Final Report ADASM

Student Paola Maria Lepore
Identification Number 2021453

In this report I have analyzed the *Quality of life 2024* dataset that contains 9 variables, measurements about some aspects of quality of life of 178 cities. The data are useful for an overall analysis of urban living standards and interesting for socio-economic comparison between wolrd's top cities in 2024.

First of all, I have loaded the matrix X from the file "QLCit24.m", then I have randomly sampled 60 cities from the total of 178 using as random seed for the permutation of the rows my data of birth, 23/12/2002.
Hence, the final matrix that I have analyzed is (60×9) where:

- The rows are the CITIES:

1	Medellin, Columbia	21	Taipei, Taiwan	41	Chicago, United States
2	Ljubljana, Slovenia	22	Rotterdam, Netherlands	42	Dubai, United Arab Emirates
3	Sofia, Bulgaria	23	Brussels, Belgium	43	Yerevan, Armenia
4	Melbourne, Australia	24	Warsaw, Poland	44	Tampa, United States
5	San Diego, United States	25	Budapest, Hungary	45	Limassol, Cyprus
6	Jeddah, Saudi Arabia	26	Ho Chi Minh City, Vietnam	46	Edinburgh, United Kingdom
7	London, United Kingdom	27	Split, Croatia	47	Dublin, Ireland
8	Lviv, Ukraine	28	Pune, India	48	Tbilisi, Georgia
9	Ankara, Turkey	29	Kathmandu, Nepal	49	Los Angeles, Unites States
10	Ottawa, Canada	30	Odessa, Ukraine	50	New York, United States
11	Oslo, Norway	31	Izmir, Turkey	51	Novosibirsk, Russia
12	Singapore, Singapore	32	Gothenburg, Sweden	52	Kiev, Ukraine
13	Washington, United States	33	Minsk, Belarus	53	Athens, Greece
14	Hamburg, Germany	34	Muscat, Oman	54	Chennai, India
15	Brasilia, Brazil	35	Rome, Italy	55	Lisbon, Portugal
16	Vilnius, Lithuania	36	Manchester, United Kingdom	56	Berlin, Germany
17	Phoenix, United States	37	Hyderabad, India	57	Vancouver, Canada
18	Winnipeg, Canada	38	Karachi, Pakistan	58	Bogota, Columbia
19	Skopje, North Macedonia	39	Reykjavik, Iceland	59	Zurich, Switzerland
20	Milan, Italy	40	Santiago, Chile	60	Colombo, Sri Lanka

- The columns are the VARIABLES:

	Name of variable	Description
1	*Quality of Life Index*	Overall index measuring quality of life, calculated based on various sub-indices.
2	*Purchasing Power Index*	Measures relative purchasing power in the city.
3	*Safety Index*	Indicates how safe the city is based on crime rates.
4	*Health Care Index*	Reflects the quality and accessibility of healthcare services.
5	*Cost of Living Index*	Represents the cost of living, including housing, food, and transportation.
6	*Property Price to Income Ratio*	A measure of housing affordability, calculated as the ratio of property prices to average incomes.
7	*Traffic Commute Time Index*	Average time spent commuting within the city.
8	*Pollution Index*	Level of environmental pollution, considering factors such as air and water quality.
9	*Climate Index*	Quality of the climate, taking into account factors such as temperature, precipitation, etc.

Introduction

In a preliminary analysis of the dataset it is important to look at the data. All variables are quantitative and continuous, so are feasible for my purposes, that are, generally, clustering and dimensional reduction.
I have computed the mean and the range, in Table 1, for the variables in order to explore the dataset and understand which are the most critical cities in the analysis.

VARIABLE	1	2	3	4	5	6	7	8	9
MINIMUM	68	18.1	35.5	49	19	2.9	18.4	15.1	16.1
MAXIMUM	214.2	181.2	83.7	86.6	100.4	42.7	59.8	95.7	99.8
MEAN	147.267	88.375	58.368	65.783	53.692	12.638	35.553	50.137	76.893

Table 1: Range and mean of variables.

It is important to remark that not all variables have a positive meaning. *Traffic Commute Time Index* and *Pollution Index*, are such that a higher value represents negative aspects. The maximum reached, respectively 59.8 and 95.7, represent the most critical values: in the first case it is the time spent in traffic and in the second case it is the level of pollution in the air. In fact, they are observed for cities that have lower *Quality of Life Index*, such as, respectively, Colombo (60) and Kathmandu (29). In addition, Colombo (60) is the worst city also according to the *Quality of Life Index*, since it is poor of infrastructures and has high criminality and lack of economic opportunities. In contrast, other variables are such that for increasing value, increases the positivity of the meaning.
The city that has the best quality of life is Rotterdam (22), which is known for its accessibility to services, environmental sustainability and economic opportunities.

The first step to achieve a good analysis is the standardization of the data: starting from the original data matrix X I have obtained the standardized data matrix Z. This step is crucial since the variables have been computed in different units of measurement, so, they, in the original asset, are not feasible for comparison.
Very interesting is the correlation matrix S_z, computed from the standardized matrix Z, its heatmap is showed in Figure 1.

Figure 1: Heatmap of correlation matrix.

The first, *Quality of life index*, and the second variable, *Purchasing Power index*, appear highly positively correlated: it is logical that for increasing purchasing power, quality of life could increase significantly. In fact, a city with high *Purchasing Power Index*, such as Zurich (59), is associated to an high value of *Quality of Life Index* which is 208.1.
Another pair of variables that are highly, but negatively, correlated are the first and the eighth variable, *Pollution Index*: even in this case it is coherent, since, as I said before, for an increasing level of pollution the evaluation of quality of life decreases. In fact the city which has the highest level of pollution, Kathmandu (29), has a value of *Quality of Life Index*, 95.7, far below average, 147.267.

Exercise 1

First of all, I want to analyze the dataset using some clustering techniques based on statistical modelling. Since, in general, semiparametric classification models are applied to the dissimilarity matrix, I need to compute the dissimilarity matrix D.
The distance matrix D is (60×60) and each element is the *Euclidean distance* between two cities. In Figure 2 there is the heatmap of the distances to better visualize the distance matrix D.

