UNIVERSITÀ DI PISA

Data Mining and Machine Learning

Quality of Life: Clustering of Italian Provinces

Iacopo Bicchierini

Github: `https://github.com/Bicchie/Quality-of-Life-Italian-Provinces-Clustering`

Contents

Introduction

This Cluster Analysis is made for educational purpose, in particular for the Data Mining and Machine Learning course, the main aim is to try different Clustering techniques in order to get familiar with them and with the concept behind machine learning in general.
The choice to use the IlSole24Ore dataset regarding the qualify of life in the italian provinces is driven by my curiosity in understand these data and the desire to extract some useful knowledge from them.

Chapter 1

Business Understanding

Usually when dealing with Data Mining process the standard is **CRISP-DM** (**CR**oss **I**ndustry **S**tandard **P**rocess for **D**ata **M**ining) that helps to focus on the right aspects.

Business Objectives

I would like to offer an extended analysis finalized in a report. I will consider as possible **stakeholders**: decision makers in both local and state government and also entrepreneurs. This analysis could be useful in order to have further understanding of province similarity and differences, give glimpse about how to direct government funding and where are possible opportunity to do business for entrepreneurs.

Data Mining Goals

From a technical data mining perspective I will consider this work has a success if some patterns and useful insight about the current and past situation of provinces will be extracted, as well as some trend in the membership and composition of clusters extracted from the same group of indicators using 2020, 2021 and 2022 data. So summarizing:

- **Cluster Analysis** : extracts some interesting patterns on how the provinces are clustered using different groups of indicators and different Clustering techniques.
- **Temporal Cluster Analysis** : extracts some interesting pattern on how this Quality of Life indicators changed during the last 3 years.

Chapter 2

Data Understanding

The Dataset is a collection of life quality indexes gathered by **IlSole24Ore** *a leading Italian multimedia publishing organization, operating in the economic, financial, professional, and cultural information sector.*
Data is publicly available from this GitHub repository: https://github.com/IlSole24ORE. Will be considered the dataset from 2022, 2021 and 2020 denominated respectively **QVD2022**, **QVD2021** and **QVD2020**.

2.1 Data Description

For the whole set of 107 Province we have 6 groups of 15 indexes regarding different aspects of Life Quality. So the rows are 9630 = 107*6*15. They are:

- **Wealth and Consumption**
- **Business and Work**
- **Justice and Security**
- **Demography and Society**
- **Environment and Service**
- **Culture and Leisure**

Considering the 3 versions of the dataset from year 2020, 2021 and 2022, It is worth mentioning how the number of indexes and groups remains constant but some indicators change in some group. For example some COVID-19 indicators disappears in 2022 indicators or the Inflation is considered only in 2022, because of the energy crisis etc. Despite this, most of the indicators do not change and in any case the group of phenomena is analysed in a coherent way in the 3 datasets.

Name of the field	Description	Format	Example
PROVINCE NAME (ISTAT)	Extended name of province Istat	String	*Monza e della Brianza*
CODICE NUTS	NUTS Code	String	*ITC4D*
PROVINCE CODE ISTAT	Code of the province	Integer	*108*
CURRENT DENOMINATION	Name by which the province is commonly known	String	*Monza e Brianza*
VALUE	Measure of phenomenon	Float	*687.7859771*
INDICATOR	Short name of the phenomenon	String	*Home e corporate banking*
UNIT OF MEASURE	Description of the unit in which the phenomenon is measured	String	*Per thousand inhabitants*
TIME REFERENCE	Period to which the measurement of the phenomenon refers	String	*31 December 2020*
ORIGINAL SOURCE	Name of the source that produced the data or from which they were taken for subsequent processing by Il Sole 24 ORE	String	*Banca d'Italia*

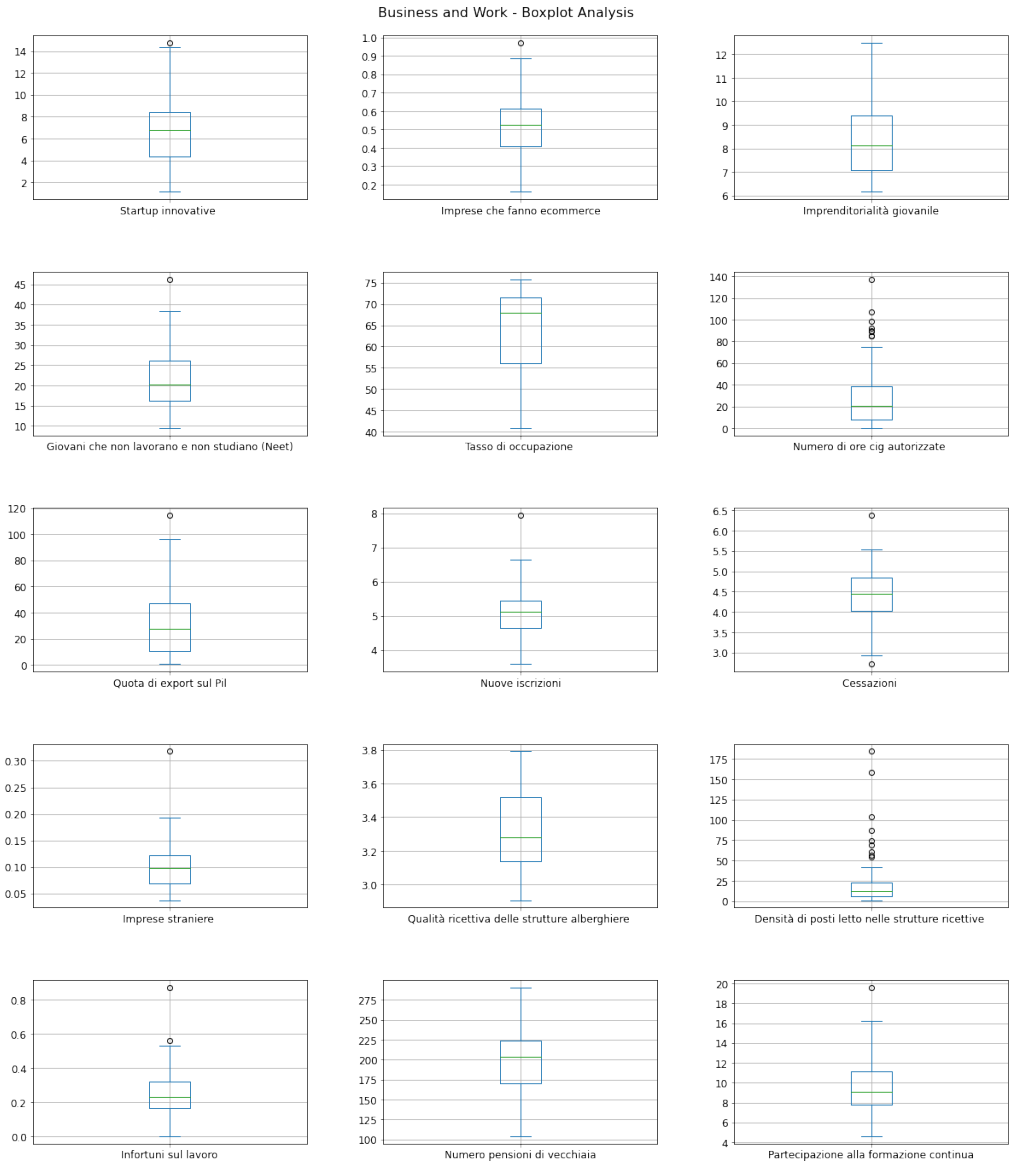
Figure 2.1: Dataset

Before considering the EDA (Exploratory Data Analysis) is worth to divide into the 6 groups the original file and prune the useless columns, keeping only **Province Name**, **Value** and **Indicator**.

2.2 Exploratory Data Analysis

Let's analyze the 6 groups of indicators concerning the Quality of Life in the italian provinces. The analyzes are done using **boxplot** and **correlation analysis**.

Business and Work



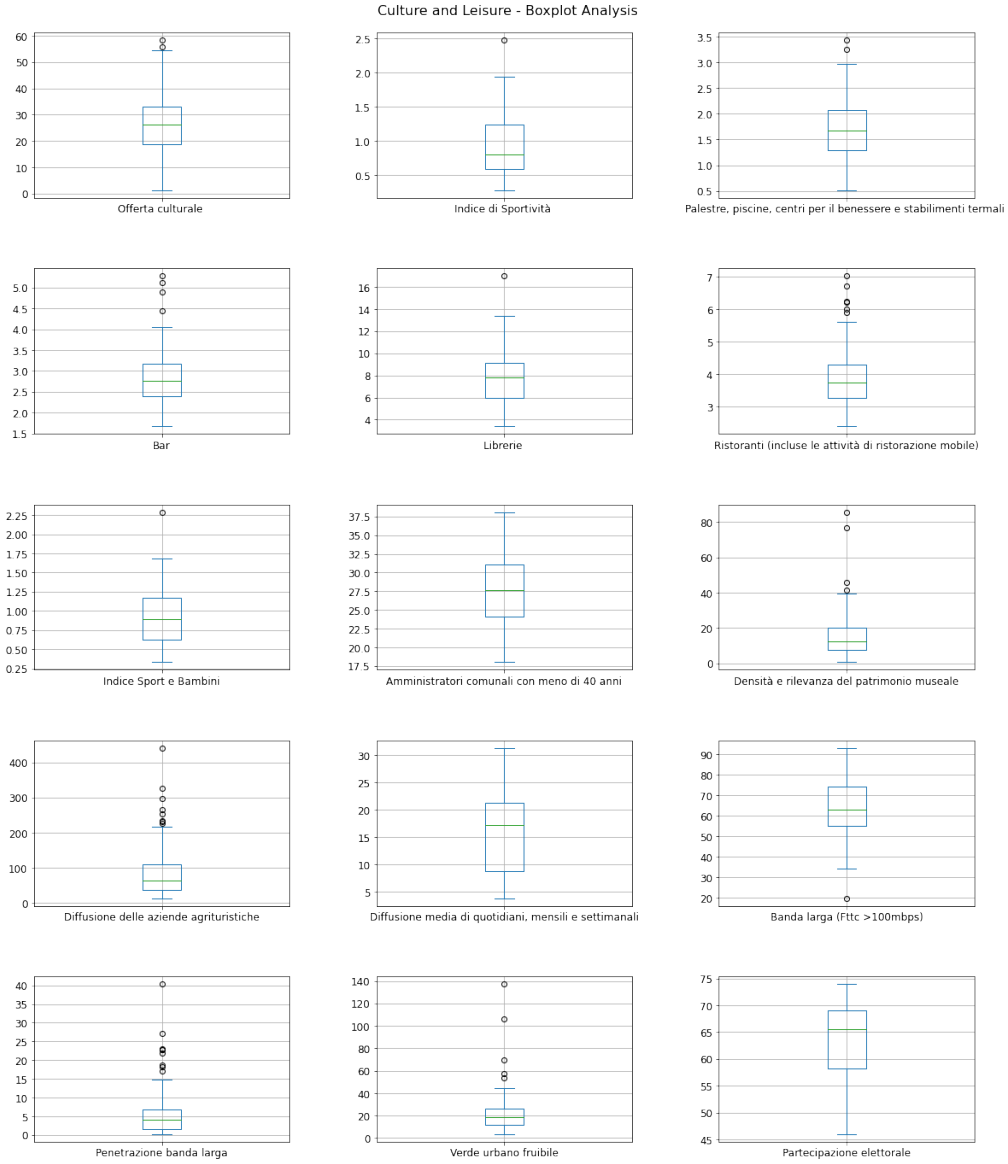
There can be plenty comments regarding these data, for example the **density of beds in accommodation facilities** indicator has a lot of outliers, this can be due to some highly touristic regions. Particular indicators are also **innovative startup** and **employment rate** that show the difference in economic development in different provinces. Also the **share of exports on GDP** varies considerably, from almost 0 to over 100.

	Cessazioni	Densità di posti letto nelle strutture ricettive	Giovani che non lavorano e non studiano (Neet)	Imprenditorialità giovanile	Imprese che fanno ecommerce	Imprese straniere	Infortuni sul lavoro	Numero di ore cig autorizzate	Numero pensioni di vecchiaia	Nuove iscrizioni	Partecipazione alla formazione continua	Qualità ricettiva delle strutture alberghiere	Quota di export sul Pil	Startup innovative	Tasso di occupazione
Cessazioni															0.701139
Densità di posti letto nelle strutture ricettive															
Giovani che non lavorano e non studiano (Neet)									-0.73334						-0.893498
Imprenditorialità giovanile															-0.77364
Imprese che fanno ecommerce															
Imprese straniere															
Infortuni sul lavoro															
Numero di ore cig autorizzate															
Numero pensioni di vecchiaia			-0.73334												0.832873
Nuove iscrizioni															
Partecipazione alla formazione continua															
Qualità ricettiva delle strutture alberghiere															
Quota di export sul Pil															
Startup innovative															
Tasso di occupazione	0.701139		-0.893498	-0.77364					0.832873						

Figure 2.2: Correlation Analysis of Business and Work

An unexpected correlation is the negative correlation between **youth entrepreneurship** and **employment rate**, instead is understandable the positive correlation between **number of old-age pensions** and **employment rate** since we expect to have more space for the young generation if the older one retire from work.

Culture and Leisure

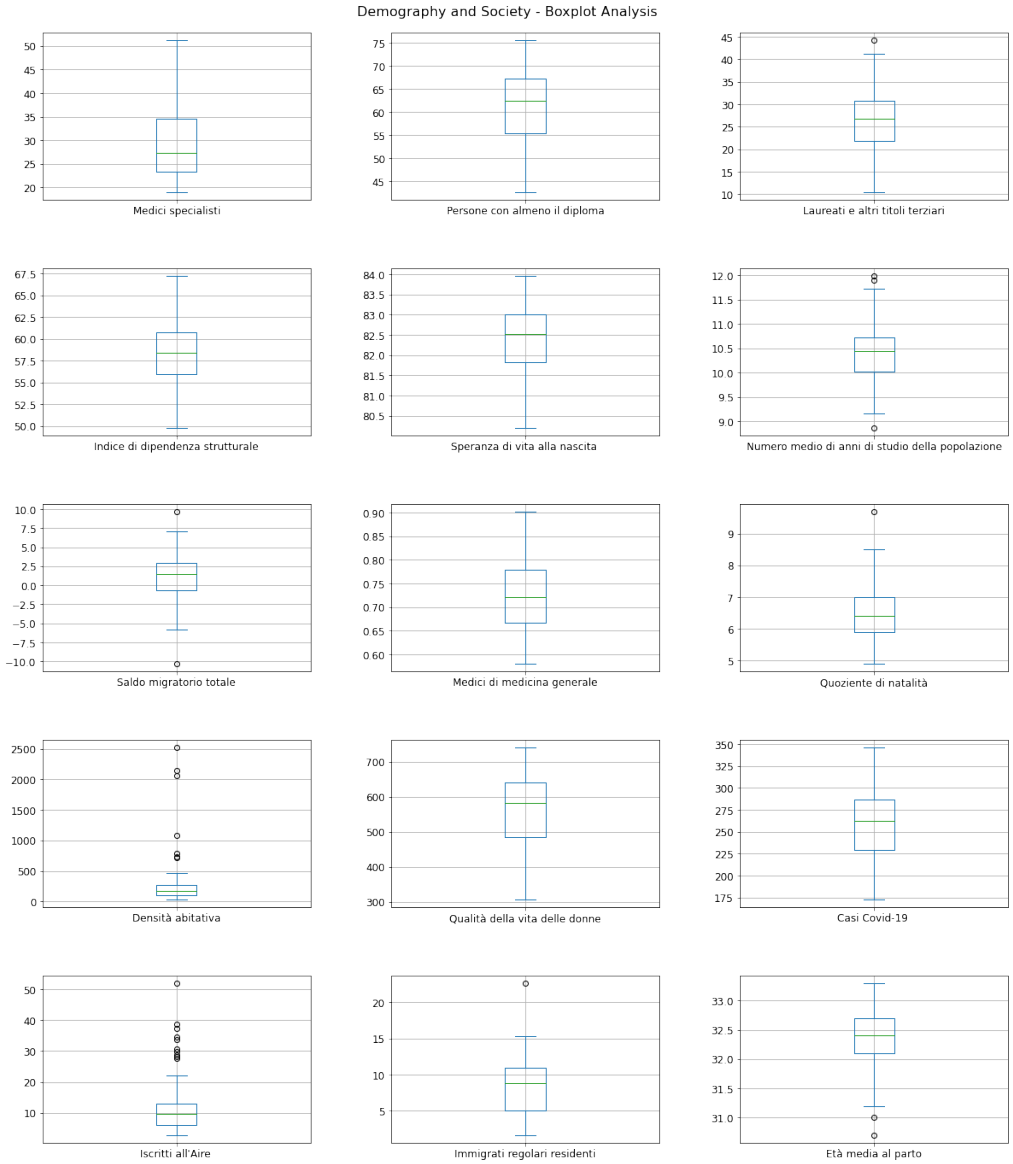


We can find an high number of outliers in the **broadband penetration**, there is a huge disparity in province regarding this type of infrastructure. Another observation is the huge differences in **electoral participation**.

	Amministratori comunali con meno di 40 anni	Banda larga (Fttc >100mbps)	Bar	Densità e rilevanza del patrimonio museale	Diffusione delle aziende agrituristiche	Diffusione media di quotidiani, mensili e settimanali	Indice Sport e Bambini	Indice di Sportività	Librerie	Offerta culturale	Palestre, piscine, centri per il benessere e stabilimenti termali	Partecipazione elettorale	Penetrazione banda larga	Ristoranti (incluse le attività di ristorazione mobile)	Verde urbano fruibile
Amministratori comunali con meno di 40 anni															
Banda larga (Fttc >100mbps)															
Bar															
Densità e rilevanza del patrimonio museale															
Diffusione delle aziende agrituristiche															
Diffusione media di quotidiani, mensili e settimanali															
Indice Sport e Bambini								0.759166							
Indice di Sportività							0.759166								
Librerie															
Offerta culturale															
Palestre, piscine, centri per il benessere e stabilimenti termali															
Partecipazione elettorale															
Penetrazione banda larga															
Ristoranti (incluse le attività di ristorazione mobile)															
Verde urbano fruibile															

Only one correlation observed and something expected like the **children sport index** and **sport index**.

