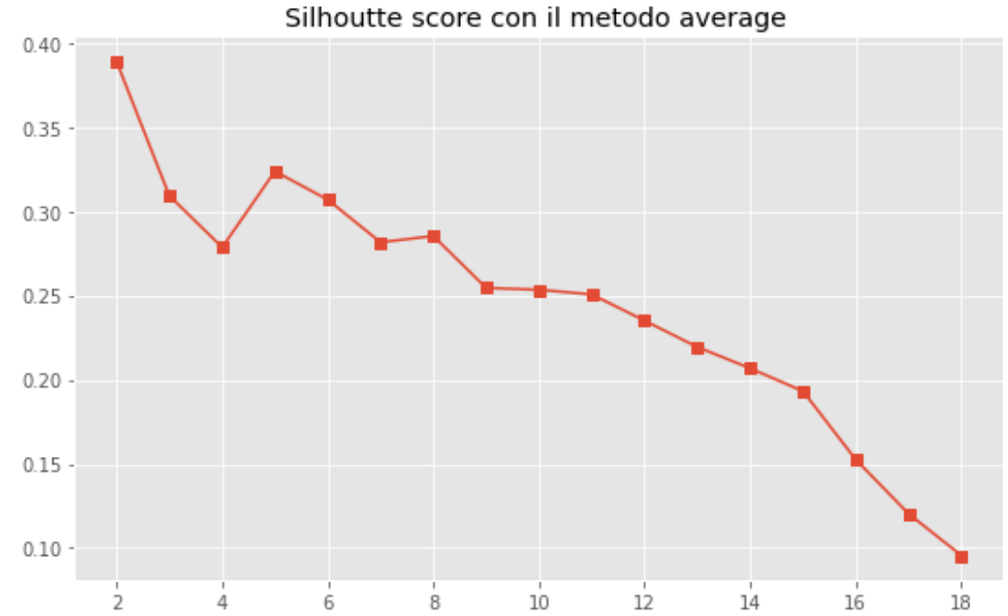
- Construction of the dendrogram: *single*, *complete*, *average* and *ward*.
- Goodness of clustering evaluated with: the *Silohoutte* and *Cophonetic* index.