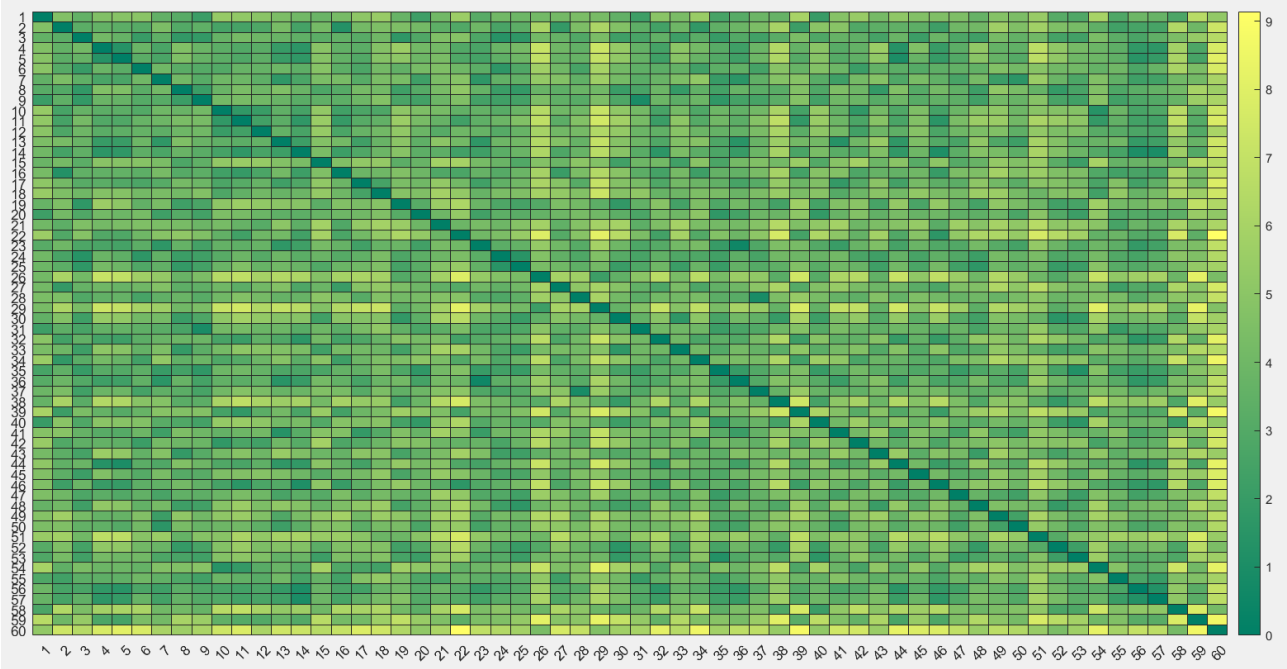


Figure 2: Heatmap of the distances between the cities.

Excluding the diagonal, which contains the distances between one city and itself, which obviously is 0, it is possible to identify the nearest and farthest cities, remembering that this is a 'mathematical' distance between measured data.

- NEAREST: Manchester (36) and Brussels (23) with a distance equal to 0.56;
- FARTHEST: Colombo (60) and Rotterdam (22) with a distance equal to 9.14.

From a general point of view, those results make sense since small Euclidean distances represent cities that have similar characteristics of quality of life. In fact, Brussels (23) and Manchester (36) have similar values for all variables, as shown in Table 2

Variable	1	2	3	4	5	6	7	8	9
Manchester	153.4	103.2	44.7	74.4	65.7	7.3	39.0	53.3	86.9
Brussels	148.4	105.0	44.40	73.6	66.6	7.0	36.8	61.6	83.8

Table 2: Values of variables for the nearest cities.

Similarly, high Euclidean distances represent cities that have very different values for the indexes taken into account, as happens for the two farthest cities, as shown in Table 3.

Variable	1	2	3	4	5	6	7	8	9
Rotterdam	214.2	139.1	71.6	81.2	61.3	5.5	22.9	24.0	87.9
Colombo	68.0	18.1	58.0	71.2	36.0	42.7	59.8	60.4	59.1

Table 3: Values of variables for the farthest cities.

In addition, the values of the *Quality of life index* for these two cities are, respectively, the maximum and the minimum. It is logical if one considers the level of technological progress and the richness achieved by the two cities:

Rotterdam has a significantly higher index's values than Colombo, meaning that it is a more rich and advanced city, which offers to its citizens a great quality of life with respect to Colombo, which is poorer and less able to guarantee good standard of life.

The first technique used to analyze the data is the *Well-structured perfect partition (WSPP)*. This model uses the dissimilarity data and is considered the most parsimonious for describing a clustering, since it is a simple model with few parameters. In fact, it assumes equal heterogeneity within clusters and equal isolation between clusters.
The idea is to minimize the following objective function with respect to the membership matrix U, the value of isolation between clusters α_2, and the value of heterogeneity within clusters α_1.
The model is:$||D - \alpha_2(1_n 1_n' - UU') - \alpha_1(UU' - I_n)||^2$, subject to: U binary and row stochastic and $0 \leq \alpha_1 \leq \alpha_2$.
Since I do not know the best number of clusters from the start, I have applied the *WSPP* function several times and computed the pseudoF statistics. It is a useful index to assess the quality of the clustering, and it is the ratio of between deviance and within deviance.
I have chosen the value of K which returns the maximum value of pseudoF, in my case 3, as shown in Table 4.

K	2	3	4
pseudoF	13.061	16.372	12.288

Table 4: PseudoF computed for different values of K for *WSPP*.

The elements required in input in the *WSPP* function are: the distance (or dissimilarity) matrix D, the number of clusters $K = 3$, and the number of iterations, 50 in my case.
The function returns in output the values that minimize the objective function of the model which are:

$$\text{heterogeneity: } \alpha_1 = 3.198 \qquad \text{isolation: } \alpha_2 = 4.696.$$

Also the membership matrix U contributes to minimize the objective function. Even in this case I choose to represent the heatmap of the membership matrix since I can have a better visualization of the matrix U. In Figure 3 rows are the 60 cities and columns are the clusters: each element is colored yellow if the city belongs to that clusters ($U_{ij} = 1$) and green otherwise ($U_{ij} = 0$).

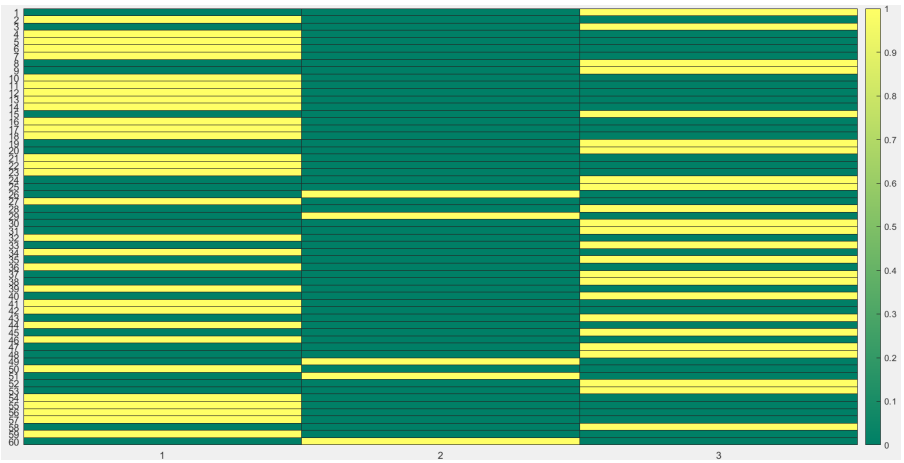


Figure 3: Heatmap of membership matrix U for *WSPP*.

From the heatmap in Figure 2 is also possible to determine how many and which cities are in each cluster: cluster 1 has 31 cities, cluster 2 has 5 cities, and cluster 3 has 24 cities. The partition is the following:

- CLUSTER 1: Ljubljana (2), Melbourne (4), San Diego (5), Jeddah (6), London (7), Ottawa (10), Oslo(11), Singapore (12), Washington (13), Hamburg (14), Vilnius (16), Phoenix (17), Winnipeg (18), Taipei (21), Rotterdam (22), Brussels (23), Split (27), Gothenburg (32), Muscat (34), Manchester (36), Reykjavik (39), Chicago (41), Dubai (42), Tampa (44), Edinburgh (46), New York (50), Chennai (54), Lisbon (55), Berlin (56), Vancouver (57), Zurich (59);
- CLUSTER 2: Ho Chi Minh City (26), Kathmandu (29), Los Angeles (49), Novosibirsk (51), Colombo (60);
- CLUSTER 3: Medellin (1), Sofia (3), Lviv (8), Ankara (9), Brasilia (15), Skopje (19), Milan (20), Warsaw (24), Budapest (25), Pune (28), Odessa (30), Izmir (31), Minsk (33), Rome (35), Hyderbad (37), Karachi

(38), Santiago (40), Yerevan (43), Limassol (45), Dublin (47), Tbilisi (48), Kiev (52), Athenes (53), Bogota (58).