Demography and Society

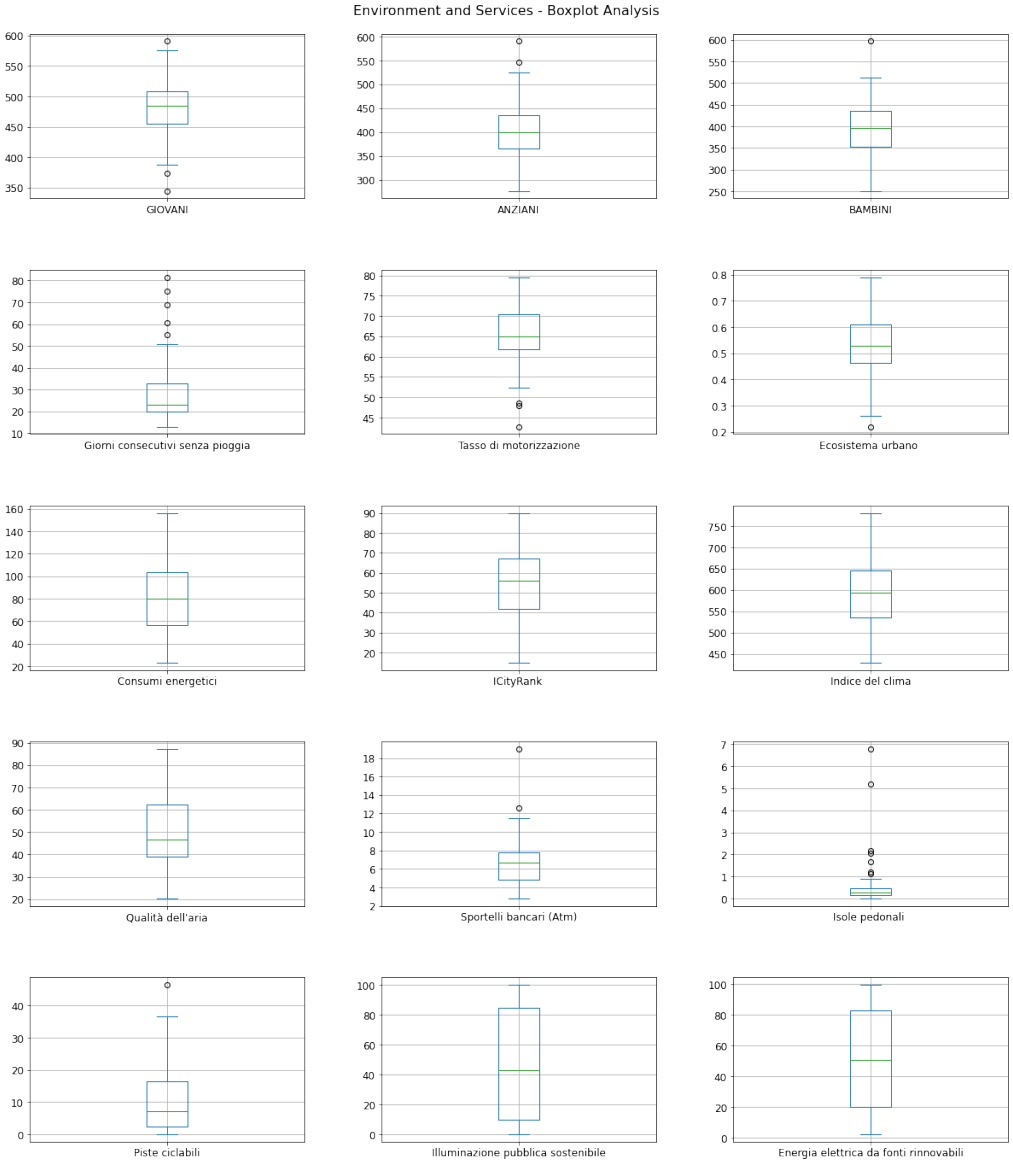


This is one of the most interesting group of indicators, it is very wide the margin between "well educated" and not provinces, this can be seen in **people with at least a high school diploma**, **graduates** and **average number of years of study**. Instead the **life expectancy at birth** is pretty similar to all the provinces, a well known result but always good to observe. Another interesting indicator is the **total migration balance**, It represents very well the migration from south to north of the younger portion of the population, with some region having positive balance and other negative.

	Casi Covid-19	Densità abitativa	Età media al parto	Immigrati regolari residenti	Indice di dipendenza strutturale	Iscritti all'Aire	Laureati e altri titoli terziari	Medici di medicina generale	Medici specialisti	Numero medio di anni di studio della popolazione	Persone con almeno il diploma	Qualità della vita delle donne	Quoziente di natalità	Saldo migratorio totale	Speranza di vita alla nascita
Casi Covid-19															
Densità abitativa															
Età media al parto															
Immigrati regolari residenti															
Indice di dipendenza strutturale															
Iscritti all'Aire															
Laureati e altri titoli terziari										0.820367	0.799255				
Medici di medicina generale															
Medici specialisti															
Numero medio di anni di studio della popolazione							0.820367				0.94109				
Persone con almeno il diploma							0.799255			0.94109					
Qualità della vita delle donne															0.801416
Quoziente di natalità															
Saldo migratorio totale															
Speranza di vita alla nascita												0.801416			

Not considering the obvious positive correlations linked to schooling, it is interesting to observe the positive correlation between **life expectancy at birth** and the **quality of life of women**.

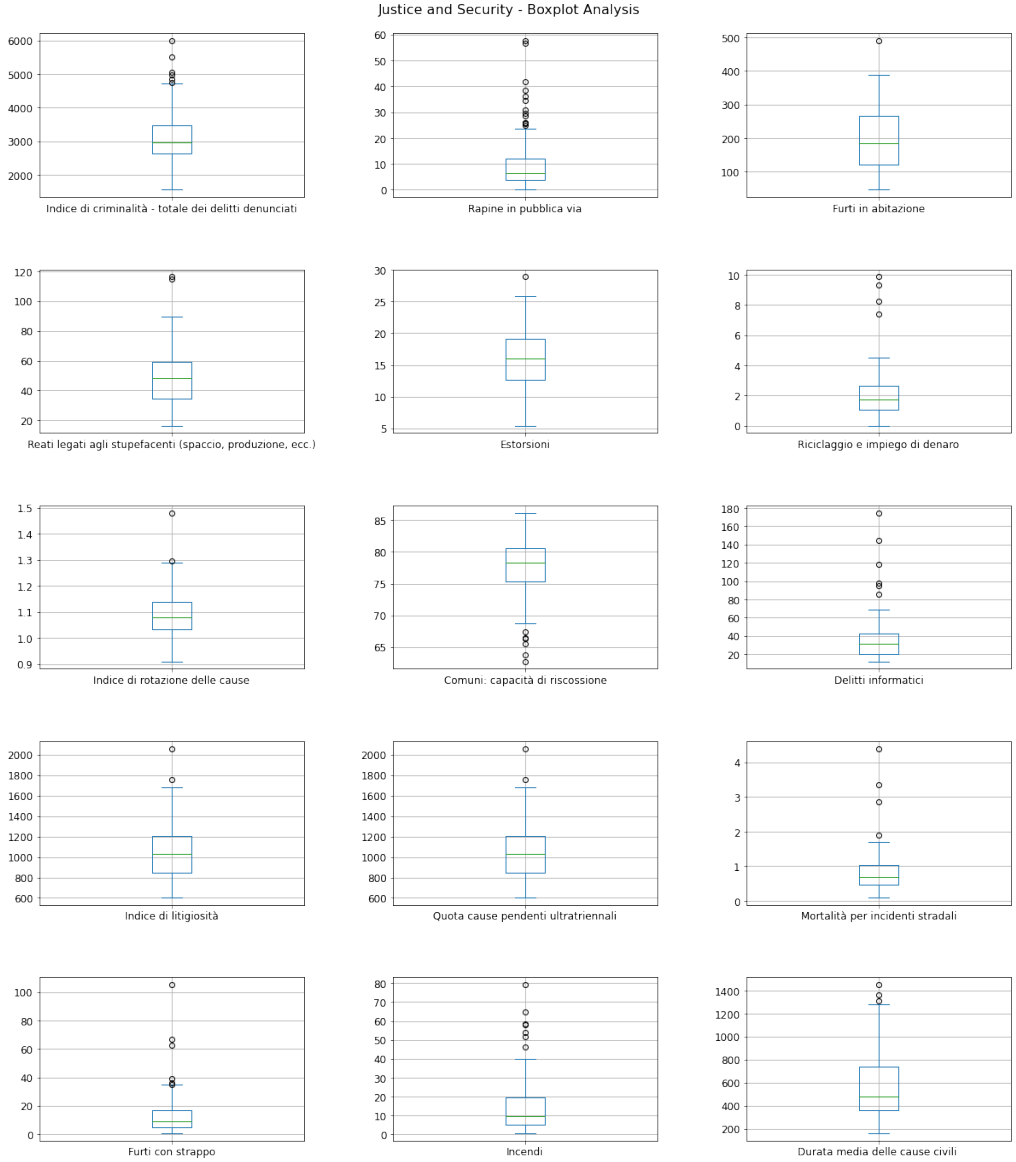
Environment and Services



Air quality and **energy consumption** are linked and vary widely between provinces, in the **consecutive days without rain** indicator can be appreciated the diversity of climate in the peninsula.

	ANZIANI	BAMBINI	Consumi energetici	Ecosistema urbano	Energia elettrica da fonti rinnovabili	GIOVANI	Giorni consecutivi senza pioggia	ICityRank	Illuminazione pubblica sostenibile	Indice del clima	Isole pedonali	Piste ciclabili	Qualità dell'aria	Sportelli bancari (Atm)	Tasso di motorizzazione
ANZIANI															
BAMBINI			0.518199											0.633719	
Consumi energetici		0.518199					-0.615041			-0.547155		0.517648	0.568335	0.622134	
Ecosistema urbano														0.506426	
Energia elettrica da fonti rinnovabili															
GIOVANI															
Giorni consecutivi senza pioggia			-0.615041											-0.539291	
ICityRank													0.507639		-0.531083
Illuminazione pubblica sostenibile															
Indice del clima			-0.547155										-0.533701		
Isole pedonali															
Piste ciclabili			0.517648												
Qualità dell'aria			0.568335					0.507639		-0.533701					
Sportelli bancari (Atm)	0.633719	0.622134	0.506426				-0.539291								
Tasso di motorizzazione								-0.531083							

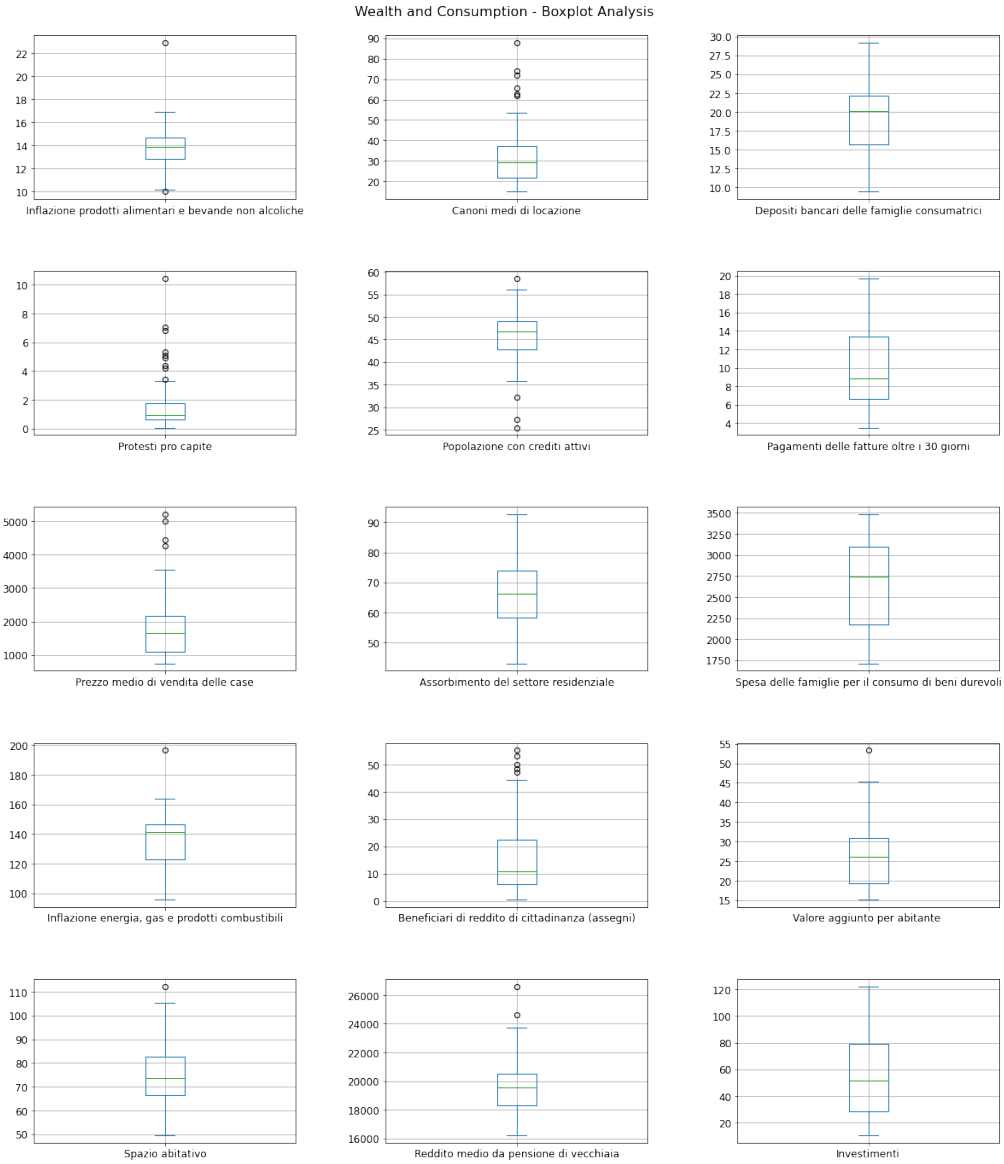
Justice and Security



It is impressive the quantity of outliers in almost the whole set of indicators in this group. This could be related to the fact that the big cities are inherently more dangerous and prone to crime.

	Comuni: capacità di riscossione	Delitti informatici	Durata media delle cause civili	Estorsioni	Furti con strappo	Furti in abitazione	Incendi	Indice di criminalità - totale dei delitti denunciati	Indice di litigiosità	Indice di rotazione delle cause	Mortalità per incidenti stradali	Quota cause pendenti ultratriennali	Rapine in pubblica via	Reati legati agli stupefacenti (spaccio, produzione, ecc.)	Riciclaggio e impiego di denaro
Comuni: capacità di riscossione															
Delitti informatici															
Durata media delle cause civili															
Estorsioni															
Furti con strappo								0.784144					0.849468		
Furti in abitazione															
Incendi															
Indice di criminalità - totale dei delitti denunciati					0.784144								0.790196		
Indice di litigiosità															
Indice di rotazione delle cause															
Mortalità per incidenti stradali															
Quota cause pendenti ultratriennali															
Rapine in pubblica via					0.849468			0.790196							
Reati legati agli stupefacenti (spaccio, produzione, ecc.)															
Riciclaggio e impiego di denaro															

Wealth and Consumption



In the **average selling price of houses** can be seen the big city as outliers, as well as **average rents**. It is remarkable the margin and the outliers in the **citizens' income beneficiaries**, some province's citizens do not use at all this benefits, others in a very huge manner.

	Assorbimento del settore residenziale	Beneficiari di reddito di cittadinanza (assegni)	Canoni medi di locazione	Depositi bancari delle famiglie consumatrici	Inflazione energia, gas e prodotti combustibili	Inflazione prodotti alimentari e bevande non alcoliche	Investimenti	Pagamenti delle fatture oltre i 30 giorni	Popolazione con crediti attivi	Prezzo medio di vendita delle case	Protesti pro capite	Reddito medio da pensione di vecchiaia	Spazio abitativo	Spesa delle famiglie per il consumo di beni durevoli	Valore aggiunto per abitante
Assorbimento del settore residenziale															
Beneficiari di reddito di cittadinanza (assegni)				-0.82385			-0.744405	0.885858						-0.844191	-0.743068
Canoni medi di locazione										0.932062					
Depositi bancari delle famiglie consumatrici		-0.82385					0.794349	-0.823832						0.797283	0.84378
Inflazione energia, gas e prodotti combustibili															
Inflazione prodotti alimentari e bevande non alcoliche															
Investimenti		-0.744405		0.794349				-0.845583						0.796291	0.764155
Pagamenti delle fatture oltre i 30 giorni		0.885858		-0.823832			-0.845583							-0.839611	-0.764684
Popolazione con crediti attivi															
Prezzo medio di vendita delle case			0.932062												0.729203
Protesti pro capite															
Reddito medio da pensione di vecchiaia															
Spazio abitativo															
Spesa delle famiglie per il consumo di beni durevoli		-0.844191		0.797283			0.796291	-0.839611							0.802524
Valore aggiunto per abitante		-0.743068		0.84378			0.764155	-0.764684		0.729203				0.802524	

The negative correlation between **citizens' income beneficiaries** and **bank deposit**, **investments**, **added value per inhabitant** are understandable. If a lot of people need state help, this means that there is no work and opportunity and investments.