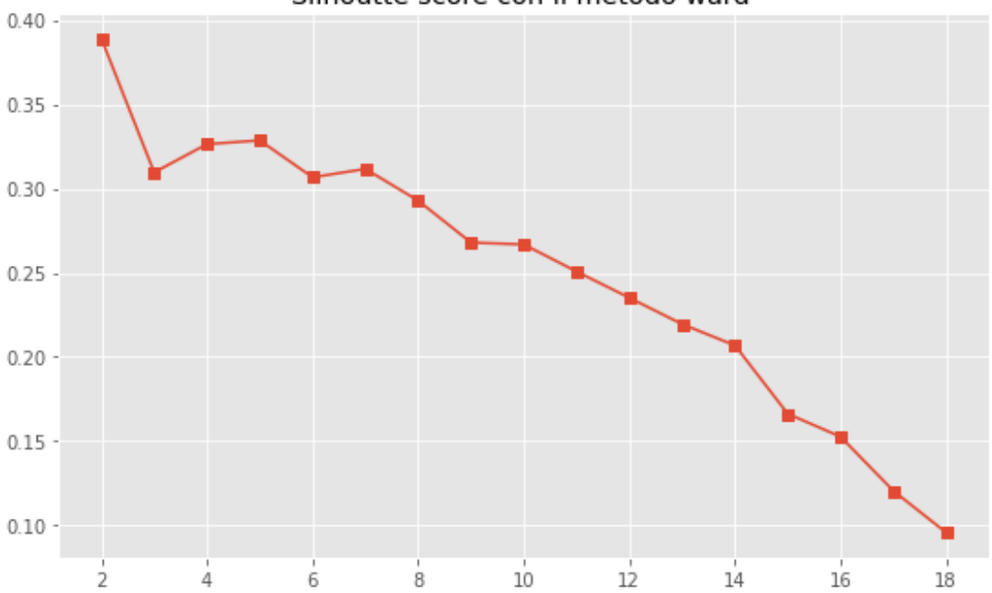


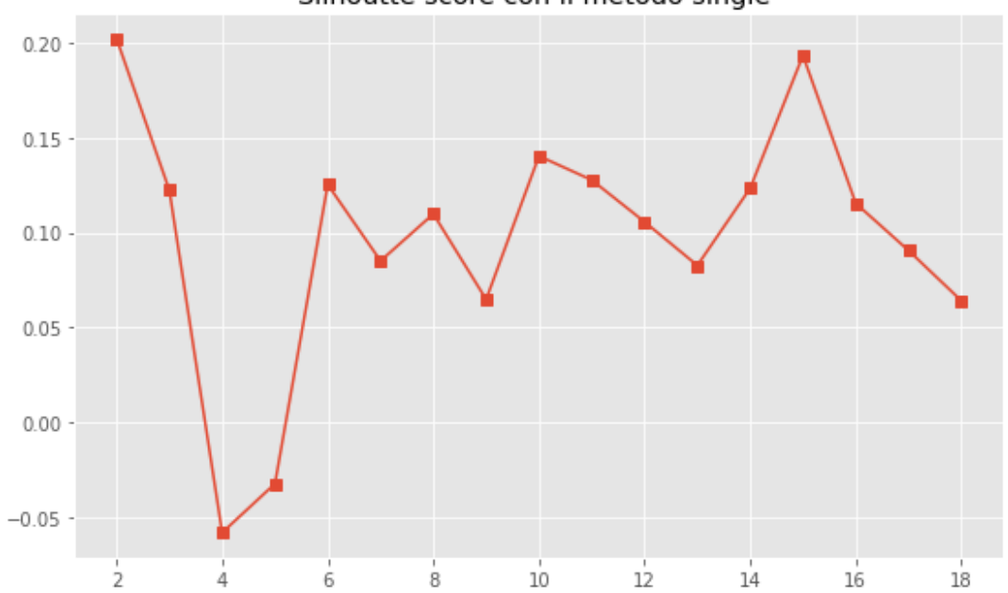
Clustering

Hierarchical clustering



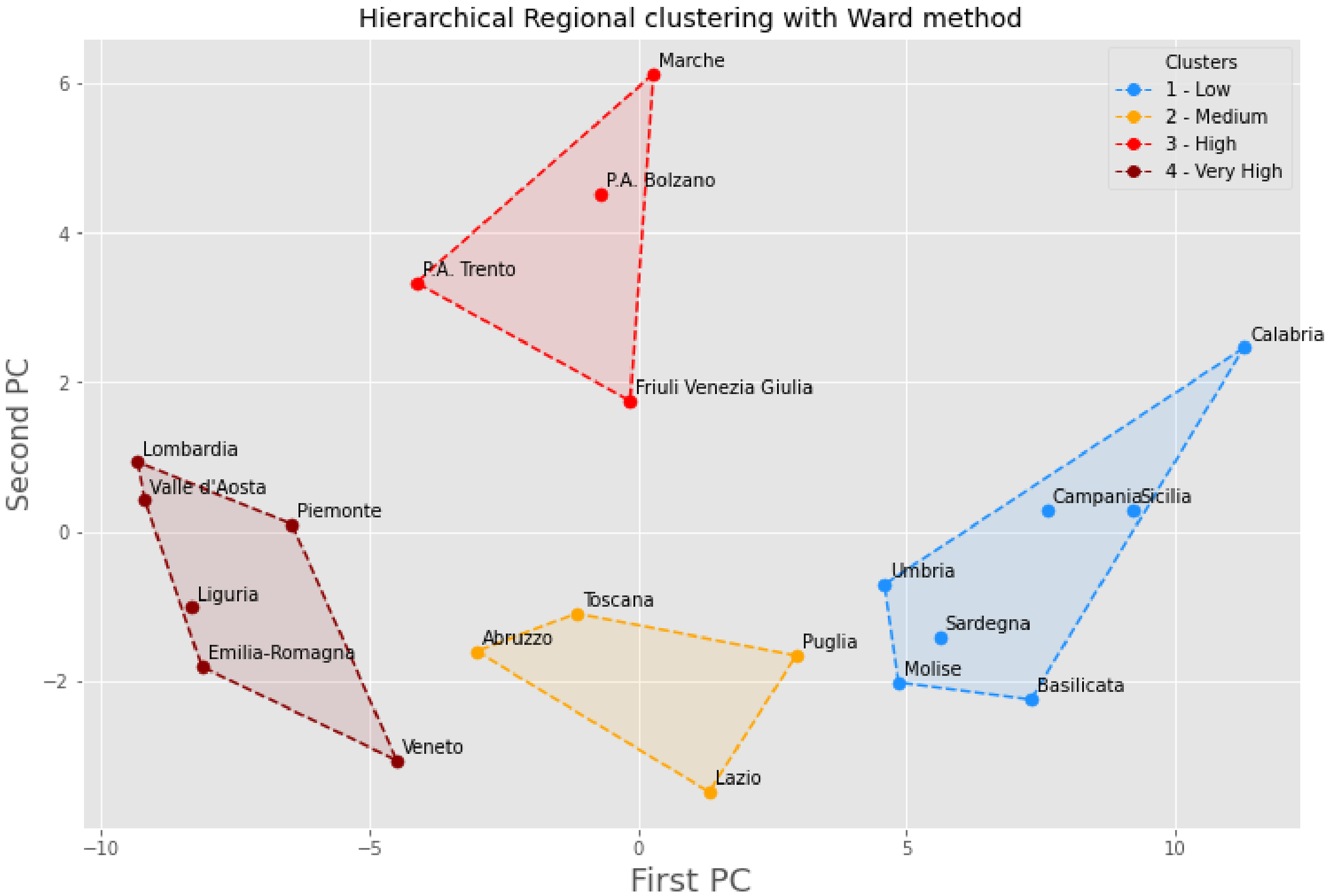






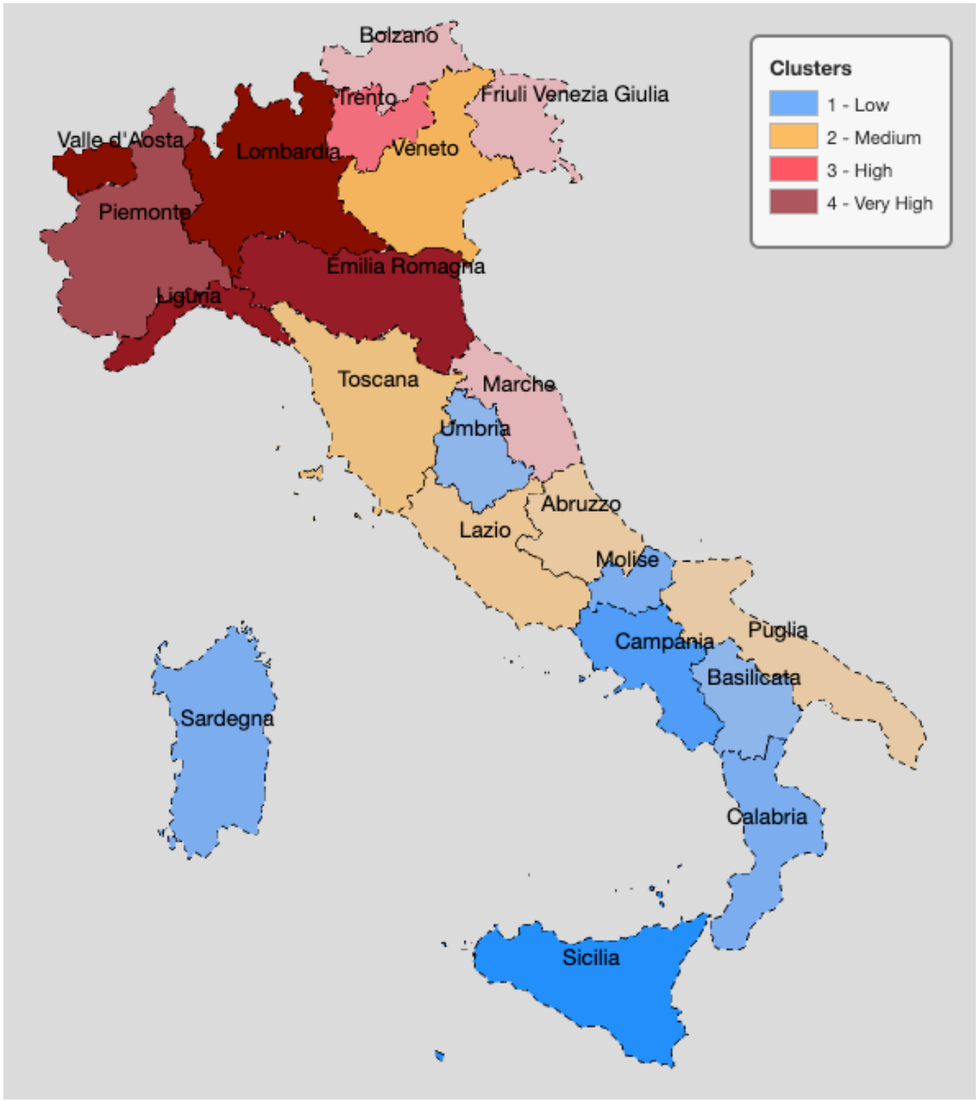
Clustering

Hierarchical clustering

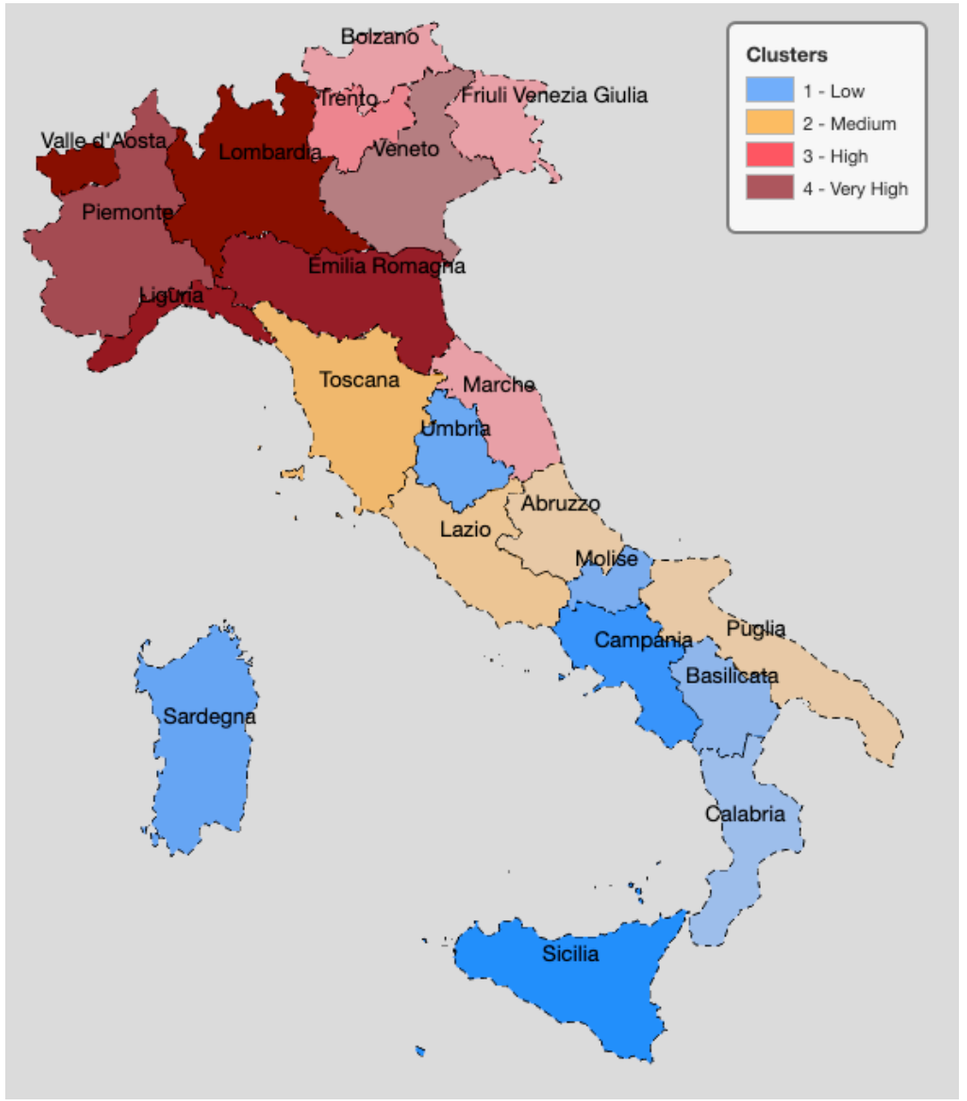


Clustering

Results



K-means



Hierarchical

Conclusion

- Two types of clustering applied to time series on Covid-19 regional mortality.
- Modelling data using B-spline with penalization curves.
- It was possible to group regions with different Covid-19 mortality trends.
- Groups:
  - **Low** - southern and island regions (Mezzogiorno).
  - **Medium** - central regions.
  - **High** - north-eastern regions.
  - **Very High** - north-western regions.
- Framework applicable for other countries and other variables.

Bibliography I


Marx BD Eilers PHC. *Flexible Smoothing with B-splines and Penalties.* Statistical Science, 1996.


D. R. Brillinger. "A biometrics invited paper with discussion: The natural variability of vital rates and associated statistics. Biometrics". In: (1986).


Marx BD Eilers PHC. *Flexible Smoothing with B-splines and Penalties.* Statistical Science, 1996.


Dipartimento Della Protezione Civile (2020). *Dataset of Covid-19 infected cases in Italy*. https://github.com/pcm-dpc/Covid-19/tree/master/.


Resident population and demographic indicators for year 2019. http://demo.istat.it/pop2019/index1.html.

Bibliography II

Carmada. *R package. MortalitySmooth - Smoothing and Forecasting Poisson Counts with P-Splines*. https://www.rdocumentation.org/packages/MortalitySmooth/.

The End