The classification obtained through the *WSPP* seems to be coherent since the two nearest cities are in the same cluster, the first, while the two farthest cities are in a different cluster, Rotterdam (22) in the first and Colombo (60) in the second one.
It is also possible to say that the cluster made up of cities with a good quality of life is the first, while the cluster that has the worst cities is the second in which I can identify poor and distressing cities such as Colombo (60) and Kathmandu (29).

The second technique used to analyze data is the *Well-structure partition (WSP)*, a more flexible model in which I must specify different values for heterogeneity and isolation between clusters.
The idea of this method is to minimize the following objective function with respect to: the membership matrix U, the isolation matrix D_B and the heterogeneity matrix D_W.
The model is: $||D - UD_BU' - UD_WU' + diag(dg(UD_WU'))||^2$, subject to: U binary and row stochastic and $max(D_W) \leq min(D_B)$.
As in the previous case, I do not know the best value for the number of clusters K from the start, so I have applied the *WSP* several times and computed the pseudoF statistics.

K	2	3	4
pseudoF	2.020	18.685	13.986

Table 5: PseudoF computed for different values of K for *WSP*.

It is clear from Table 7 that the number of clusters which return the maximum value for pseudoF is $K = 3$, and therefore I choose 3 as the desired number of clusters.
Below are the matrices D_B and D_W that I have obtained in output on *Matlab* after applying the *WSP* function.

$$D_B = \begin{bmatrix} 0 & 6.592 & 4.247 \\ 6.592 & 0 & 4.639 \\ 4.247 & 4.639 & 0 \end{bmatrix} \quad D_W = \begin{bmatrix} 3.155 & 0 & 0 \\ 0 & 4.115 & 0 \\ 0 & 0 & 3.174 \end{bmatrix} \tag{1}$$

In order to understand if the partition is good, it is necessary to check if the *WSP* property holds: that is, since the $max(D_W) = 4.115 \leq min(D_B) = 4.247$. So there is lower distance between cities within clusters, meaning that they have similar distances. On the opposite side, there are higher distances between cities that are in different clusters, since the idea is to isolate each cluster from the others.
For matrix U I have plotted the heatmap in Figure 3 to better visualize the partition: each element is colored yellow if the city belongs to that cluster ($U_{ij} = 1$) and green otherwise ($U_{ij} = 0$).

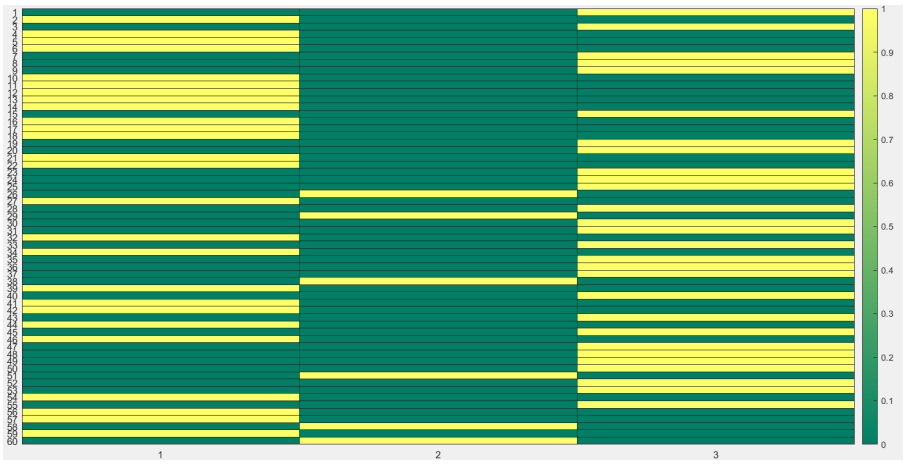


Figure 4: Heatmap of membership matrix U for WSP.

Also in this case, it is possible to determine from matrix U and, as a consequence from the heatmap in Figure 4, how many cities are in each cluster and which are the cities that belong to each of them: cluster 1 has 6 cities, cluster 2 has 28 cities and cluster 3 has 26 cities.
They are clustered in this way:

- CLUSTER 1: Medellin (1), Sofia (3), London(7), Lviv (8), Ankara (9), Brasilia (15), Milan (20), Brussels (23), Warsaw (24), Budapest (25), Pune (28), Odessa (30), Izmir (31), Minsk (33), Rome (35), Manchester (36), Hyderabad (37), Santiago (40), Yerevan (43), Limassol(45), Dublin (47), Tbilisi (48), Los Angeles (49), New York (50), Kiev (52), Athenes (53), Lisbon (55);
- CLUSTER 2: Ho Chi Minh City (26), Kathmandu (29), Karachi (38), Novosibirsk (51), Bogota (58), Colombo (60);
- CLUSTER 3: Ljubljana (2), Melbourne (4), San Diego (5), Jeddah (6), Ottawa (10), Oslo (11), Singapore (12), Washington (13), Hamburg (14), Vilnius (16), Phoenix (17), Winnipeg (18), Taipei (21), Rotterdam (22), Split (27), Gothenburg (32), Muscat (34), Rome (35), Chicago (41), Dubai (42), Tampa (44), Edinburgh (46), Chennai (54), Berlin (56), Vancouver (57), Zurich (59).

The partition obtained through *WSP* is similar to the partition obtained through *WSPP*. In fact, also in this case the two farthest cities are in a different cluster, Colombo (60) in the first and Rotterdam (22) in the third. Instead, the two nearest cities, Manchester (36) and Brussels (23), are both in cluster 2.
In this case the best cluster is the third one, since is made up of rich and wealthy cities, such as Rotterdam (22), Zurich (59) and Tampa (44). The worst cluster is the second one, as happens in the *WSPP*.

The third technique is the *Parsimonious dendrogram (PD)*. It aims to simplify dendrograms that have many levels and appear to be more complex. The idea is to minimize the following objective function with respect to: the membership matrix U, the isolation matrix D_B and the heterogeneity matrix D_W.
The model is: $||D - UD_BU' - UD_WU' + diag(dg(UD_WU'))||^2$, subject to: U binary and row stochastic and D_B ultrametric.
As before I did not know the best number of clusters, so I applied the *PD* technique several times and computed the pseudoF statistics. The highest value of the pseudoF is achieved with $K = 3$, as shown in Table 6.

K	2	3	4
pseudoF	2.02	18.271	12.034

Table 6: PseudoF computed for different values of K for *PD*.

The most interesting *Matlab* output of the function is the membership matrix U, plotted in the heatmap in Figure 5: each element is colored yellow if the city belongs to that cluster ($U_{ij} = 1$) and green otherwise ($U_{ij} = 0$).

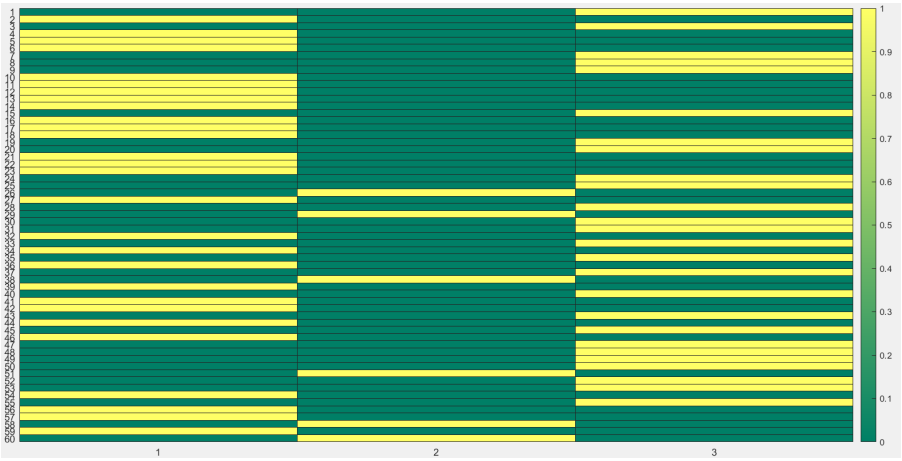


Figure 5: Heatmap of membership matrix U for PD.