Conclusion

The indicators vary widely, with outlier values. There can be the possibility to extract meaningful observations after clustering, now we normalize data and reduce the number of features describing each group of indicators.

Chapter 3

Data Preparation

The preprocessing stage of **Data Cleaning** and **Data Integration** are not needed since there is no missing values or noisy data, this is due to the fact that the Dataset was made available ready to use.

3.1 Normalization and Feature Reduction

The Normalization used is the **Z-score**:

$$z = \frac{x - \mu}{\sigma}$$

With μ the means and σ the standard deviation.
The **Feature Reduction** is applied using the **PCA** technique.
The choice of the parameter *n_components* = 0.95 in the sklearn implementation of PCA set that are selected the number of principal components to get 95% of variance explained.

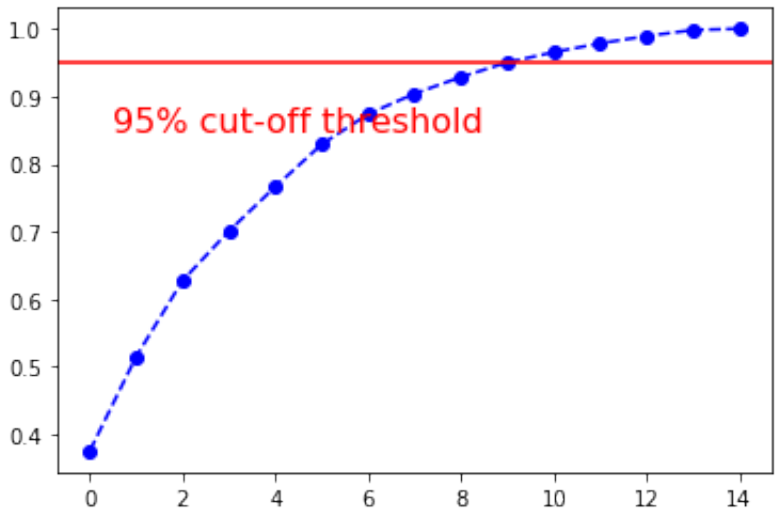


Figure 3.1: Variance Graph of Demography and Society per number of components

As we can see here the number of components selected are 10. The number of components for each group of indicators after PCA is:

- **Wealth and Consumption** : from 15 to 9
- **Business and Work** : from 15 to 11
- **Justice and Security** : from 15 to 11
- **Demography and Society** : from 15 to 10
- **Environment and Service** : from 15 to 12
- **Culture and Leisure** : from 15 to 11

3.2 Cluster Tendency

Clustering tendency assessment determines whether a given data set has a non-random structure, which may lead to meaningful clusters. One of the most used metrics is the **Hopkins Statistic** that is the one that It is used in this work. It tests the spatial randomness of a variable as distributed in a space.

$$H = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i}$$

With y_i points from a random uniform distribution and x_i points sampled from the Dataset. Furthermore to alleviate the inherent randomness of this metric, the hopkins statistics is computed 50 times and is taken as result the mean. The results of this assessment for each group of indicators :

- **Wealth and Consumption** : H = 0.69
- **Business and Work** : H = 0.7
- **Justice and Security** : H = 0.7
- **Demography and Society** : H = 0.698
- **Environment and Service** : H = 0.69
- **Culture and Leisure** : H = 0.6825

The values are closed to 0.7 so there could be a good confidence level of cluster tendency.

Chapter 4

Modeling

In this Chapter there will be presented some clustering approach based on different ideas like **Partitioning**, **Hierarchical** based and **Density** based techniques. Here the analysis will be done using the 2022 Dataset.
The visualization, when considered useful to highlight some interesting aspects will be displayed using a 2D Scatter Plot, generated reducing the number of features to 2 with PCA.

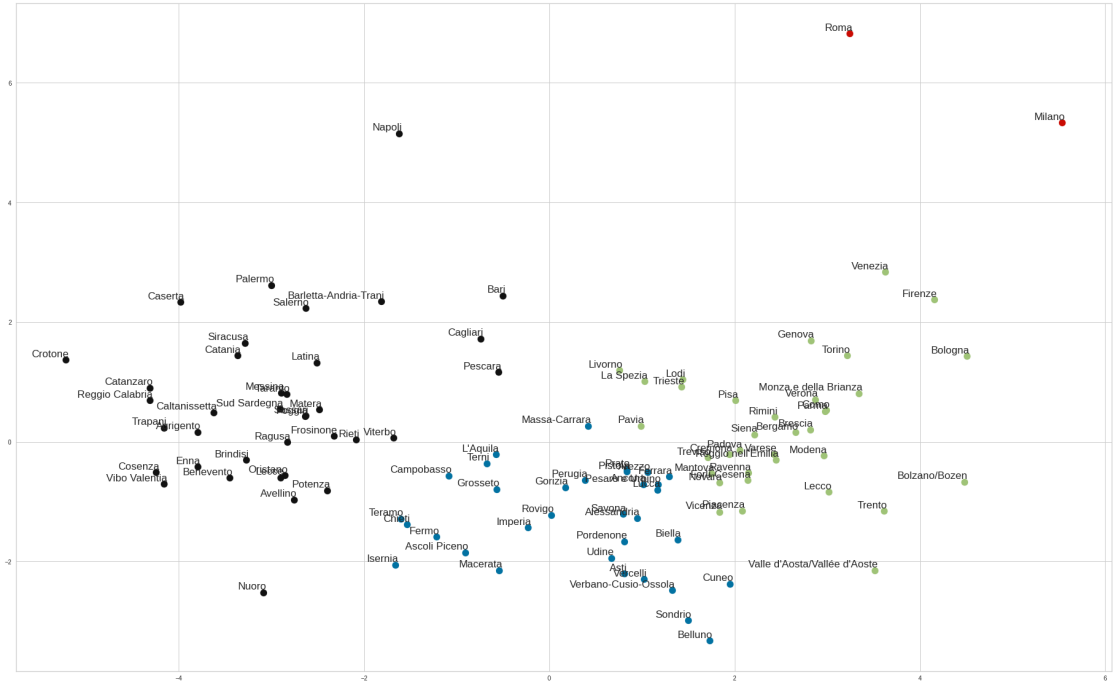


Figure 4.1: 2D Scatter Plot of Wealth and Consumpiton clustered using K-Means

4.1 Partitioning

The main idea behind this group of techniques is to find a specified number of partition usually denoted as k, such that this partitioning of the original Dataset D optimizes the partition criterion. Since It is impossible to enumerate all the possible partitions of the dataset there are

some heuristic approach like the k-means algorithm or the k-medoid algorithm.
The first problem that emerge when dealing with Partitioning methods is the choice of the number of clusters parameter.

4.1.1 Choosing Number of Clusters

Two methods were considered in order to find this parameter:

Elbow Method

The **Elbow method** is a very popular technique and the idea is to run k-means clustering for a range of clusters k (for example 1 to 10) and for each value, calculate the sum of squared distances from each point to its assigned center named distortion. After calculating these distortions for each *k*, It is selected the number of clusters in the "elbow" of the curve, when the distortion value starts to decrease less rapidly.

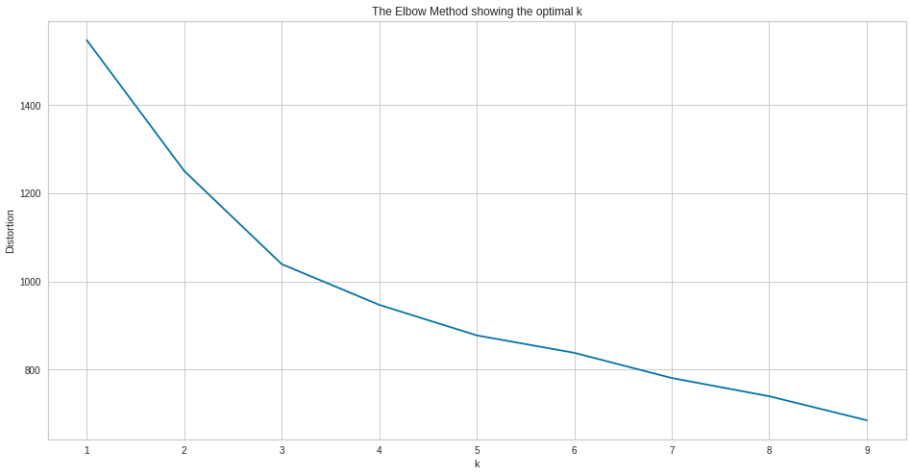


Figure 4.2: Elbow Method for Justice and Security group

Unfortunately this analysis showed that in almost all groups of indicators, the curve does not present the expected "elbow", therefore not giving particular advice on how to choose the number of clusters.

Silhouette Analysis

Another approach to find the right number of clusters is the **Silhouette Analysis** that refers to a method of interpretation and validation of consistency within clusters of data. The silhouette value is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation). The silhouette ranges from -1 to +1, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters. Computing this metrics for a range of *k* clusters, we could choose the number with the maximum Silhouette Score. Unfortunately also in this case this technique does not give us great help, since the silhouette score decreases with the increase of the clusters, it was therefore decided to take *4 clusters* as a good tradeoff value, trying to avoid too specific clusters.

4.1.2 K-Means

Approaching the problem using **K-Means** the results are these:

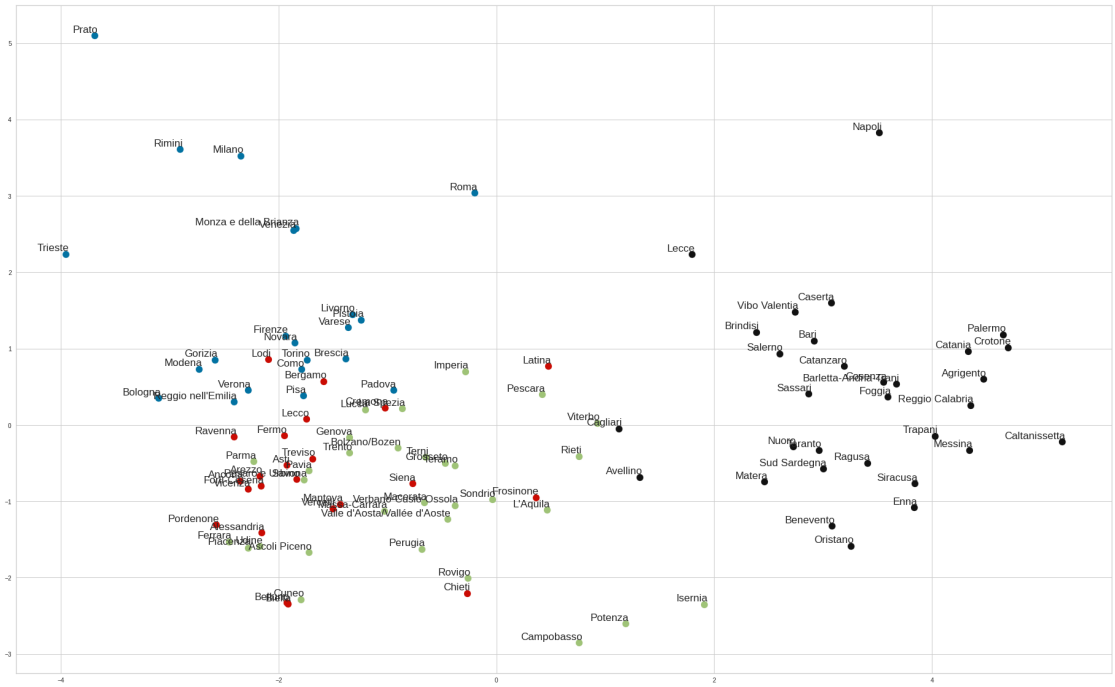


Figure 4.3: Business and Work with K-Means

Here can be seen the *black* cluster containing mainly the provinces of southern Italy that present similar behavior in term of Business and Work. They are well separated and clearly distinguishable. In *blue* the provinces that are more productive in term of Business and Work like Bologna, Roma, Milano, Firenze and so on. It is interesting to observe the *green* cluster that contains provinces with different background like the little Isernia and Genova, this is perhaps the more heterogeneous cluster.

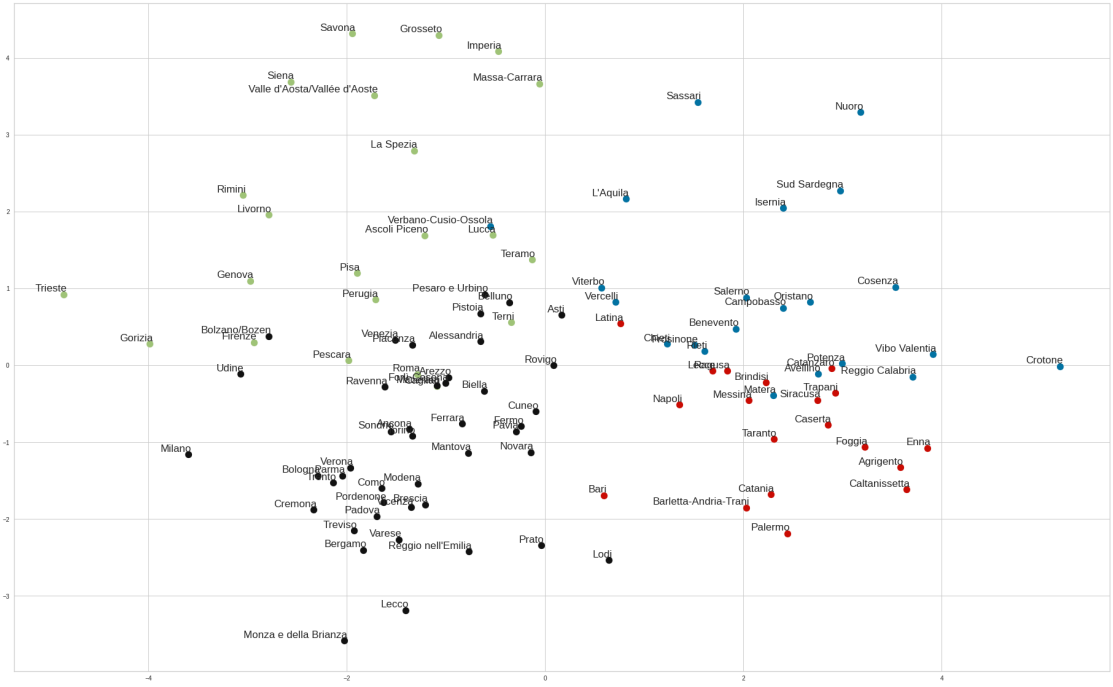


Figure 4.4: Culture and Leisure with K-Means

Demography and Society clusters are well separated, It is interesting to observe such differences in tuscanian provinces like Livorno and Lucca that are clustered not in the same cluster of Pisa, Siena and Firenze.

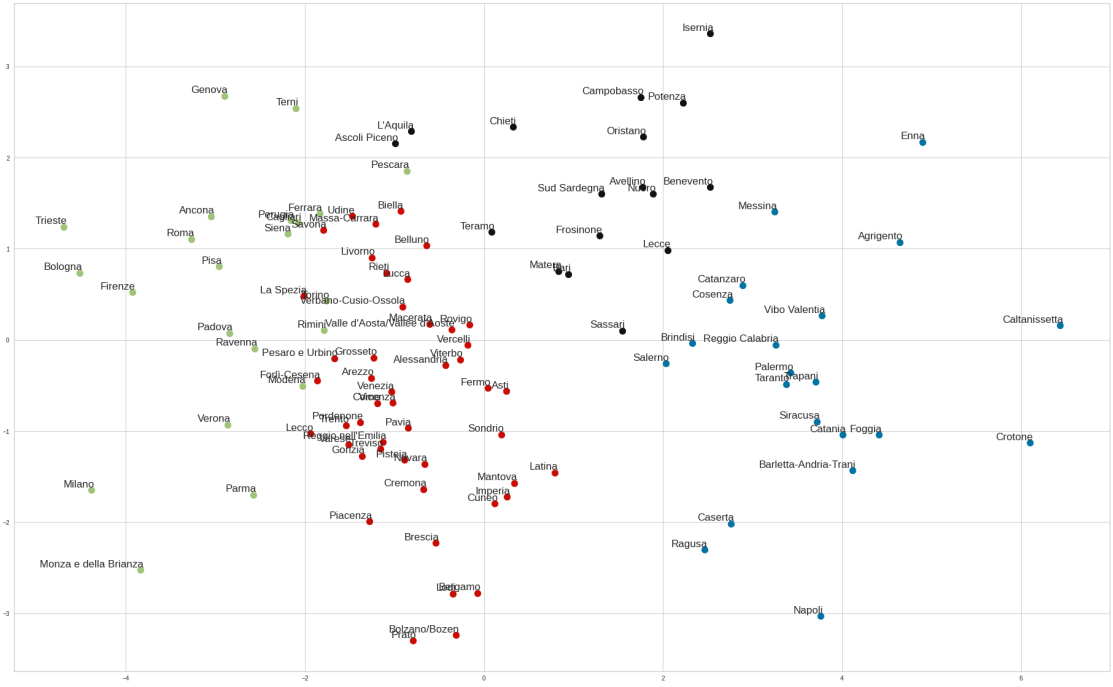


Figure 4.5: Demography and Society with K-Means

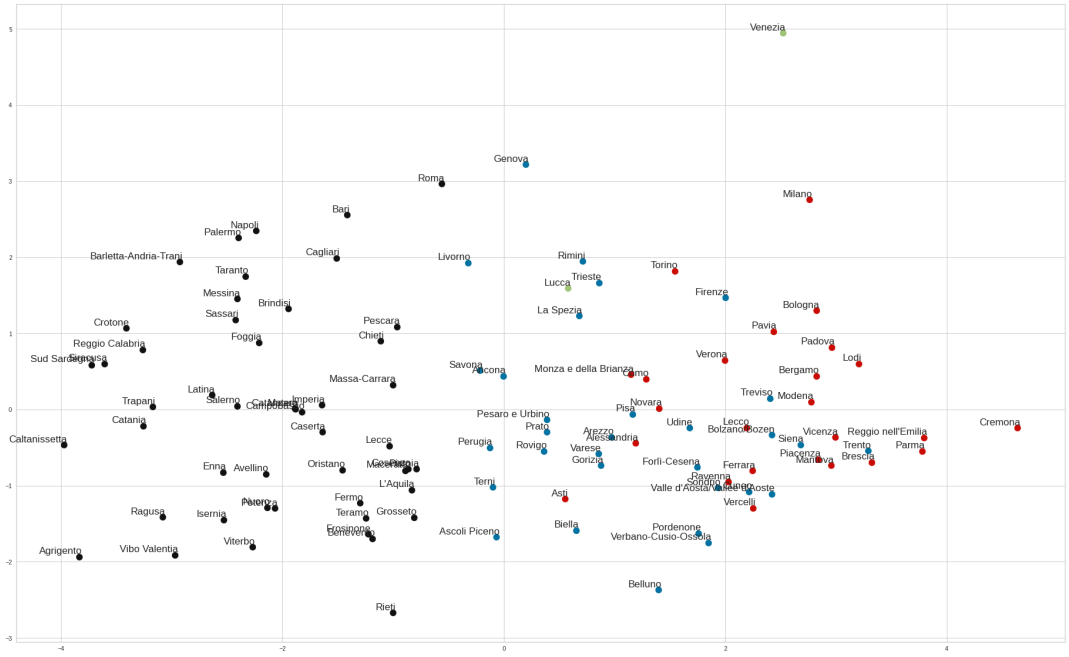


Figure 4.6: Environment and Services with K-Means

In Justice and Security clustering, the *red* cluster contain Milano, Napoli, Torino, Bologna, Palermo and Roma, this is expected since they are the biggest cities in Italy with an high rate of criminality. But the presence in the same cluster of Rimini and Prato, suggest that even if they are less populous cities, their crime rate rivals the larger ones. Despite its size, Firenze is not grouped in the red cluster, demonstrating virtuous behavior.

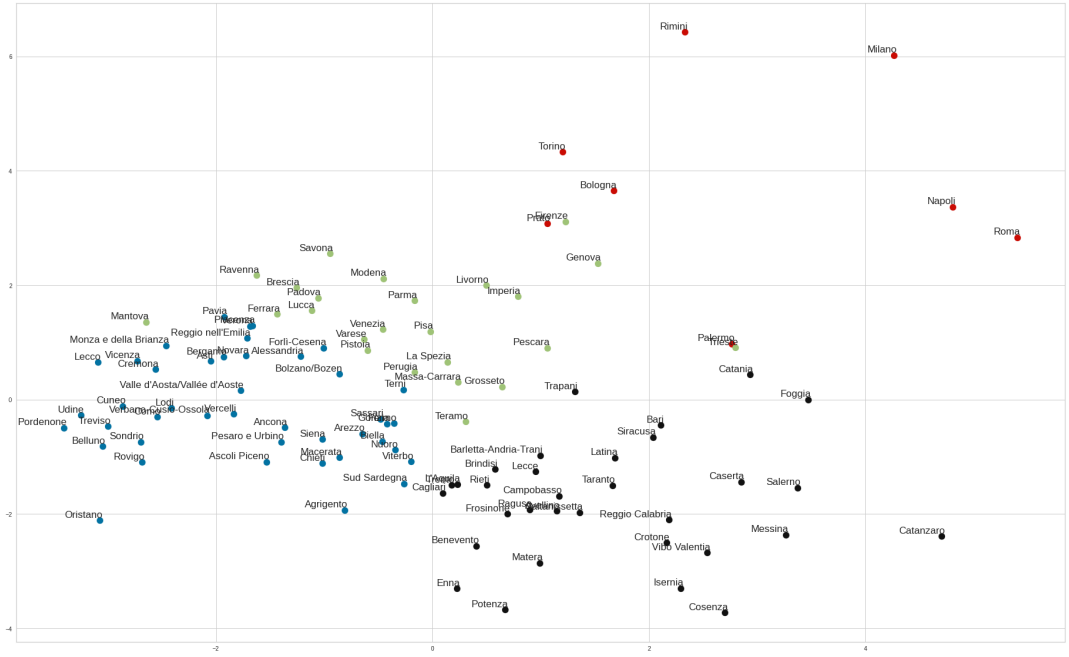


Figure 4.7: Justice and Security with K-Means

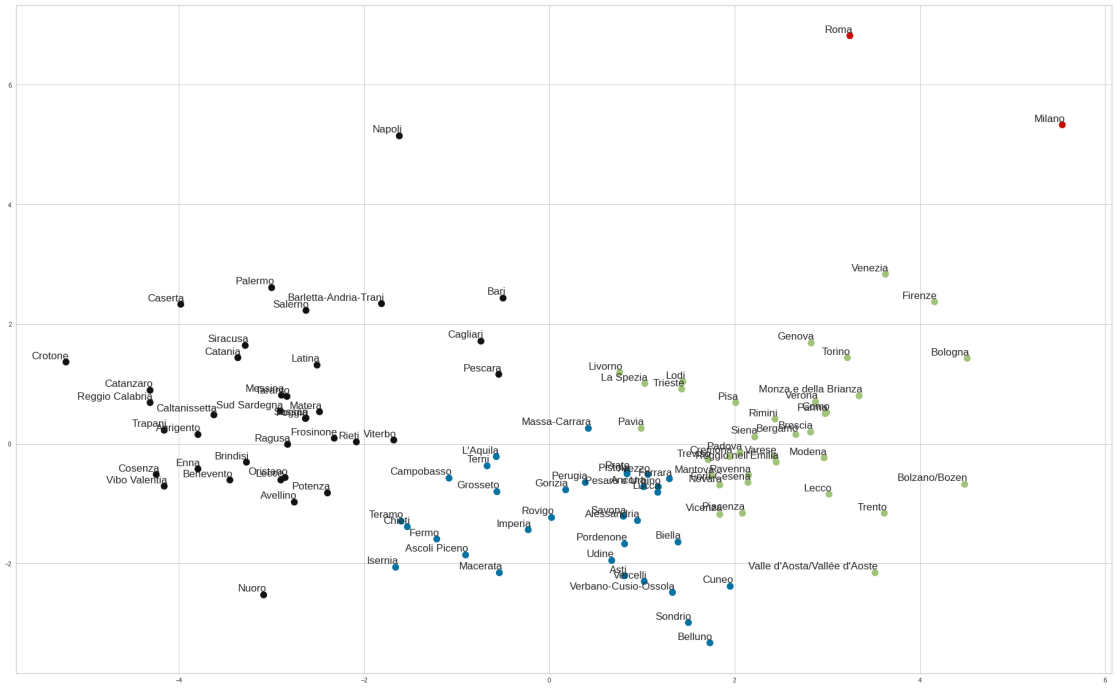


Figure 4.8: Wealth and Consumption with K-Means

It is clear how the economic and political italian capitals Roma and Milano provide a substantial contribution to Italy's wealth and consumption, being clustered alone and distant from the others. The *green* cluster contains small or middle town with good consumption rate.

4.1.3 CLARANS

This partitioning method is an implementation of the idea K-Medoids, a partioning technique.