From matrix U, as before, it is possible to determine how many cities are in each cluster: cluster 1 has 28 cities, cluster 2 has 6 cities and cluster 3 has 26 cities.
They are clustered in this way:

- CLUSTER 1: Ljubljana (2), Melbourne (4), San Diego (5), Jeddah (6), Ottawa (10), Oslo (11), Singapore (12), Washington (13), Hamburg (14), Vilnius (16), Phoenix (17), Winnipeg (18), Taipei (21), Rotterdam (22), Brussels (23), Split (27), Gothenburg (32), Muscat (34), Manchester (36), Reykjavik (39), Chicago (41), Dubai (42), Tampa (44), Edinburgh (46), Chennai (54), Berlin (56), Vancouver (57), Zurich (59).

- CLUSTER 2: Ho Chi Minh City (26), Kathmandu (29), Karachi (38), Novosibirsk (51), Bogota (58), Colombo (60);
- CLUSTER 3: Medellin (1), Sofia (3), London(7), Lviv (8), Ankara (9), Brasilia (15), Skopje (19), Milan (20), Warsaw (24), Budapest (25), Pune (28), Odessa (30), Izmir (31), Minsk (33), Rome (35), Hyderabad (37), Santiago (40), Yerevan (43), Limassol(45), Dublin (47), Tbilisi (48), Los Angeles (49), New York (50), Kiev (52), Athenes (53), Lisbon (55);

Also in this partition there is coherence with what I have observed from the Euclidean distances: Rotterdam (22) and Colombo (60) are in a different cluster, while Manchester (36) and Brussels (23) are in the same.
In this case the best cluster is the first one, that is made up of almost the same cities of the first cluster in *WSPP* and the third cluster in *WSP*. The worst cluster is the second which is similar to the second cluster of the previous techniques.
An automatic output of the *PD* function is the dendrogram, a tree diagram used to represent the hierarchical structure between units, in this case cities. In a dendrogram the nodes are the clusters that have been formed, while the edges are the relationships between clusters or units.
The parsimonious dendrogram of my dataset is shown in Figure 6.

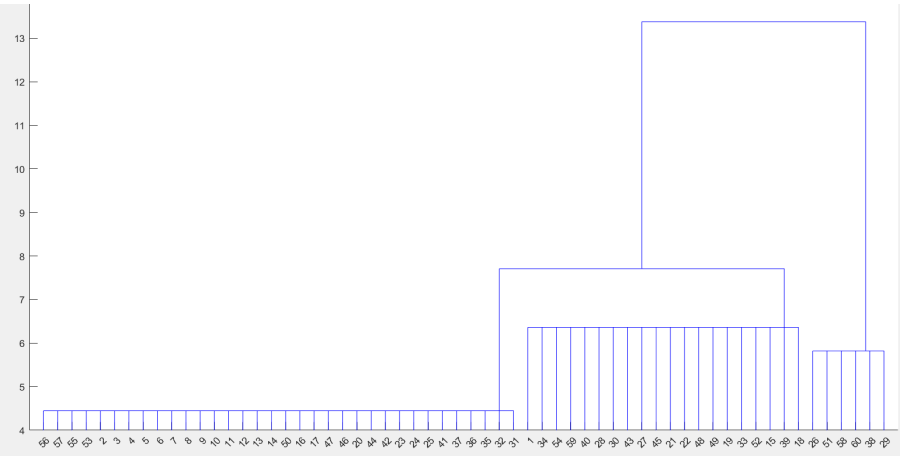


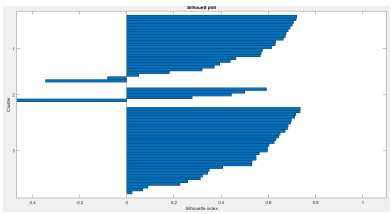
Figure 6: Parsimonious Dendrogram.

From the plot in Figure 6 the classification structure is clear: there are three large clusters, the ones computed before, and the hierarchical structure allows us to unify the two biggest group into one bigger cluster. That means that the cities in the first and third cluster have similar characteristics, so they are aligned in the evaluation of quality of life in different aspects.
A remark to add in the *PD* application is the ultrametricity check. I want the classification matrix to be ultrametric, since this property allows a better interpretation of results and guarantees the hierarchical representation.
So, I have applied the function that verifies ultrametricity to the reconstructed matrix through estimations:

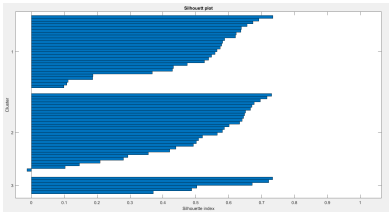
$$\hat{Q} = \hat{U}\hat{D}_B\hat{U}' + \hat{U}\hat{D}_W\hat{U}' - diag(dg(\hat{U}\hat{D}_W\hat{U}')).$$

Since the value of the function is almost 0, the ultrametricity is verified and my computation can be considered correct.

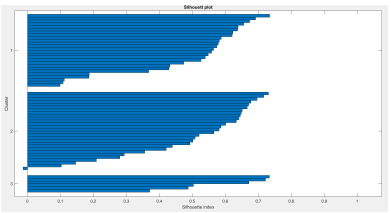
The comparison between these three methods could be conducted through the silhouette plots, represented in Figure 7(a), 7(b) and 7(c).
The silhouette plot is a very useful instrument to understand the goodness of the partition. It is helpful to state if the data fit well the clustering structure computed, and hence, each value measures if the unit is grouped into the right cluster. If the assignment is correct, the value of the silhouette plot is near 1, otherwise it is near -1.



((a)) Silhouette plot for *WSPP*.



((b)) Silhouette plot for *WSP*.



((c)) Silhouette plot for *PD*.

It is useful for the interpretation also to consider the mean of silhouette values, computed on distances, which are:

$$s_{WSPP} = 0.361 \qquad s_{WSP} = 0.347 \qquad s_{PD} = 0.347$$

Looking at the silhouette plot of *WSPP* in Figure 7(a), it is clear that there are some miss-classified cities, such as Budapest (25), Ho Minh City (26) and Minsk (33) since they have negative values meaning that are assigned to the wrong cluster.
Looking at the silhouette plot of the *WSP* in Figure 7(b), the partition seems better, since there is only one miss-classified city which is Bogota (58).
And looking at the last silhouette plot, it is equal to the previous one with only one miss-classified city, Bogota (58).
Therefore, since the highest value is achieved by *WSPP*, I can say that it is the best clustering methodology to apply to this data, although the other values are very close to it. In fact, I can say that the three partitions are very similar even in the definition of cluster since there are very small differences. For example, Brussels (23), Manchester (36), Reykjavik (39) and Rome (35) are in a cluster in *WSP* which is different from cluster in *PD*.