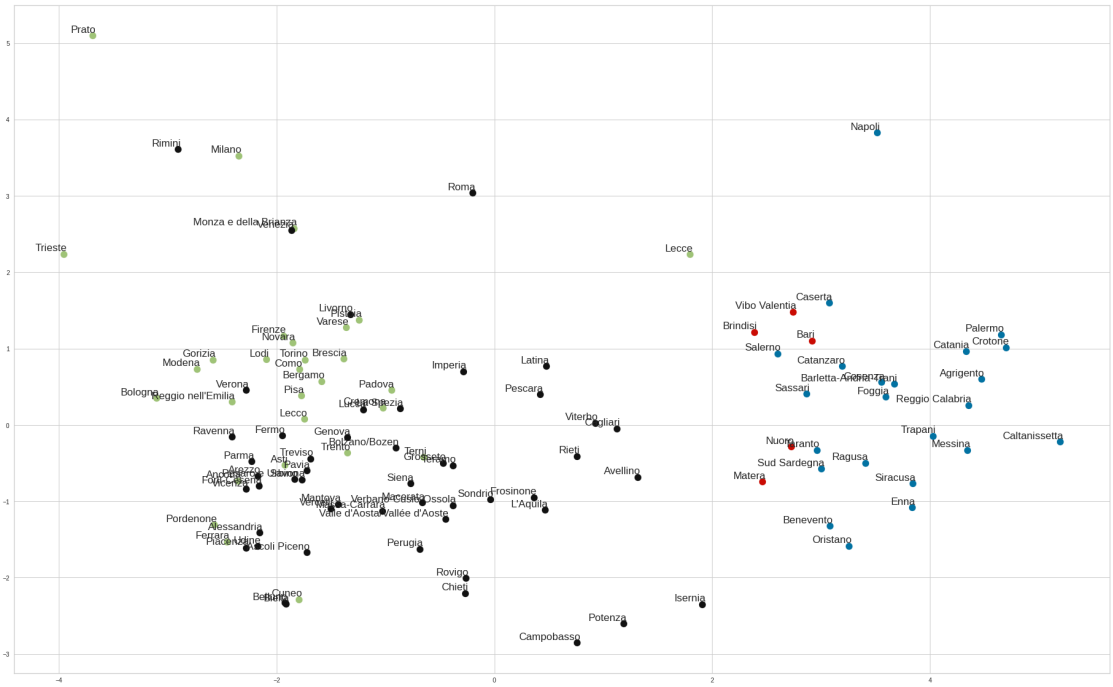


Figure 4.9: Business and Work with CLARANS

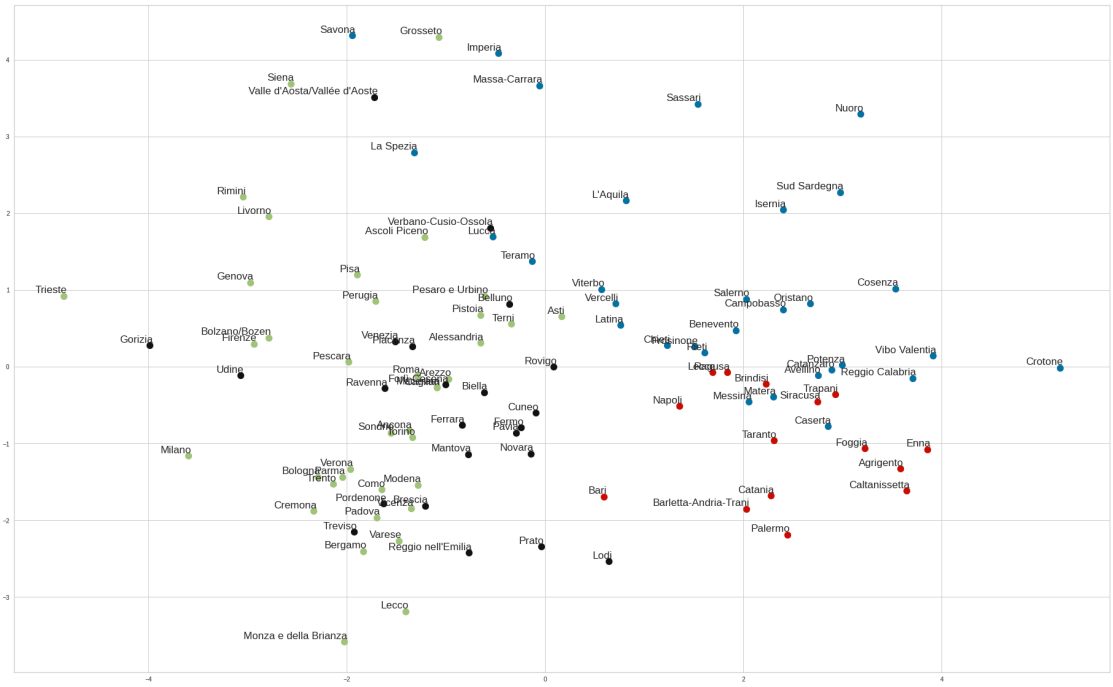


Figure 4.10: Culture and Leisure with CLARANS

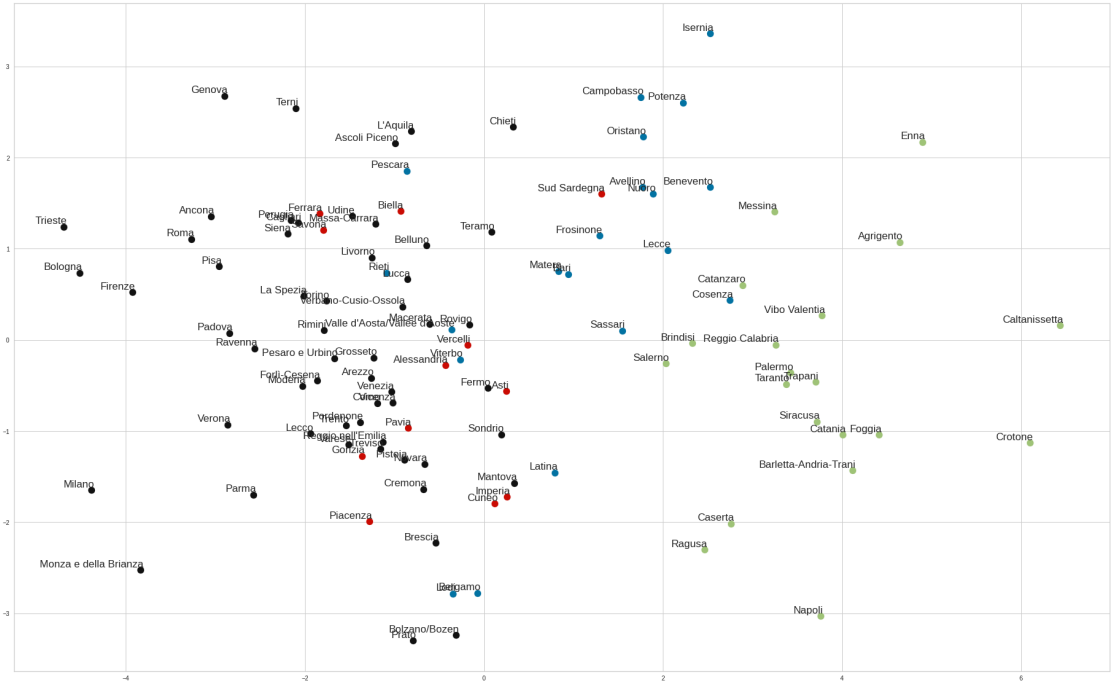


Figure 4.11: Demography and Society with CLARANS

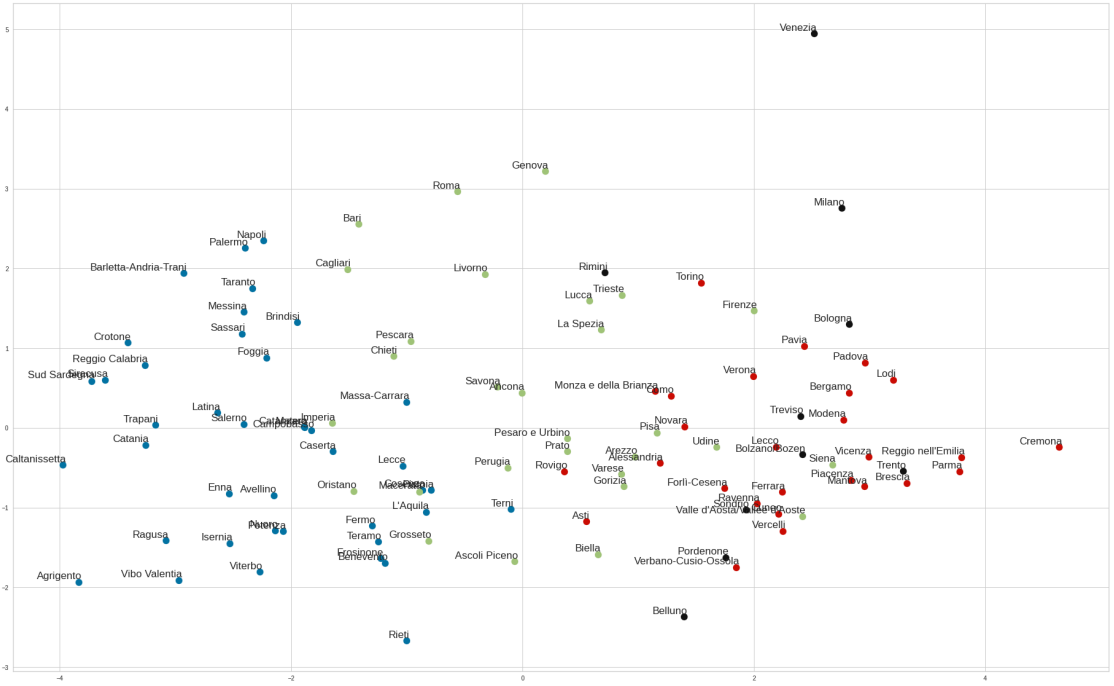


Figure 4.12: Environment and Services with CLARANS

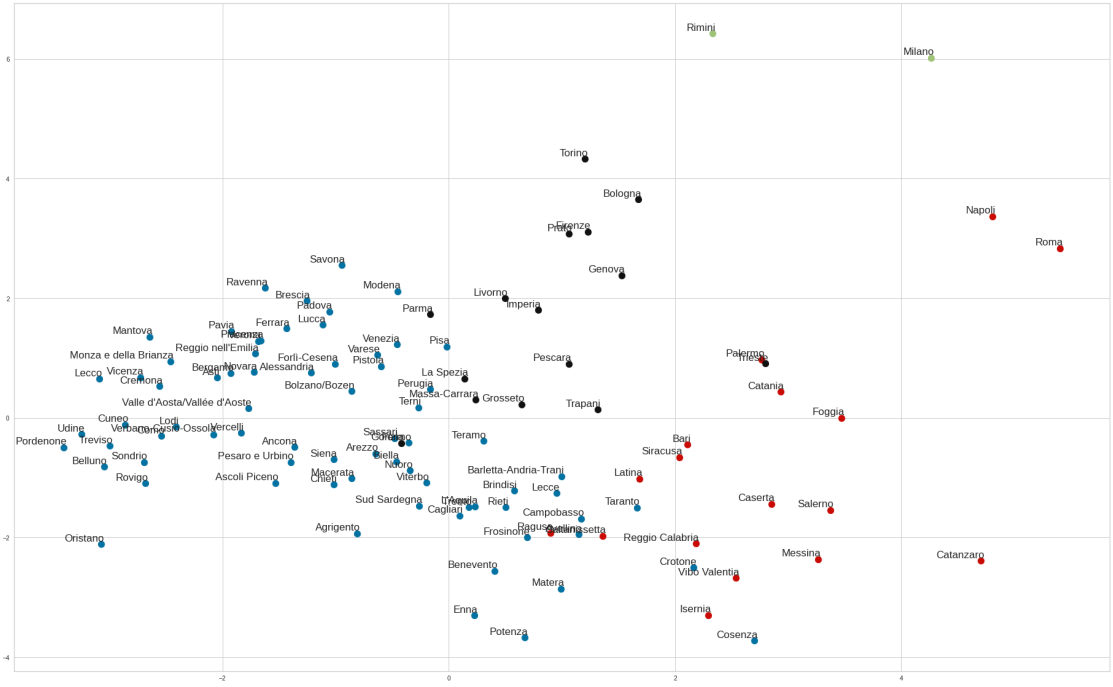


Figure 4.13: Justice and Security with CLARANS

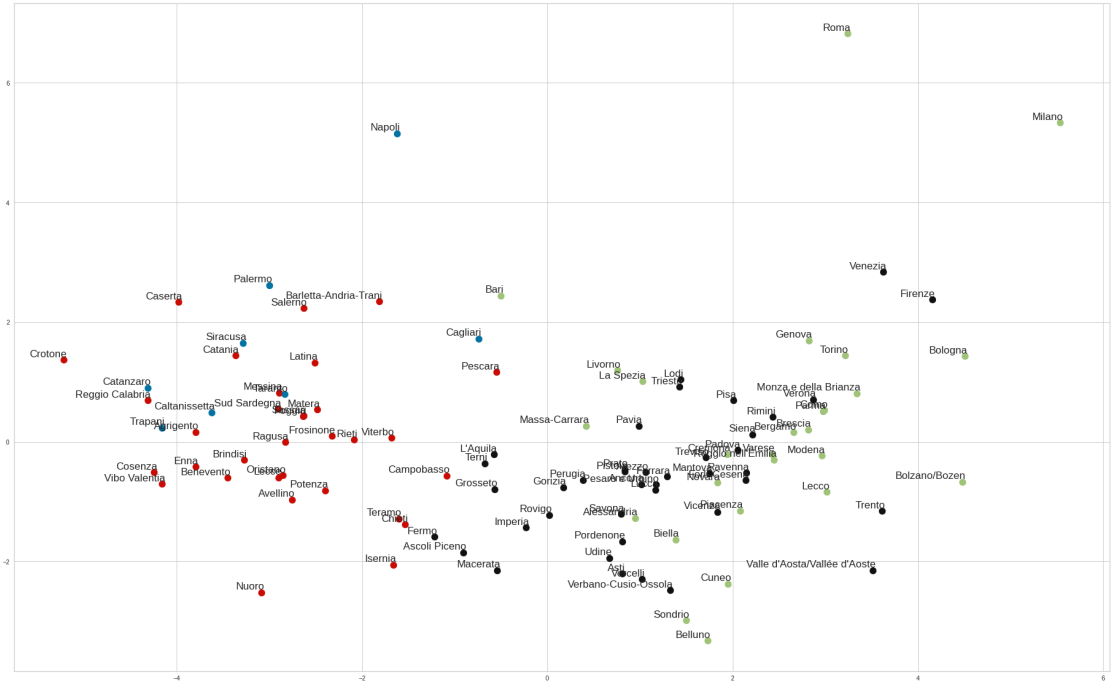


Figure 4.14: Wealth and Consumption with CLARANS

4.2 Hierarchical

Another family of Clustering methods is the **Hierarchical** based methods. The main idea is to create a hierarchical decomposition of the set of data using some criterion.

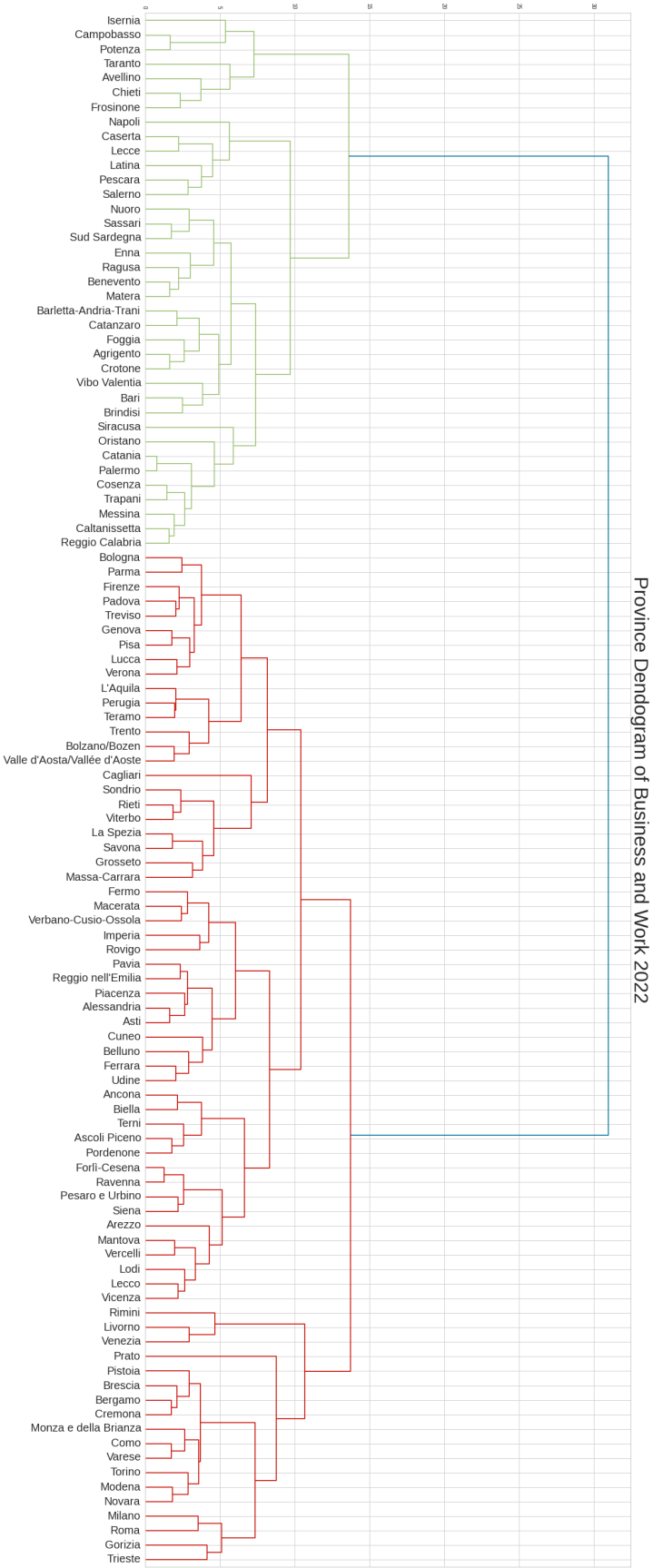
4.2.1 Agglomerative

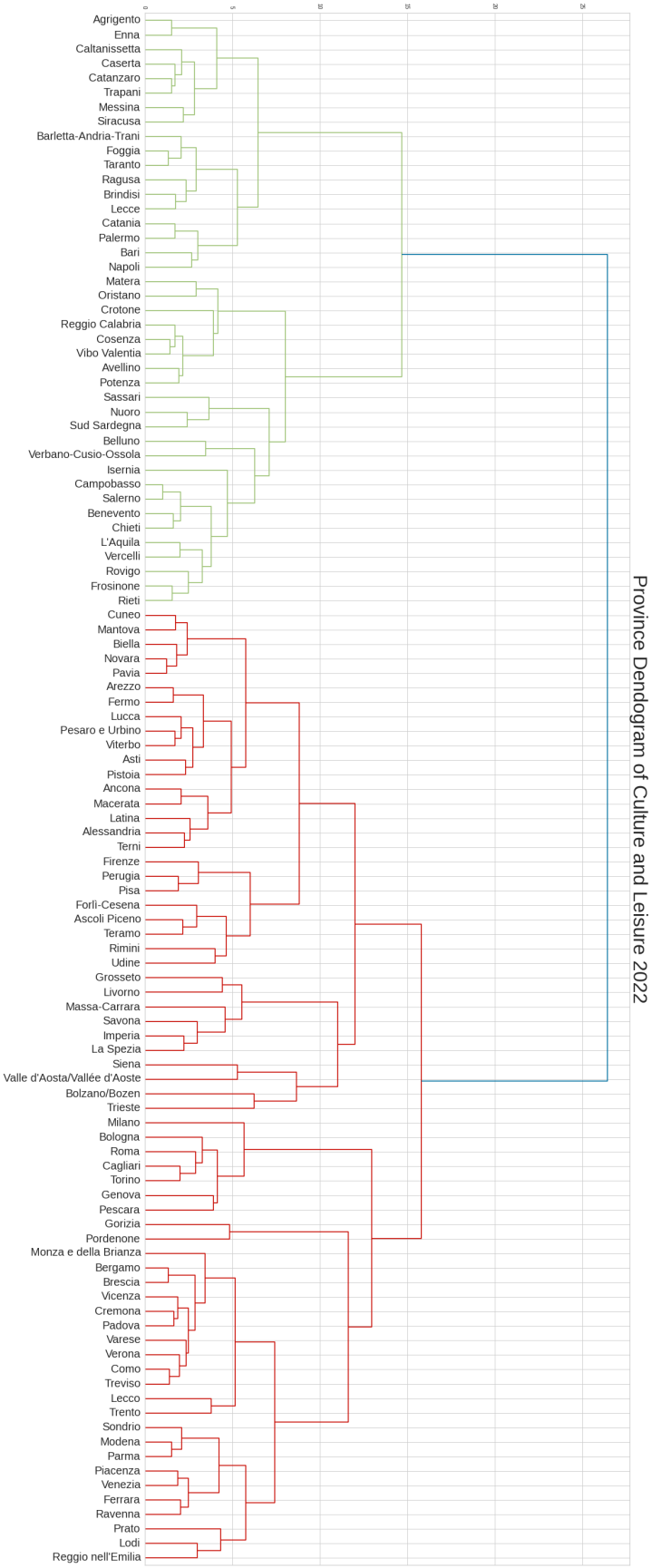
In this approach we start with each sample in its cluster and over the algorithm, clusters are merged using the so called *linkage metric*. The linkage metric determines which distance to use between sets of samples. The algorithm will merge the pairs of cluster that minimize this criterion:

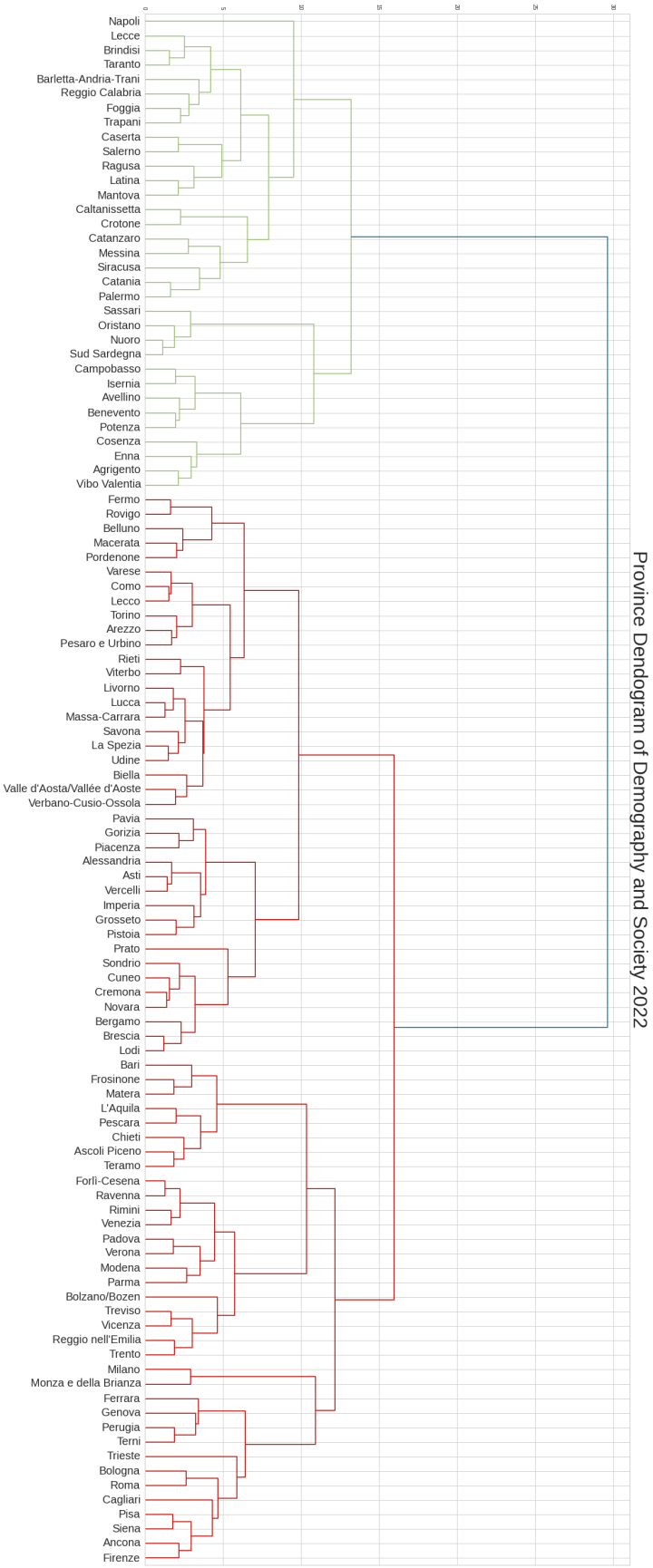
- **Ward** : minimizes the variance of the clusters being merged.
- **Average** : uses the average of the distances of each observation of the two sets.
- **Single** : uses the minimum of the distances between all observations of the two sets.
- **Complete** : uses the maximum distances between all observations of the two sets.

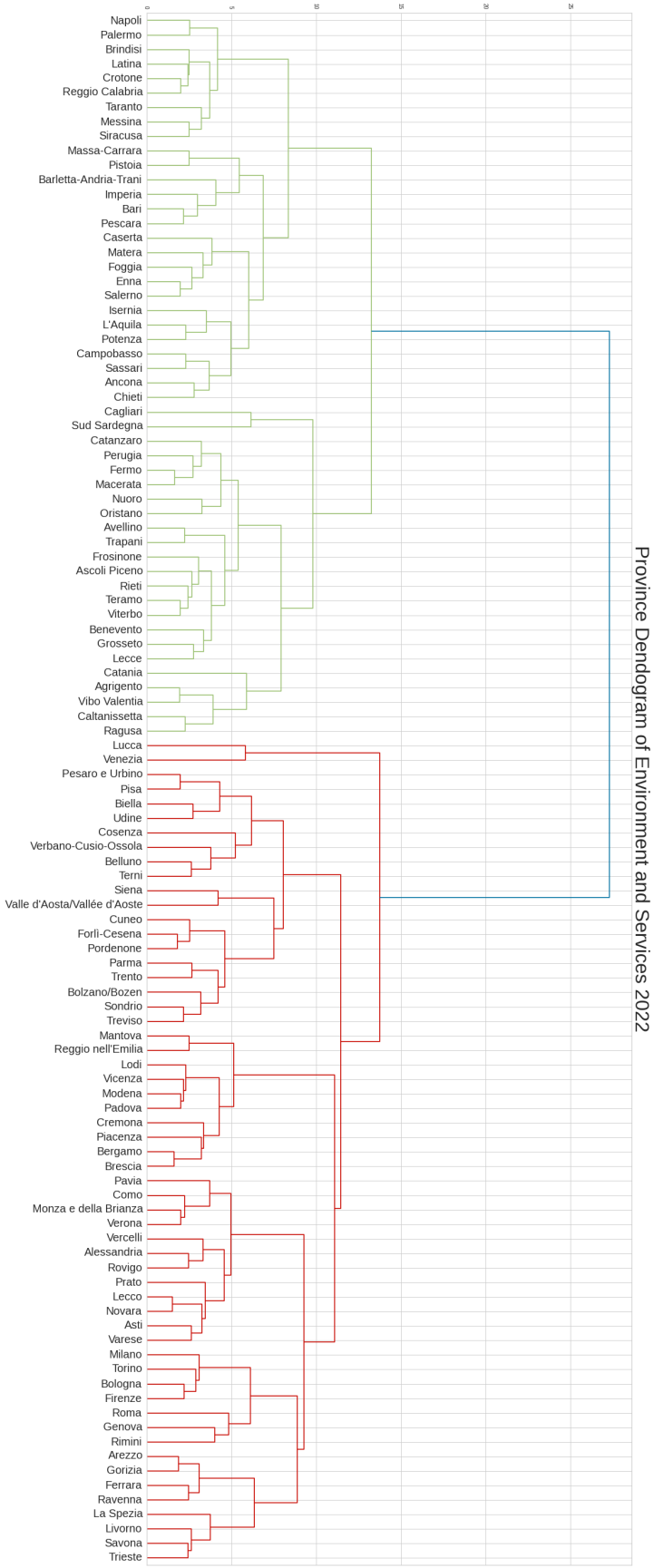
After testing the whole set of linkage, It is chosen the *ward*, that gives the better results. Here some comments on the following dendograms:

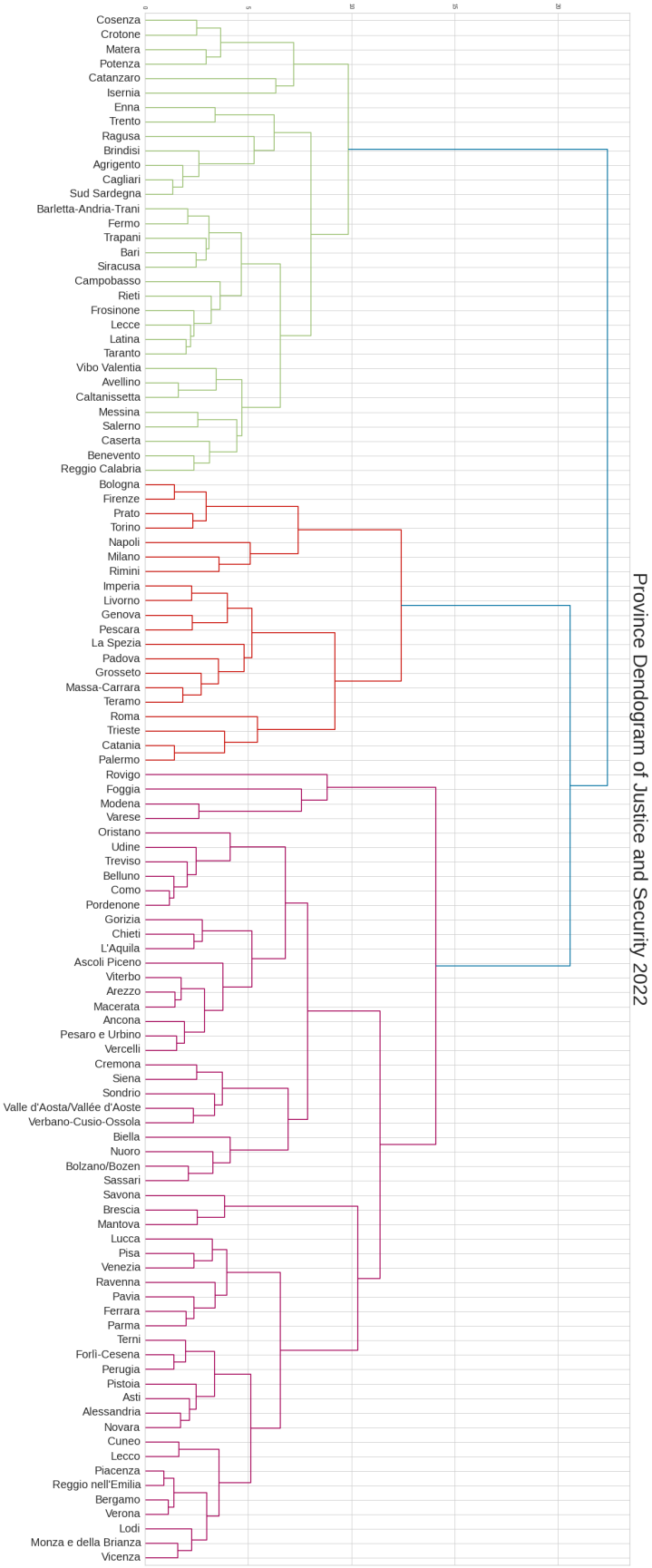
- **Business and Work** : interstingly Livorno, Rimini and Venezia are group together in th every first stage, they are city on the seas with Livorno and Venezia having also an imporant commercial port.
- **Justice and Security** : a curious coupling is Bolzano-Sassari, it should be investigated.
- **Environment and Service** : Trieste, Savona, Livorno and LaSpezia that are port cities are grouped together sharing similar environment and services characteristics.
- **Culture and Leisure** : Milano, Bologna, Roma, Genova and Torino grouped together, they are important cultural centers.

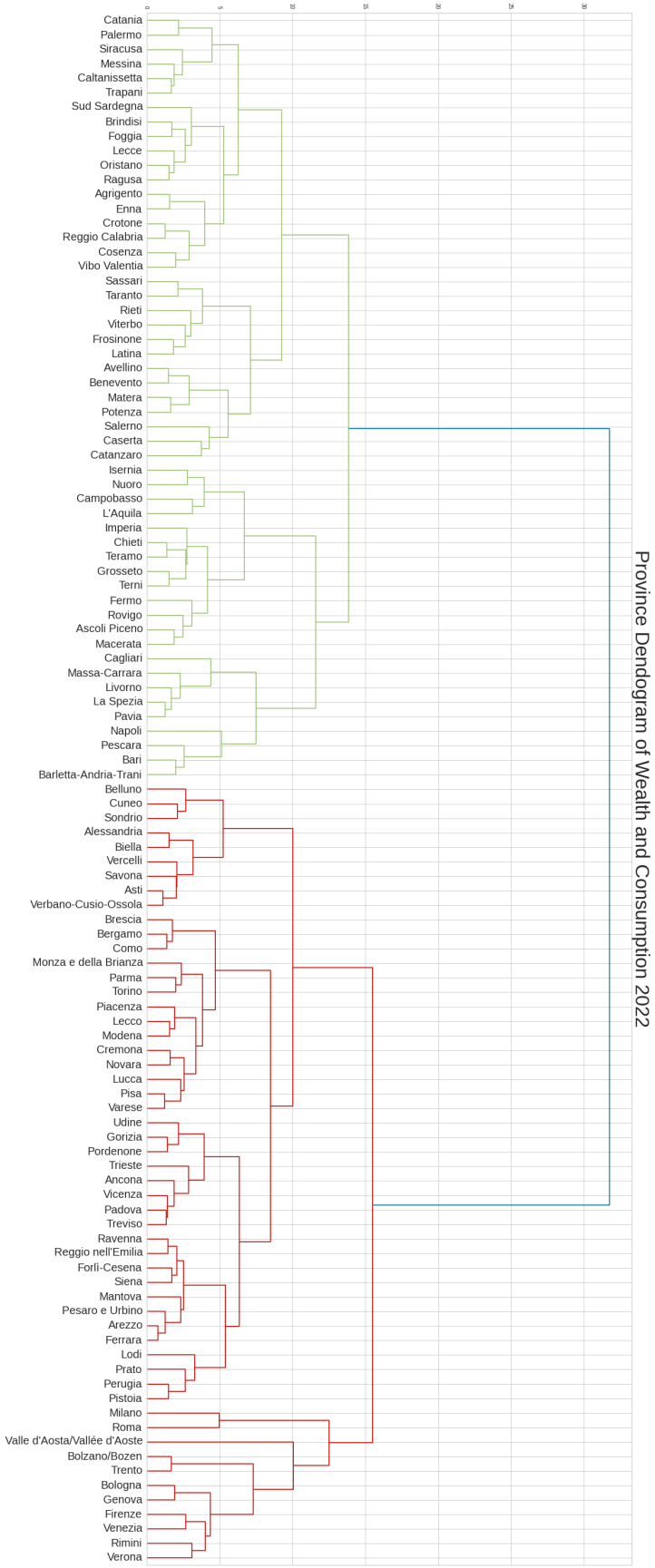

Province Dendogram of Business and Work 2022
Isernia
Campobasso
Potenza
Taranto
Avellino
Chieti
Frosinone
Napoli
Caserta
Lecce
Latina
Pescara
Salerno
Nuoro
Sassari
Sud Sardegna
Enna
Ragusa
Benevento
Matera
Barletta-Andria-Trani
Catanzaro
Foggia
Agrigento
Crotone
Vibo Valentia
Bari
Brindisi
Siracusa
Oristano
Catania
Palermo
Cosenza
Trapani
Messina
Caltanissetta
Reggio Calabria
Bologna
Parma
Firenze
Padova
Treviso
Genova
Pisa
Lucca
Verona
L'Aquila
Perugia
Teramo
Trento
Bolzano/Bozen
Valle d'Aosta/Vallée d'Aoste
Cagliari
Sondrio
Rieti
Viterbo
La Spezia
Savona
Grosseto
Massa-Carrara
Fermo
Macerata
Verbano-Cusio-Ossola
Imperia
Rovigo
Pavia
Reggio nell'Emilia
Piacenza
Alessandria
Asti
Cuneo
Belluno
Ferrara
Udine
Ancona
Biella
Terni
Ascoli Piceno
Pordenone
Forlì-Cesena
Ravenna
Pesaro e Urbino
Siena
Arezzo
Mantova
Vercelli
Lodi
Lecco
Vicenza
Rimini
Livorno
Venezia
Prato
Pistoia
Brescia
Bergamo
Cremona
Monza e della Brianza
Como
Varese
Torino
Modena
Novara
Milano
Roma
Gorizia
Trieste


0
5
10
15
20
25
Agrigento
Enna
Caltanissetta
Caserta
Catanzaro
Trapani
Messina
Siracusa
Barletta-Andria-Trani
Foggia
Taranto
Ragusa
Brindisi
Lecce
Catania
Palermo
Bari
Napoli
Matera
Oristano
Crotone
Reggio Calabria
Cosenza
Vibo Valentia
Avellino
Potenza
Sassari
Nuoro
Sud Sardegna
Belluno
Verbano-Cusio-Ossola
Isernia
Campobasso
Salerno
Benevento
Chieti
L'Aquila
Vercelli
Rovigo
Frosinone
Rieti
Cuneo
Mantova
Biella
Novara
Pavia
Arezzo
Fermo
Lucca
Pesaro e Urbino
Viterbo
Asti
Pistoia
Ancona
Macerata
Latina
Alessandria
Terni
Firenze
Perugia
Pisa
Forlì-Cesena
Ascoli Piceno
Teramo
Rimini
Udine
Grosseto
Livorno
Massa-Carrara
Savona
Imperia
La Spezia
Siena
Valle d'Aosta/Vallée d'Aoste
Bolzano/Bozen
Trieste
Milano
Bologna
Roma
Cagliari
Torino
Genova
Pescara
Gorizia
Pordenone
Monza e della Brianza
Bergamo
Brescia
Vicenza
Cremona
Padova
Varese
Verona
Como
Treviso
Lecco
Trento
Sondrio
Modena
Parma
Piacenza
Venezia
Ferrara
Ravenna
Prato
Lodi
Reggio nell'Emilia
Province Dendogram of Culture and Leisure 2022


0
5
10
15
20
25
30
Napoli
Lecce
Brindisi
Taranto
Barletta-Andria-Trani
Reggio Calabria
Foggia
Trapani
Caserta
Salerno
Ragusa
Latina
Mantova
Caltanissetta
Crotone
Catanzaro
Messina
Siracusa
Catania
Palermo
Sassari
Oristano
Nuoro
Sud Sardegna
Campobasso
Isernia
Avellino
Benevento
Potenza
Cosenza
Enna
Agrigento
Vibo Valentia
Fermo
Rovigo
Belluno
Macerata
Pordenone
Varese
Como
Lecco
Torino
Arezzo
Pesaro e Urbino
Rieti
Viterbo
Livorno
Lucca
Massa-Carrara
Savona
La Spezia
Udine
Biella
Valle d'Aosta/Vallée d'Aoste
Verbano-Cusio-Ossola
Pavia
Gorizia
Piacenza
Alessandria
Asti
Vercelli
Imperia
Grosseto
Pistoia
Prato
Sondrio
Cuneo
Cremona
Novara
Bergamo
Brescia
Lodi
Bari
Frosinone
Matera
L'Aquila
Pescara
Chieti
Ascoli Piceno
Teramo
Forlì-Cesena
Ravenna
Rimini
Venezia
Padova
Verona
Modena
Parma
Bolzano/Bozen
Treviso
Vicenza
Reggio nell'Emilia
Trento
Milano
Monza e della Brianza
Ferrara
Genova
Perugia
Terni
Trieste
Bologna
Roma
Cagliari
Pisa
Siena
Ancona
Firenze
Province Dendrogram of Demography and Society 2022


Napoli
Palermo
Brindisi
Latina
Crotone
Reggio Calabria
Taranto
Messina
Siracusa
Massa-Carrara
Pistoia
Barletta-Andria-Trani
Imperia
Bari
Pescara
Caserta
Matera
Foggia
Enna
Salerno
Isernia
L'Aquila
Potenza
Campobasso
Sassari
Ancona
Chieti
Cagliari
Sud Sardegna
Catanzaro
Perugia
Fermo
Macerata
Nuoro
Oristano
Avellino
Trapani
Frosinone
Ascoli Piceno
Rieti
Teramo
Viterbo
Benevento
Grosseto
Lecce
Catania
Agrigento
Vibo Valentia
Caltanissetta
Ragusa
Lucca
Venezia
Pesaro e Urbino
Pisa
Biella
Udine
Cosenza
Verbano-Cusio-Ossola
Belluno
Terni
Siena
Valle d'Aosta/Vallée d'Aoste
Cuneo
Forlì-Cesena
Pordenone
Parma
Trento
Bolzano/Bozen
Sondrio
Treviso
Mantova
Reggio nell'Emilia
Lodi
Vicenza
Modena
Padova
Cremona
Piacenza
Bergamo
Brescia
Pavia
Como
Monza e della Brianza
Verona
Vercelli
Alessandria
Rovigo
Prato
Lecco
Novara
Asti
Varese
Milano
Torino
Bologna
Firenze
Roma
Genova
Rimini
Arezzo
Gorizia
Ferrara
Ravenna
La Spezia
Livorno
Savona
Trieste
Province Dendogram of Environment and Services 2022


0
5
10
15
20
Cosenza
Crotone
Matera
Potenza
Catanzaro
Isernia
Enna
Trento
Ragusa
Brindisi
Agrigento
Cagliari
Sud Sardegna
Barletta-Andria-Trani
Fermo
Trapani
Bari
Siracusa
Campobasso
Rieti
Frosinone
Lecce
Latina
Taranto
Vibo Valentia
Avellino
Caltanissetta
Messina
Salerno
Caserta
Benevento
Reggio Calabria
Bologna
Firenze
Prato
Torino
Napoli
Milano
Rimini
Imperia
Livorno
Genova
Pescara
La Spezia
Padova
Grosseto
Massa-Carrara
Teramo
Roma
Trieste
Catania
Palermo
Rovigo
Foggia
Modena
Varese
Oristano
Udine
Treviso
Belluno
Como
Pordenone
Gorizia
Chieti
L'Aquila
Ascoli Piceno
Viterbo
Arezzo
Macerata
Ancona
Pesaro e Urbino
Vercelli
Cremona
Siena
Sondrio
Valle d'Aosta/Vallée d'Aoste
Verbano-Cusio-Ossola
Biella
Nuoro
Bolzano/Bozen
Sassari
Savona
Brescia
Mantova
Lucca
Pisa
Venezia
Ravenna
Pavia
Ferrara
Parma
Terni
Forlì-Cesena
Perugia
Pistoia
Asti
Alessandria
Novara
Cuneo
Lecco
Piacenza
Reggio nell'Emilia
Bergamo
Verona
Lodi
Monza e della Brianza
Vicenza
Province Dendogram of Justice and Security 2022


Province Dendrogram of Wealth and Consumption 2022
Catania
Palermo
Siracusa
Messina
Caltanissetta
Trapani
Sud Sardegna
Brindisi
Foggia
Lecce
Oristano
Ragusa
Agrigento
Enna
Crotone
Reggio Calabria
Cosenza
Vibo Valentia
Sassari
Taranto
Rieti
Viterbo
Frosinone
Latina
Avellino
Benevento
Matera
Potenza
Salerno
Caserta
Catanzaro
Isernia
Nuoro
Campobasso
L'Aquila
Imperia
Chieti
Teramo
Grosseto
Terni
Fermo
Rovigo
Ascoli Piceno
Macerata
Cagliari
Massa-Carrara
Livorno
La Spezia
Pavia
Napoli
Pescara
Bari
Barletta-Andria-Trani
Belluno
Cuneo
Sondrio
Alessandria
Biella
Vercelli
Savona
Asti
Verbano-Cusio-Ossola
Brescia
Bergamo
Como
Monza e della Brianza
Parma
Torino
Piacenza
Lecco
Modena
Cremona
Novara
Lucca
Pisa
Varese
Udine
Gorizia
Pordenone
Trieste
Ancona
Vicenza
Padova
Treviso
Ravenna
Reggio nell'Emilia
Forlì-Cesena
Siena
Mantova
Pesaro e Urbino
Arezzo
Ferrara
Lodi
Prato
Perugia
Pistoia
Milano
Roma
Valle d'Aosta/Vallée d'Aoste
Bolzano/Bozen
Trento
Bologna
Genova
Firenze
Venezia
Rimini
Verona

4.2.2 BIRCH

Also this Hierarchical clustering technique is an Agglomerative one. This implementation try to avoid the problem of dataset that not fit the memory, reducing also the overall complexity. It is an iterative approach that can be also used in data that arrives in stream. Another interesting aspect is that there is no need to set the number of clusters but they are automatically discovered during the execution of the algorithm. The parameters to be set are the *threshold* and the *branching factor*. It is used the same parameters for all the groups of indicators.