Exercise 2

In this second part of the report, I will analyze data with different kind of techniques, and, since I am also interested in a comparison between variables, I have worked on standardized data, taking into account the matrix Z from "QLCit24.m" dataset.
A very interesting technique for the dimensional reduction of original data is a sequential procedure that I can call *Tandem Analysis*. This method is made of two steps: the first step induces a dimensional reduction of original variables of Z through *PCA* and the second step induces a clustering for units, cities in my case, on the reduced matrix Y, through *K-means*.
First of all, I have applied the *Principal Component Analysis (PCA)* on Z: I have computed eigenvectors and eigenvalues of the correlation matrix S_z, shown below. I have determined the number of principal components thanks to the *Kaiser's Rule*: the number of principal components is equal to the number of eigenvalues larger than or equal to 1. Therefore, since there are 3 eigenvalues larger than 1 in the matrix L, colored red in equation 2, the number of principal components is 3.

$$A = \begin{bmatrix} -0.490 & -0.043 & -0.042 & -0.143 & -0.112 & 0.027 & -0.1690 & -0.173 & 0.8150 \\ -0.424 & 0.2830 & 0.1540 & 0.051 & -0.317 & -0.029 & 0.154 & -0.669 & -0.375 \\ -0.201 & -0.589 & -0.191 & 0.162 & -0.605 & -0.322 & -0.001 & 0.235 & -0.157 \\ -0.286 & 0.058 & -0.518 & 0.398 & 0.036 & 0.660 & 0.149 & 0.150 & -0.080 \\ -0.365 & 0.363 & -0.077 & 0.123 & 0.237 & -0.527 & 0.489 & 0.372 & 0.049 \\ 0.332 & -0.181 & -0.505 & 0.312 & 0.228 & -0.328 & 0.152 & -0.538 & 0.179 \\ 0.212 & 0.581 & -0.134 & 0.385 & -0.371 & -0.178 & -0.507 & 0.123 & 0.095 \\ 0.410 & 0.139 & 0.126 & -0.029 & -0.500 & 0.199 & 0.637 & 0.023 & 0.317 \\ 0.042 & 0.215 & -0.612 & -0.729 & -0.153 & -0.050 & -0.021 & 0.019 & -0.142 \end{bmatrix} \quad L = \begin{bmatrix} 3.987 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1.451 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1.130 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.920 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.578 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.478 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.296 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.160 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1.166e-06 \end{bmatrix} \quad (2)$$

Here A is the matrix where each column is an eigenvector of correlation matrix S_z, while L is a diagonal matrix where the diagonal elements are the eigenvalues of correlation matrix S_z.
The graphical representation of the components may be interesting, so I have made three scatter plot, in Figure 8, each one for a couple of components computed before.

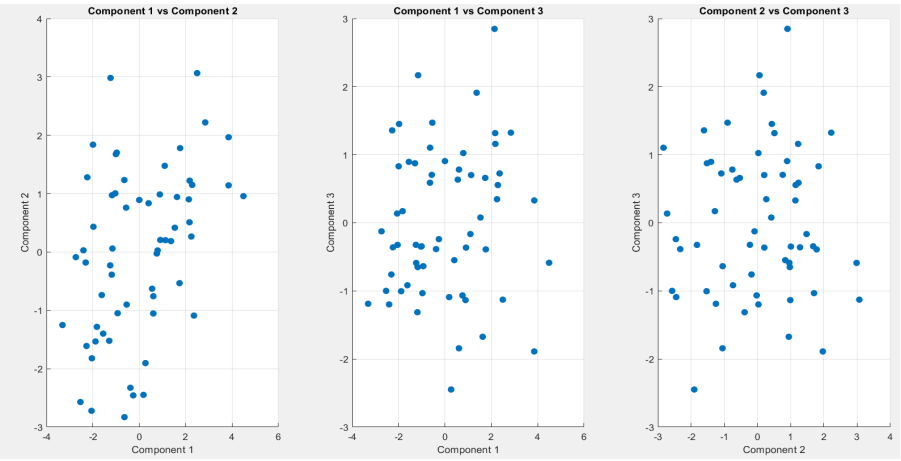


Figure 8: Scatter plot for the principal components.

These plots allow exploration of the distribution of the data in the new principal component space. Since the goal of the *PCA* is the dimensional reduction of the original data, here, I have reduced the space to dimension 3, but maintaining the most significant variance, that, in this case, is captured mainly by the first two components.
In fact, the first plot shows an higher variability and dispersion of data, meaning that these components, the first and the second one, capture an high variability of the data.
The most limited dispersion is in the third plot since capture less variance.
So I can construct the new data matrix Y, called *principal components*, of dimension (60×3): $Y = ZA_3$, with Z the standardized data and A_3 the matrix of loadings made of the first three columns of A which correspond to the largest eigenvalues.
Although there are some cross-loadings, such as for variable 5, *Cost of Living Index*, the other variables can reconstruct the components:

- PC 1: variable 1, *Quality of life index*, variable 2, *Purchasing Power Index*, variable 8, *Pollution Index*;
- PC 2: variable 3, *Safety Index*, variable 7, *Traffic Commute Time Index*;
- PC 3: variable 4, *Health Care Index*, variable 6, *Property Price to Income Ratio*, variable 9, *Climate Index*.

Since, looking at the loading matrix A_3 in equation 3, the interpretation of the components seems a bit difficult. To simplify it, I have rotated the matrix A_3 with *varimax*, an orthogonal rotation which maximizes loadings variance. The two matrices, A_3 and A_r, the rotated matrix, are reported below.

$$A = \begin{bmatrix} -0.490 & -0.043 & -0.042 \\ -0.424 & 0.283 & 0.154 \\ -0.201 & -0.589 & -0.191 \\ -0.286 & 0.058 & -0.518 \\ -0.365 & 0.363 & -0.0770 \\ 0.3320 & -0.181 & -0.505 \\ 0.212 & 0.581 & -0.134 \\ 0.410 & 0.139 & 0.126 \\ 0.042 & 0.215 & -0.612 \end{bmatrix} \qquad A_r = \begin{bmatrix} -0.401 & -0.279 & -0.070 \\ -0.525 & 0.058 & 0.070 \\ 0.127 & -0.631 & -0.100 \\ -0.200 & -0.149 & -0.540 \\ -0.476 & 0.129 & -0.166 \\ 0.449 & -0.056 & -0.440 \\ -0.066 & 0.591 & -0.217 \\ 0.274 & 0.334 & 0.130 \\ 0.027 & 0.135 & -0.635 \end{bmatrix} \tag{3}$$

In this case the variables that reconstruct the factors are:

- PC 1: variable 1, *Quality of life index*, variable 2, *Purchasing Power Index*, variable 5, *Cost of Living Index*, variable 6, *Property Price to Income Ratio*;
- PC 2: variable 3, *Safety Index*, variable 7, *Traffic Commute Time Index*, variable 8, *Pollution Index*;
- PC 3: variable 4, *Health Care Index*, variable 9, *Climate Index*.

Hence, the first component it is a summary of economic items. In fact, the variables that contribute most to the reconstruction of this component evaluate economic characteristic in terms of cost of living, purchasing power and accessibility to properties, which obviously have influence on quality of everyday life.
The second component summarizes environmental and safety factors. In fact, the variables that contribute most to it are the ones that evaluate the quality of life in terms of traffic, security and pollution. Those variables have something in common, even if they investigate different aspects: the safety of a city may depend on traffic, and greater time spent in city traffic can influence the safety of the city; traffic is one of the principal causes of pollution.
The third component summarizes two variables that evaluate different characteristics, since one measures accessibility to health care and the other climate index. Hence, it is the most heterogeneous component and is less useful for interpretation.

The second step of the sequential approach is the application of the *K-means (KM)* on the number of components identified with the *PCA*.
I have applied *KM* several times in order to determine the best value for K using pseudoF statistics.
The best value for K is 2, as shown in Table 7.

K	2	3	4	5
pseudoF	43.712	38.850	35.022	32.154

Table 7: PseudoF computed for different values of K for *KM*.