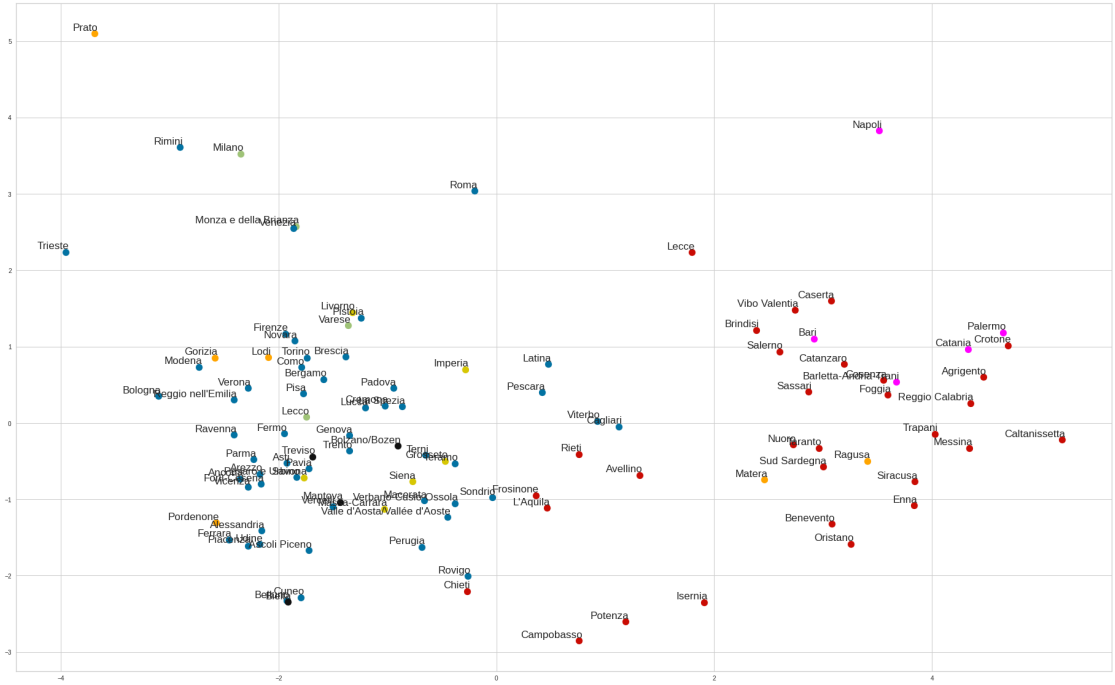


Figure 4.15: Business and Work with BIRCH

Its operation that uses only one scan and reduces memory usage, shows quite different clusters coming out compared to those seen so far, providing a new vision

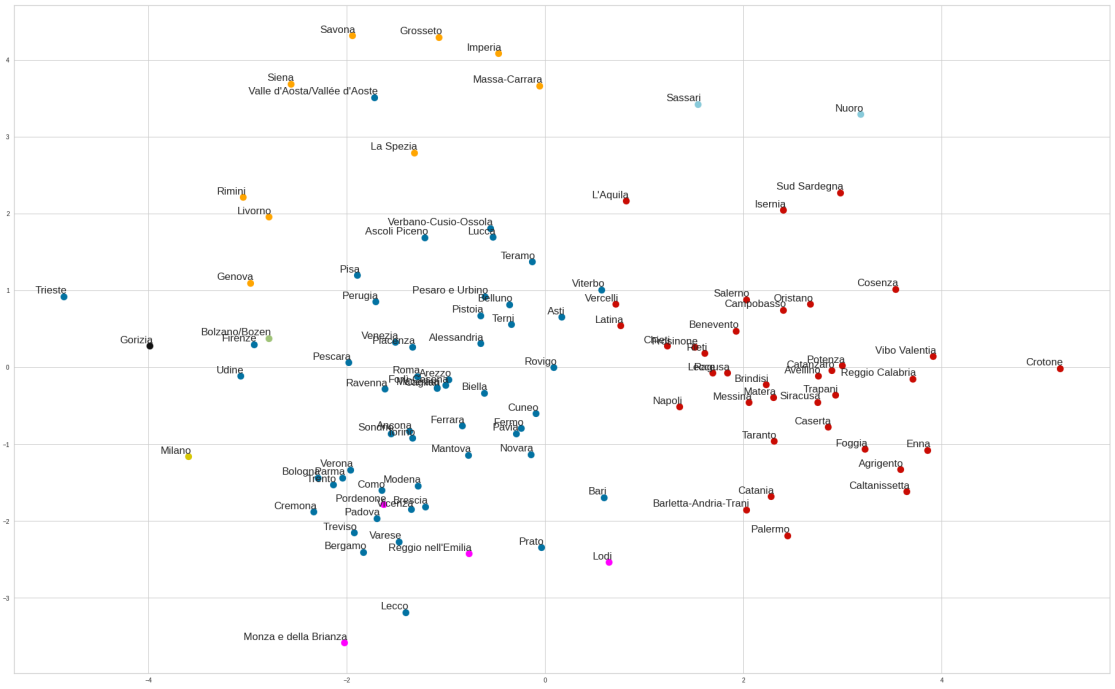


Figure 4.16: Culture and Leisure with BIRCH

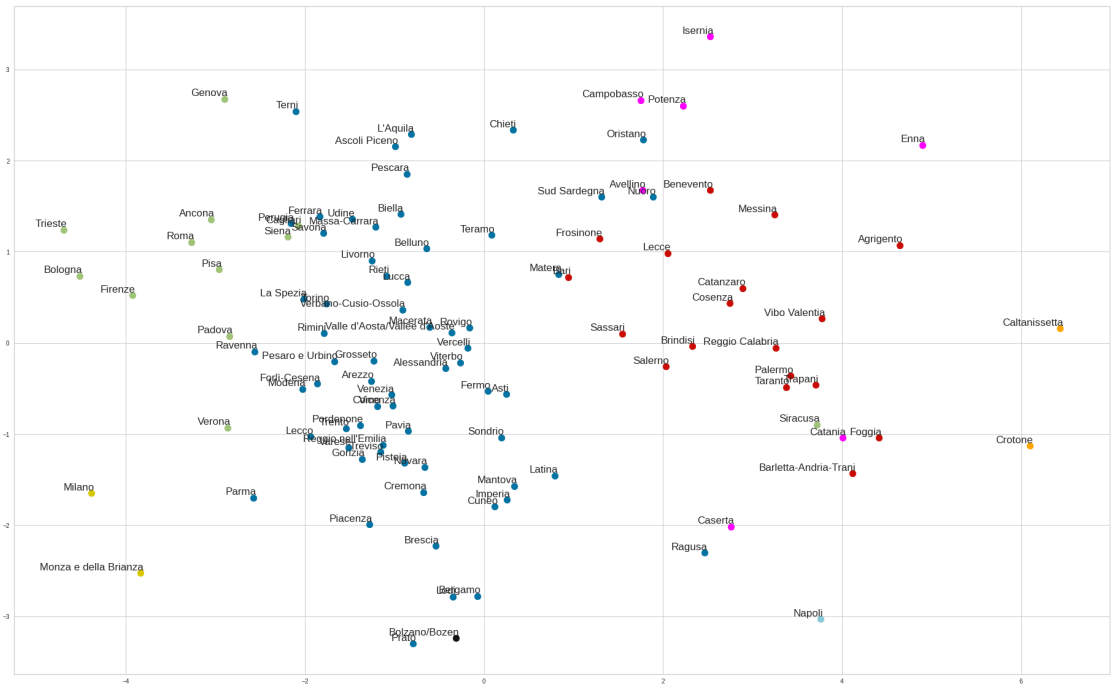


Figure 4.17: Demography and Society with BIRCH

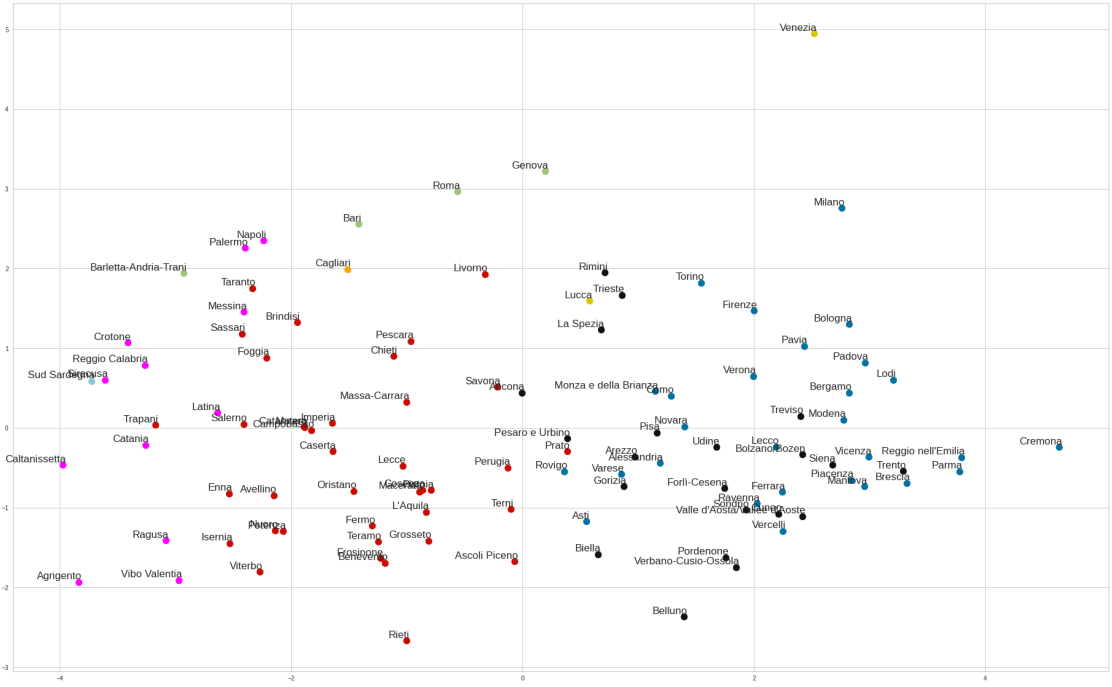


Figure 4.18: Environment and Services with BIRCH

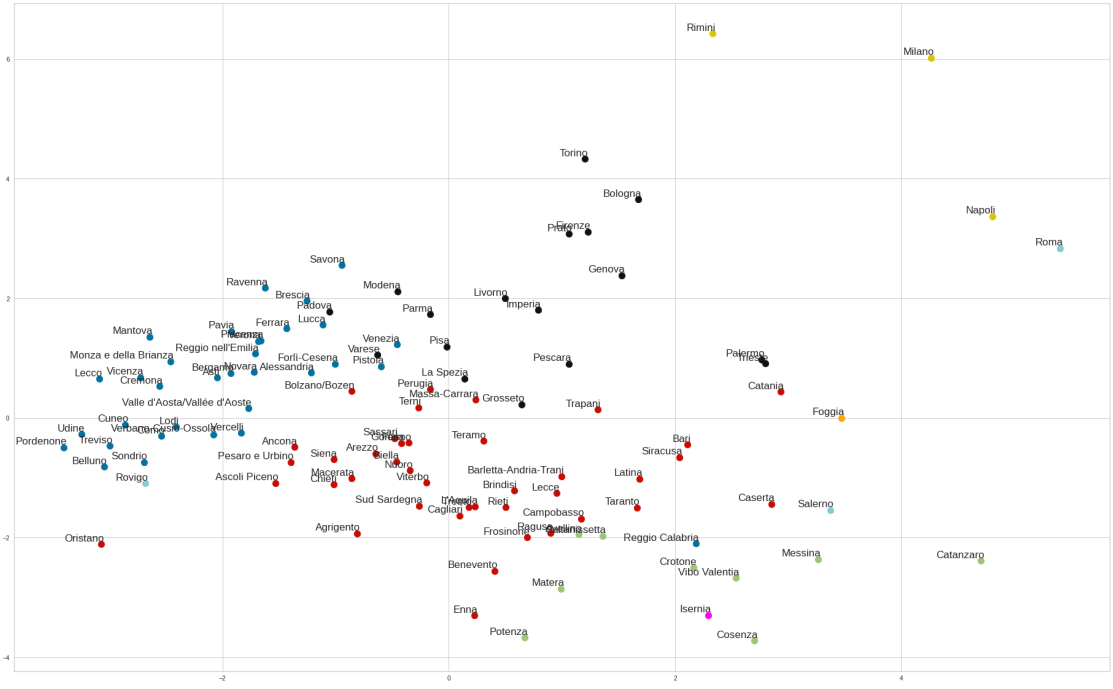


Figure 4.19: Justice and Security with BIRCH

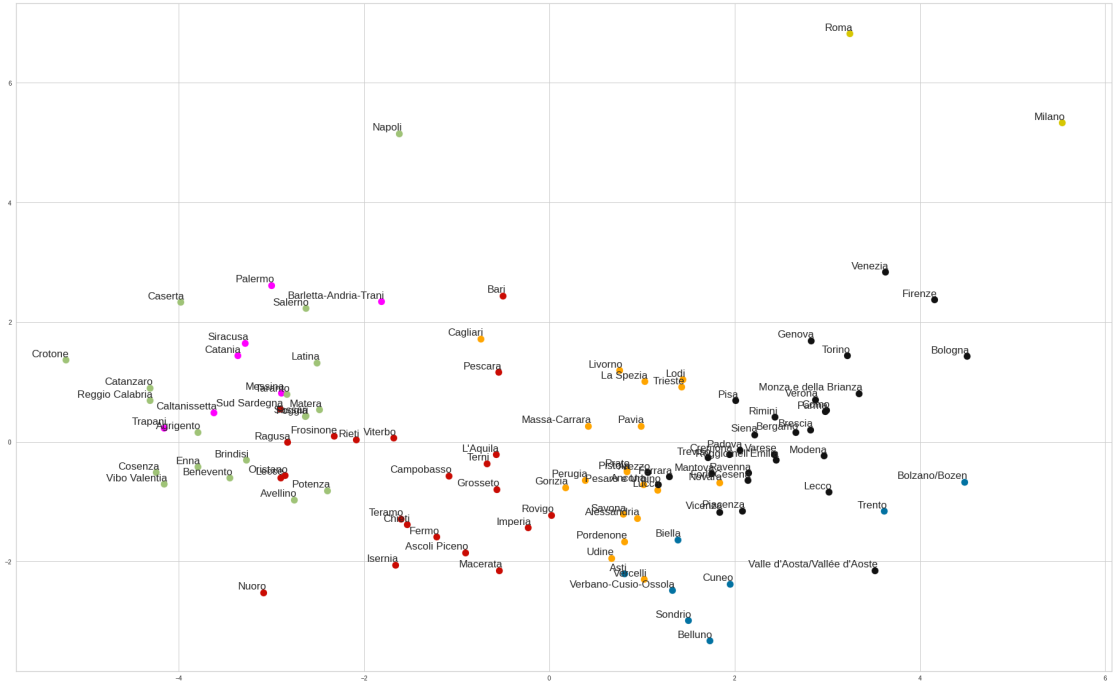


Figure 4.20: Wealth and Consumption with BIRCH

4.3 Density Based

This family of clustering approach is based on the concept of Density connected points. In this type of algorithm, It is not required the number of clusters as parameter, but the number is discovered using the density parameters like *Eps* and *MinPts* that together specified when consider an object a **Core Object**.

How to determine Eps and MinPts

An Heuristic approach to determine these parameters is computing the **Distance of the k-th nearest neighbor** :for a given k we define a function k dist , mapping each point to the distance from its k-th nearest neighbor. Then the points are sorted in descending order of their k-dist values: the graph of this function gives some hints concerning the density distribution.

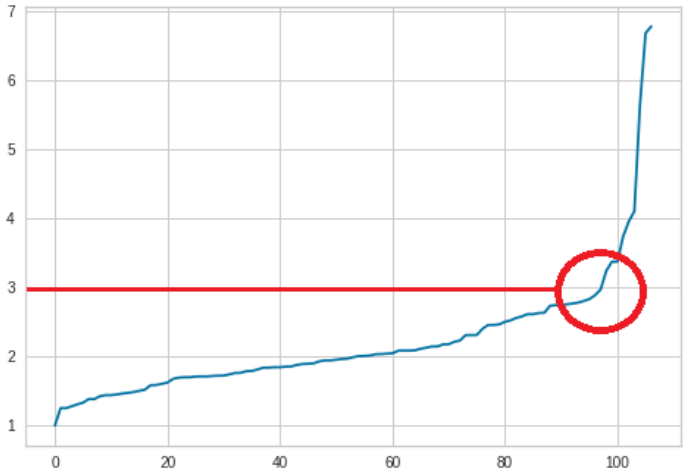


Figure 4.21: 3-dist graph for Wealth and Consumption 2022

The graph (k=3) would suggest **eps** = 3 and **minpts** = 3, but if we use such parameters the almost all objects are inserted in the same cluster with 4/5 outliers. By lowering **eps** = 2, 3 clusters are formed but one contains almost all the provinces. If It is tried to drop eps below 2, pretty much all provinces become outliers. Even doing other manipulations, you never find a condition where either you don't have almost all outliers, or there is a large cluster that contains most of the provinces.
This means that it is not possible to find density values that are good for the whole space, this is a very famous downside of **DBSCAN**, in particular the concept of global density. Certainly the structure of the data makes it very difficult to use the concept of density.
Unfortunately that happens also for the others groups of indicators, so the results are not represented since no meaningful.