Recalling the usefulness of the silhouette plot, I have plotted two of them, each one for a different clustering, to interpret the goodness of the partition with *KM*.
The plot in Figure 9 represents the classification in 2 clusters: it seems quite good since most units are close to

0.7, a medium-high value. The plot in Figure 10 represents the classification in 3 clusters.
I have reported both plots for a comparison. In fact, the first plot is not what I expected since I hoped for higher values, but is better than the second one in which the maximum value achieved by units is lower, about 0.6.
This result is also supported by computing the mean of silhouette values.

$$s_2 = 0.43 \qquad\qquad s_3 = 0.35.$$

Although the mean of silhouette returns a better value for partition in 2 clusters than for partition in 3 clusters the value is not high. Therefore, since one generally hopes for a silhouette index at least equal to 0.60 or 0.70, the clustering structure identified here is not completely clear.
In any case, the silhouette method supports the decision taken with the pseudoF statistics and, finally, the best value for K is 2.

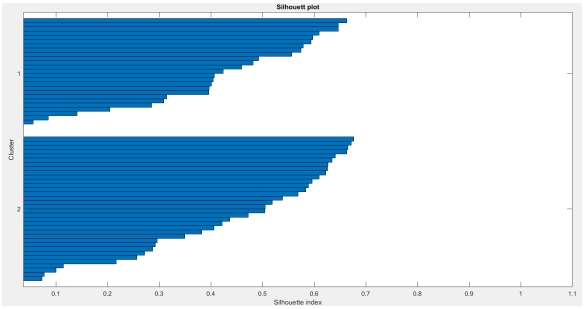


Figure 9: Silhouette plot for *KM* with K=2.

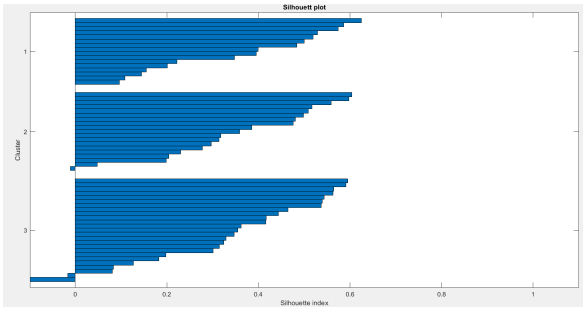


Figure 10: Silhouette plot for *KM* with K=3.

The clustering structure that I have obtained from the application of the *KM* is represented in the heatmap of the membership matrix U in Figure 11.

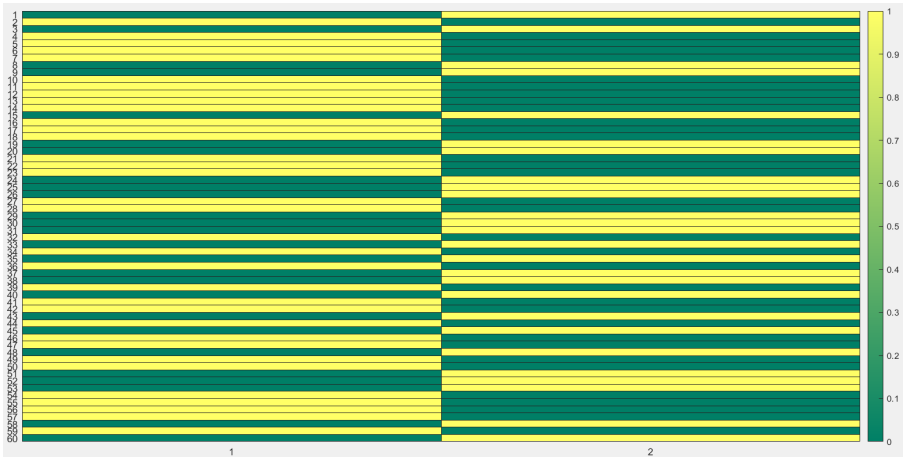


Figure 11: Heatmap of membership matrix U for *KM*.

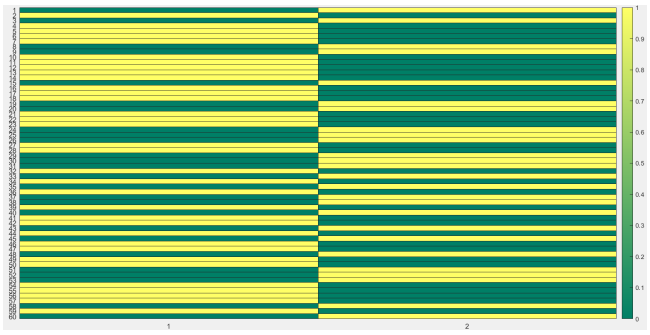
The clustering is such that cluster 1 has 34 cities and cluster 2 has 26 cities. The cities are classified as follows:

- CLUSTER 1: Ljubljana (2), Melbourne (4), San Diego (5), Jeddah (6), London (7), Ottawa (10), Oslo (11), Singapore (12), Washington (13), Hamburg (14), Vilnius (16), Phoenix (17), Winnipeg (18), Taipei (21), Rotterdam (22), Brussels (23), Split (27), Pune (28), Gothenburg (32), Muscat (34), Manchester (36), Reykjavik (39), Chicago (41), Dubai (42), Tampa (44), Edinburgh (46), Dublin (47), Los Angeles (49), New York (50), Chennai (54), Lisbon (55), Berlin (56), Vancouver (57), Zurich (59);
- CLUSTER 2: Medellin (1), Sofia (3), Lviv (8), Ankara (9), Brasilia (15), Skopje (19), Milan (20), Warsaw (24), Budapest (25), Ho Chi Minh City (26), Kathmandu (29), Odessa (30), Izmir (31), Minsk (33), Rome (35), Hyderabad (37), Karachi (38), Santiago (40), Yerevan (43), Limassol (45), Tbilisi (48), Novosibirsk (51), Kiev (52), Athens (53), Bogota (58), Colombo (60).

This classification seems coherent since in the first cluster there are rich cities that have achieved higher progress and that have a better quality of life according to values of variables. For example, there is the city that has the highest value for *Quality of Life index* (1), Rotterdam (22), with a value equal to 214.2, and also the city with the

second highest value, Zurich (59), with 208.1. I can also consider another variable, for example *Pollution Index* (8). The smallest value, which is the best since, as I said before, it has a negative meaning, are in Reykjavik (39), with a value equal to 15.1, in Gothenburg (32), with a value equal to 19.4 and in Oslo (11), with a value equal to 22.1, all cities of the first cluster.
Instead in the second cluster there are cities that have not good values for most of variables. In fact, for example, the worst value for *Health Care*, variable 4, is in Minsk (33), and is equal to 49.0; the worst value for *Quality of Life index* (1), is in Colombo (60), and is equal to 68. The worst value for *Safety Index* (3), is in Bogota (59), and is equal to 33.5, showing that this city is very unsafe and dangerous. All of them are in the second cluster, which groups the poorest and most underdeveloped cities.
This classification with 2 clusters can discriminate in a good way cities which have a better quality of life from cities that have lower quality of life in terms of safety, pollution, health care and economic issues.

The other two interesting clustering techniques are variations of the *KM*: the *Reduced K-means (RKM)* and *Factorial K-mean (FKM)*.
These techniques are methods used for the simultaneous clustering of cities and for the dimensional reduction for variables. In *RKM* the partition of cities is achieved through the *KM*, while the partition of variables is achieved through *PCA*. In *FKM* incorporates dimensionality reduction through *PCA* or *FA* before applying *KM*.
For both techniques I needed to choose a number of clusters for cities, K, and a number of clusters for variables, Q, since I wanted a reduction of the two dimensions.
I choose the number of clusters for variables, so columns, from the *PCA* results obtained before, which suggest $Q = 3$. For as concern K, I have taken into account the results of *KM* application that I did in the previous step, and hence I have chosen $K = 2$.
The most interesting output of the *RKM* are the membership matrix U for cities, in heatmap in Figure 12(a) and the loading matrix A for variables in Figure 12(b).



((a)) Heatmap of membership matrix U for *RKM*.

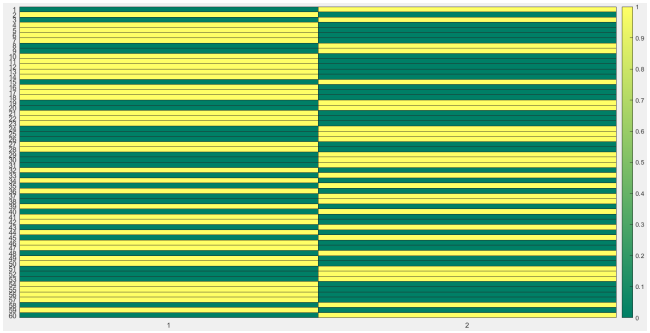
$$A = \begin{bmatrix} -0.032 & 0.075 & -1.006 \\ -0.546 & -0.064 & -0.010 \\ -0.154 & -0.017 & -0.003 \\ -0.386 & -0.074 & -0.005 \\ -0.372 & -0.356 & 0.009 \\ 0.560 & -0.460 & 0.036 \\ 0.033 & 0.215 & -0.010 \\ 0.140 & 0.777 & -0.035 \\ -0.002 & 0.189 & -0.0090 \end{bmatrix}$$

((b)) Loading matrix A for *RKM*.

For as concern cities, they are clustered into 2 groups, and structure is the same as in the application of *KM*. Instead, for as concern variables, I can define the clusters in the following way, through loadings in matrix A, assigning each variable to the component for which it has the highest loading.

- PC 1: *Purchasing Power index* (2), *Safety Index* (3), *Health Care* (4), *Property Price to Income Ratio* (6);
- PC 2: *Traffic Commute Time Index* (7), *Pollution Index* (8), *Climate Index* (9);
- PC 3: *Quality of Life Index* (1).

The identified components seem coherent with respect to the aspects that variables evaluate. The global variable, *Quality of Life Index*, is alone. However, I can recognize the first principal component as an index which measures economic factors and also something connected to health and safety, so it is a very heterogeneous component.
The second principal component is an overall variable of climate, in fact evaluates pollution and climate aspects.
For variable 5, *Cost of Living index*, there are cross-loadings, in fact, it is suitable for two clusters, the first and the second one, since, as shown in matrix A, has similar values for component 1 and 2. That means that it is an ambiguous variable and is not clear how to classify it.
Also in *FKM* it is interesting to comment on the membership matrix U, through its heatmap in Figure 13(a) and the loading matrix A in Figure 13(b).



((a)) Heatmap of membership matrix U for *FKM*.

$$A = \begin{bmatrix} -0.463 & -0.223 & -0.126 \\ -0.466 & 0.298 & -0.0290 \\ -0.131 & -0.066 & -0.030 \\ -0.339 & 0.697 & -0.385 \\ -0.429 & -0.597 & -0.367 \\ 0.288 & 0.104 & -0.177 \\ 0.104 & 0 & 0.071 \\ 0.392 & -0.056 & -0.814 \\ 0.067 & -0.007 & -0.023 \end{bmatrix}$$

((b)) Loading matrix A for *FKM*.

Even in this case, the classification of cities is the same obtained with *KM*.
In order to improve the interpretation of loadings, avoiding, if possible, cross-loadings, I have applied a *varimax* rotation on A_3, obtaining the following matrix.

$$A_3 = \begin{bmatrix} -0.494 & 0.105 & 0.159 \\ -0.129 & 0.480 & 0.244 \\ -0.139 & 0.025 & 0.050 \\ 0.042 & 0.857 & -0.117 \\ -0.811 & -0.119 & -0.061 \\ 0.172 & -0.001 & -0.309 \\ 0.099 & -0.078 & -0.001 \\ -0.143 & 0.049 & -0.893 \\ 0.028 & -0.031 & -0.0570 \end{bmatrix}$$

Table 8: Rotated loading matrix A for *FKM*.

The components obtained by interpreting the loadings in A_3, which seem quite heterogeneous, and are:

- PC 1: *Quality of Life Index* (1), *Safety Index* (3), *Cost of Living Index* (5), *Traffic Commute Time Index* (7);
- PC 2: *Purchasing Power Index* (2), *Health Care* (4);
- PC 3: *Property Price to Income Ratio* (6), *Pollution Index* (8), *Climate Index* (9).

Only the third component seems to incorporate similar variables since they all concern climate.
The first one mostly incorporates economic variables that, however, are not alone, since other variables that contribute most to it evaluate very different aspects, such as traffic and safety.
The second component seems the most heterogeneous, in fact, is made of two variables that measure very different aspects, one is economic, (2), the other concern health, (4). Actually, could be find a link between the two variables, since higher purchasing power can allow someone to pay for their health care, and hence, to have more accessibility to health services of the city.
Therefore, if I compare these two analyses, the most reliable seems to be the *RKM*, in fact the classification of variables seems more coherent with respect to the aspects of quality of life that they evaluate.

Very interesting is also the *Clustering and Disjoint PCA (CDPCA)*. It is a methodology that aims at a clustering of cities and a partitioning of variables, identifying the components with maximum variance. The idea is to minimize the within variance and the model is the following: $||X - U\bar{Y}V'B||^2$, subject to U and V binary and row stochastic and B diagonal and such that $V'BBV = I_Q$.
I have given in input the number of clusters for units, $K = 2$ from results of *KM*, and the number of clusters for variables, $Q = 3$, from results of *PCA*.
The most important output of the function *CDPCA* are the membership matrix U for cities in Figure 14 and the membership matrix V for variables in Figure 15. In fact, this technique uses a re-parametrization, $A = BV$, of the loading matrix which allows a better interpretation of results, assigning each variable at most to one component, trying to avoid cross-loadings.

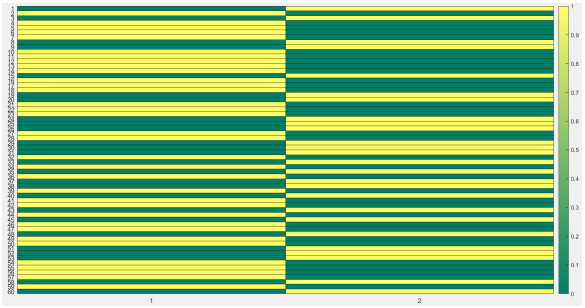


Figure 14: Heatmap of membership matrix U for *CDPCA*.

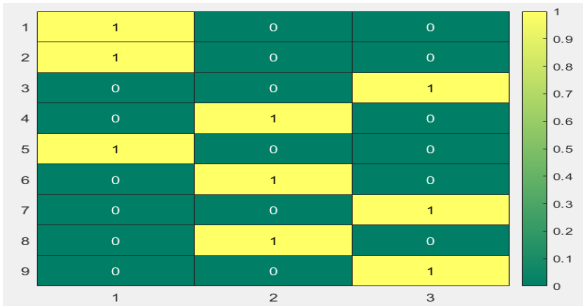


Figure 15: Heatmap of membership matrix V for *CDPCA*.

Since the number of clusters for units is the same of the *KM*, the partition is the same. For variables, the partition is different from the previous; in fact in this case each component has 3 variables.