Chapter 5

Temporal Cluster Analysis

In literature can be found some technique to assess the evolution of clusters over time, a problem named **Temporal Cluster Analysis**. For example in 'A News-Based Framework for Uncovering and Tracking City Area Profiles: Assessment in Covid-19 Setting' by Ducange, Marcelloni and Renda this problem is considered and approached using concepts of *Purity*, *Coverage* and *Preservation*. This approach exploits the *virtual assignment* that do the clustering of the previous set of data in time t-1, using the clusters discovered at window t.
In this work initially It was tried to exploit this approach but since the indicators inside each group change even considerably in some group during 2020, 2021 and 2022, It is considered forced or even wrong utilize this approach.
The idea in this work, to consider the evolution of the clusters, is much simpler and only uses the composition of the clusters in different time windows. The first operation is to gather for each group of indicators and for each time windows the clustering result. For the sake of simplicity are considered only the **K-Means** and **Agglomerative Hierarchical** approach in this analysis. From the result in the form of an array of 107 integers, It is generated a list of lists, the number of lists is equals to the number of clusters discovered, and in each list there are the numeric id of the provinces. $ex : [01120] -- > [[04], [12], [3]]$

Then It is computed the *matching matrix*.
It is calculated by sliding through clusters i, of time window t-1, and by sliding through clusters j of time window t, calculating how many objects in the clusters of row i go as a percentage of the total objects in the clusters of column j.
Then to graphically observe the evolution of the clusters, are computer the 2 matrices from 2020 to 2021 and 2021 to 2022, the results are showed in a **Sankey Diagram**.

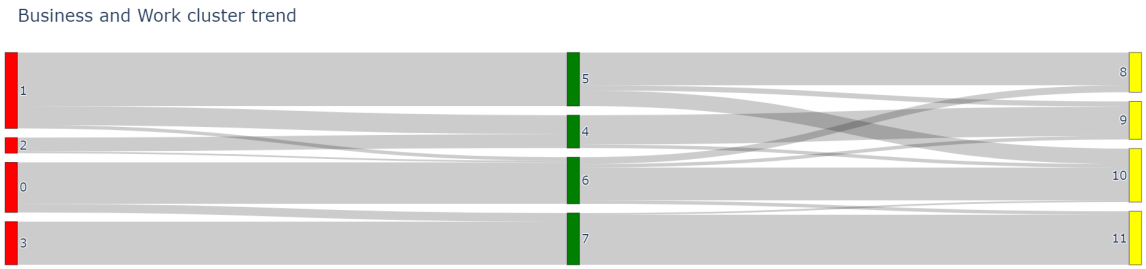
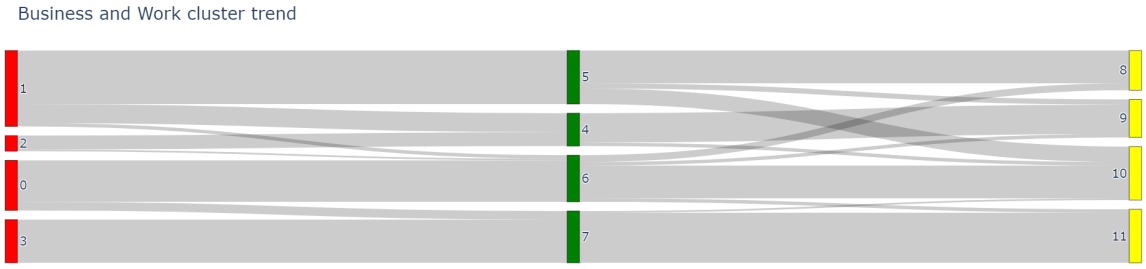


Figure 5.1: Example of Sankey Diagram

In order to analyze this diagram, can be easily observed some interesting pattern or behavior directly from the diagram and then using the intersection between cluster can be highlighted who is the responsible of the transition.

5.1 K-Means Trend

Business and Work

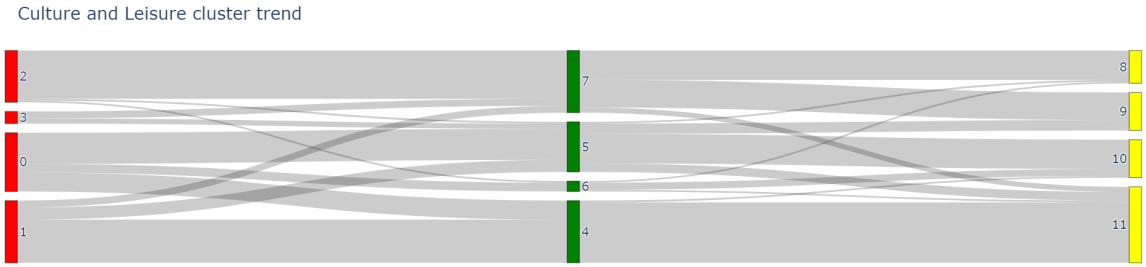


For example here we can highlight the transition of a small group of province from the cluster 0 in red to the cluster 7, when the majority of the provinces in cluster 0 go to cluster 6 in green. Analyzing who are these provinces we discover that:

- *cluster 0* : Avellino, Cagliari, Campobasso, Chieti, Frosinone, Grosseto, Imperia, Isernia, L'Aquila, La Spezia, Latina, Livorno, Lucca, Massa-Carrara, Matera, Nuoro, Oristano, Perugia, Pescara, Potenza, Rieti, Savona, Siena, Sondrio, Sud Sardegna, Teramo, Terni.
- *intersection between cluster 0 and cluster 7* : **Matera, Nuoro, Oristano, Sud Sardegna**.
- *cluster 7* : Agrigento, Bari, Barletta-Andria-Trani, Benevento, Brindisi, Caltanissetta, Caserta, Catania, Catanzaro, Cosenza, Crotone, Enna, Foggia, Isernia, Lecce, Matera, Messina, Napoli, Nuoro, Oristano, Palermo, Ragusa, Reggio Calabria, Salerno, Sassari, Siracusa, Sud Sardegna, Taranto, Trapani, Vibo Valentia.

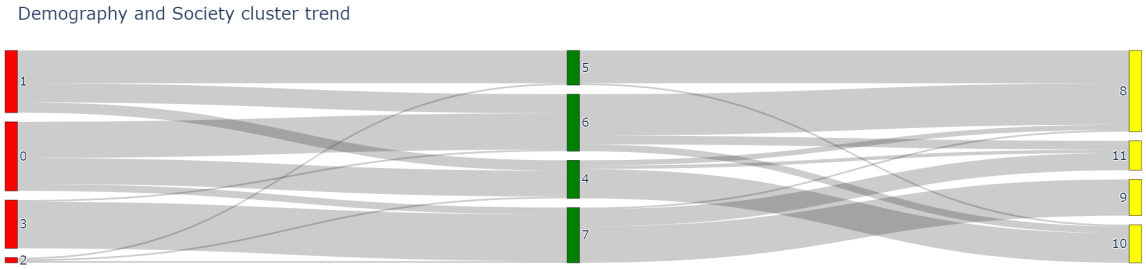
This can be seen as a worsening in Business and Work condition of the highlighted provinces between 2020 and 2021, maybe due to the coronavirus pandemic and the drop of the touristic sector.

Culture and Leisure

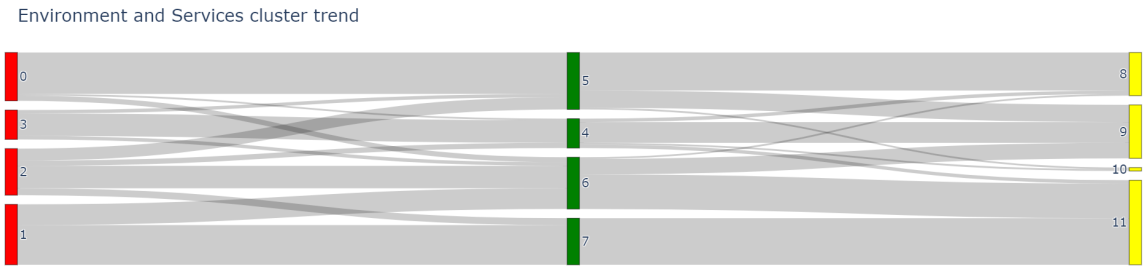


Here can be said that even some frgamentation and transition of provinces in different clusters, most objects remain clustered together.

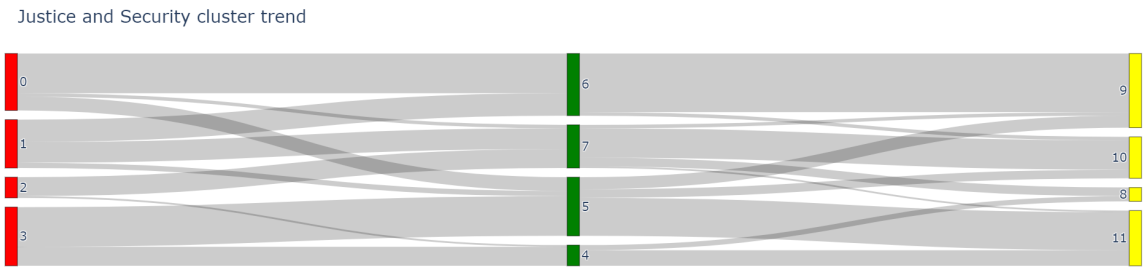
Demography and Society



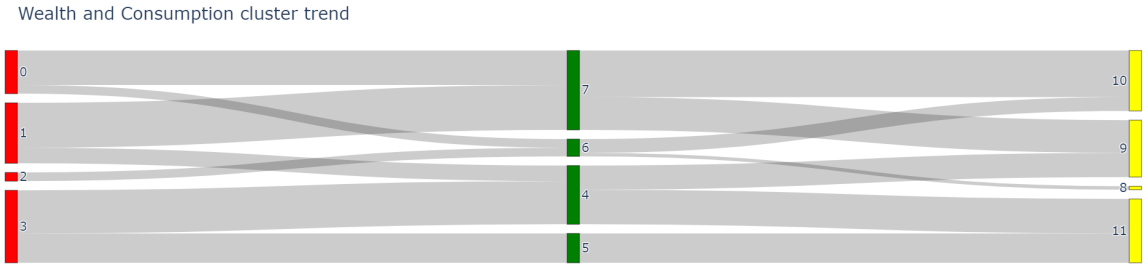
Environment and Services



Justice and Security



Wealth and Consumption

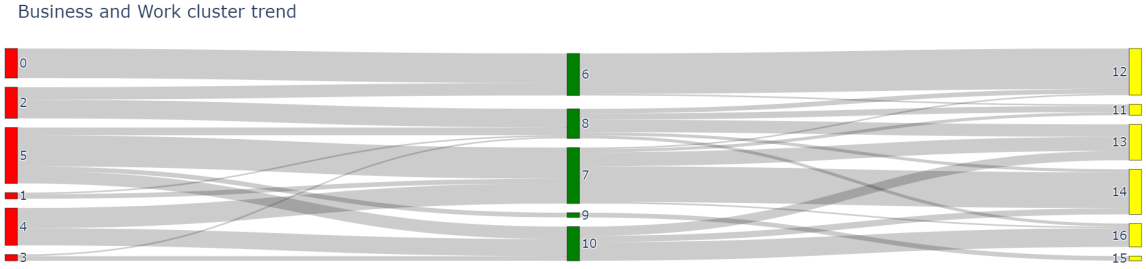


Also Wealth and Consumption seems to be pretty consistent over time.

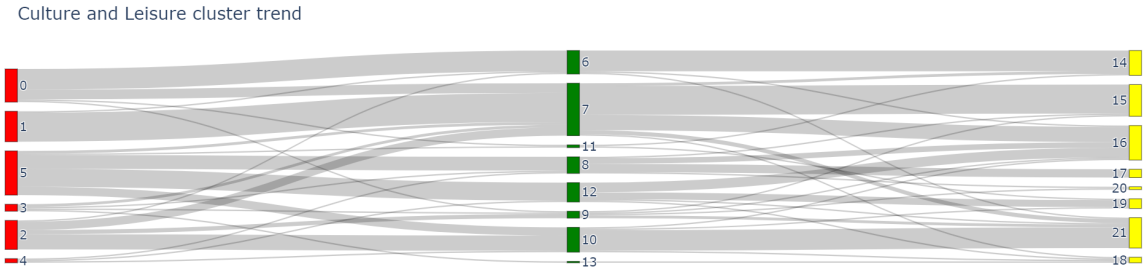
5.2 Hierarchical Agglomerative Trend

Note that to extract from the dendogram resulting from the agglomerative technique, It must be selected a distance threshold. The threshold decided is d=10 that allows to give an optimal number of clusters for each indicators, in the range of 4-8 clusters.

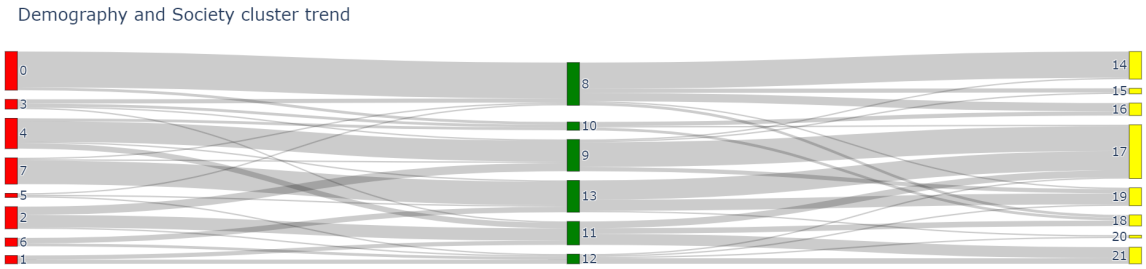
Business and Work



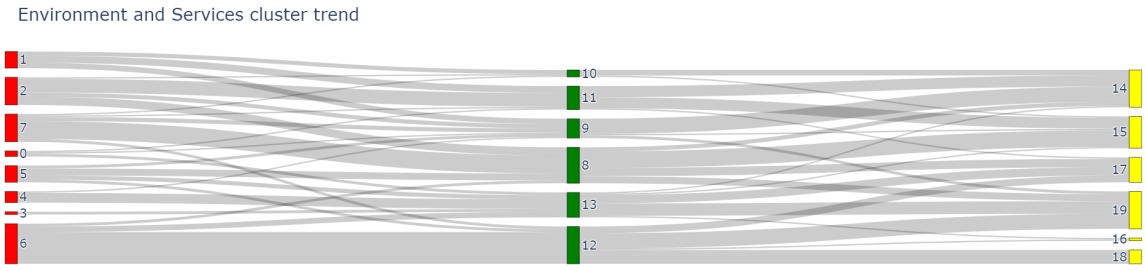
Culture and Leisure



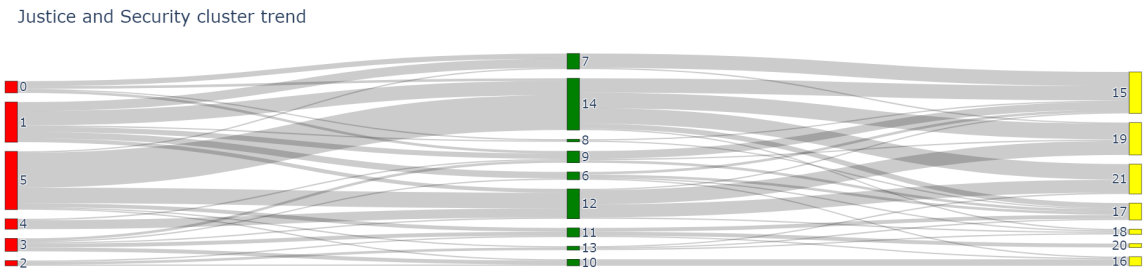
Demography and Society



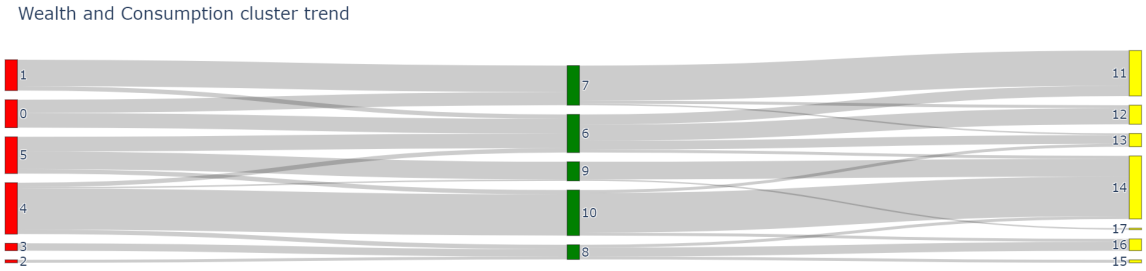
Environment and Services



Justice and Security



Wealth and Consumption



Chapter 6

Evaluation

The metric used in this evaluation section is the **Silhouette Coefficent**, the most used Intrinsic method (when you don't have the labels). The best clustering method is the **K-Means** with an average Silhouette Coefficent on the 6 indicators of 0.1782. With respect to the single group of indicators, the most suitable for the clustering seems to be **Demography and Society** . The worst method is CLARANS that heavily struggle to cluster Environment and Services and Business and Work.

	K-Means	CLARANS	Agglomerative	BIRCH
Business and Work	0.1567	0,096	0.1582	0.1614
Culture and Leisure	0.1913	0.1753	0.1816	0.1519
Demography and Society	0.199	0.1748	0.184	0.1796
Environment and Services	0.1557	0.085	0.1182	0.1097
Justice and Security	0.17	0.1	0.133	0.136
Wealth and Consumption	0.1965	0.1815	0.1785	0.123
Average	0.1782	0.1354	0.159	0.1436

Table 6.1: Silhouette Coefficent for each clustering methods and group of indicators

Chapter 7

Conclusion

Analyzing this data gave me the opportunity to use different clustering techniques and understand their pros and cons. The data probably didn't have a huge tendency to clustering, but some interesting results can be extracted anyway. The clustering technique that I personally appreciated the most with was the agglomerative hierarchical one, since it gave an overview of the situation in the various indicators that merge similar provinces effectively. It was also interesting to note that density-based techniques were not suitable for this type of dataset. The temporal cluster analysis using that simple technique is not so satisfying, It could be refined extracting some metrics, for example in the 'stability' of the clusters, this could be analyzed in future works.