- CLUSTER 1: *Quality of Life Index* (1), *Purchasing Power Index* (2), *Cost of Living Index* (5);
- CLUSTER 2: *Health Care Index* (4), *Property Price to Income Ratio* (6), *Pollution Index* (8);
- CLUSTER 3: *Safety Index* (3), *Traffic Commute Time Index* (7), *Climate Index* (9).

Even in this case, the classification of variables seems very heterogeneous, since the components are made up of variables that evaluate different aspects of quality of life. Sketchily, the first component is the one that unifies the economic indexes, since there are two variables that concern money and their power. The second is very heterogeneous since there are three variables that evaluate three completely different aspects concerning healthcare, economy and climate. At least, the third one can be interpreted as a measure of safety related to traffic and climate aspects.

The last methodology applied in this analysis is the *Double K-means (DKM)*, which is a technique used for simultaneous classification of cities and variables.
The idea is to minimize the within deviance, $||X - U\bar{Y}V'||^2$ subject to V and U binary and row stochastic.
In input, I gave the number of clusters for cities, $K = 2$ from *KM*, and the number of clusters of variables Q. In this case, I have applied *KM* on the transposed matrix Z to understand which would be a good partition for variables. As always, I have computed it several times and then I have chosen the value of Q which returns the maximum value of the pseudoF statistics, which is 2 as shown in Table 9. So I have applied the *DKM* with $K = 2$ and $Q = 2$.

Q	2	3	4
pseudoF	6.755	4.534	3.889

Table 9: PseudoF computed for different values of Q for *KM* on variables of Z.

The interesting output to comment on are the city's membership matrix, U, in heatmap of Figure 16 and the variable's membership matrix, V, in heatmap of Figure 17.

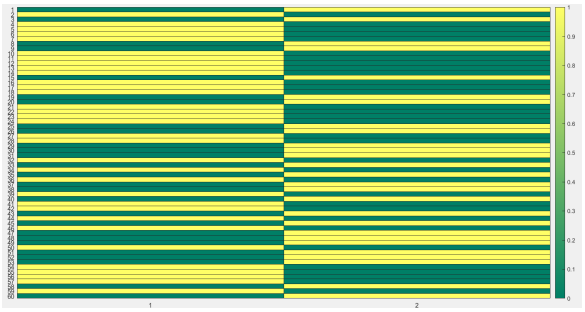


Figure 16: Heatmap of membership matrix U for *DKM*.

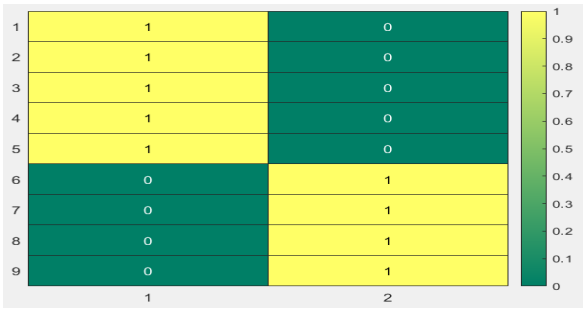


Figure 17: Heatmap of membership matrix V for *DKM*.

The clustering of cities is a bit different from the previous, in fact, here, the first cluster has 33 cities, while the second one has 27 cities. They are clustered in this way:

- CLUSTER 1: Ljubljana (2), Melbourne (4), San Diego (5), Jeddah (6), London (7), Ottawa (10), Oslo (11), Singapore (12), Washington (13), Hamburg (14), Vilnius (16), Phoenix (17), Winnipeg (18), Taipei (21), Rotterdam (22), Brussels (23), Warsaw (24), Split (27), Pune (28), Gothenburg (32), Muscat (34), Manchester (36), Reykjavik (39), Chicago (41), Dubai (42), Tampa (44), Edinburgh (46), New York (50), Chennai (54), Lisbon (55), Berlin (56), Vancouver (57), Zurich (59);
- CLUSTER 2: Medellin (1), Sofia (3), Lviv (8), Ankara (9), Brasilia (15), Skopje (19), Milan (20), Budapest (25), Ho Chi Minh City (26), Kathmandu (29), Odessa (30), Izmir (31), Minsk (33), Rome (35), Hyderabad (37), Karachi (38), Santiago (40), Yerevan (43), Limassol (45), Dublin (47), Tbilisi (48), Los Angeles (49), Novosibirsk (51), Kiev (52), Athens (53), Bogota (58), Colombo (60).

The cities that have changed cluster with respect to the previous analysis are Warsaw (24) and Los Angeles (49), but also Dublin (47). In particular, in the previous analysis Los Angeles (49) was in the cluster with the cities that guarantee a good quality of life, while, according to this methodology, it has changed cluster. On the other hand, Warsaw (24) and Dublin (47) has moved from the worst cluster to the better one. Maybe this behavior suggests that the variable's values of this city are in a situation in the middle between the wealthy cities and the more uncomfortable cities.
Also in this case the partition seems coherent with values of variables, and in particular with *Quality of Life Index* (1), since the first cluster groups cities with higher indexes, while the second has cities with lower level of quality life.
The variables are partitioned into 2 clusters as shown in the heatmap of the membership matrix V in Figure 17, and it is the following.

- CLUSTER 1: *Quality of Life Index* (1), *Purchasing Power Index* (2), *Safety Index* (3), *Health Care Index* (4), *Cost of Living Index* (5);
- CLUSTER 2: *Property Price to Income Ratio* (6), *Traffic Commute Time Index* (7), *Pollution Index* (8), *Climate Index* (9).

The partition of variables seems good, since the first cluster unifies the economic variables into a global indicator of that aspect, taking into account also healthcare and safety. Instead, the second cluster is mainly concentrated on climate aspects, such as traffic and pollution.
Finally another interesting result is the centroid matrix in the reduced Q-dimensional space $\bar{Y}$ in matrix shown in Table 10.

$$\bar{Y} = \begin{bmatrix} 0.563 & -0.354 \\ -0.688 & 0.432 \end{bmatrix}$$

Table 10: Centroid matrix $\bar{Y}$ for *DKM*.

From the matrix $\bar{Y}$ I can say that cities in the first cluster, which are the richest and most developed, have associated a positive value, meaning that it tends to assume higher values than the mean, equal to 0 since they are standardized, for as concern quality of life in terms of economic items, safety and accessibility to healthcare. While, it tends to assume lower values with respect to the mean for variables of the second cluster. It seems coherent considering that this negative value is associated to variables which have the negative meaning, *Traffic Commute Time Index* (7) and *Pollution Index* (8) and hence, values below the mean represent a good evaluation.
For as concern the second cluster, made up of poorer and less wealthy cities, the value is negative for the first cluster of variables, meaning that they tend to assume value significantly lower than the mean. They in fact are associated to lower value of *Quality of Life Index* (1). The second value for cluster 2 is positive, meaning that cities tend to assume higher value of the variables than the mean. Even in this case it is coherent with my previous analyses, since they assume higher value for variables that evaluate negative aspects, such as *Traffic Commute Time Index* (7) and *Pollution Index* (8).

The first instrument used to evaluate the consistency of the methodology's results is the confusion matrix, which returns a representation of the statistical classification accuracy. Each element of the matrix indicates the overlap between the clusters computed with the first technique and those calculated with the other. The idea is to count points assigned to both clusters.
To compare the results of the sequential procedure and of *RKM* I have computed the confusion matrix between the membership matrices of the output of the two functions, the matrix in Table 11.

$$U'_{TA}U_{RKM} = \begin{bmatrix} 34 & 0 \\ 0 & 26 \end{bmatrix}$$

Table 11: Confusion matrix for sequential procedure and *RKM*.

The result is a square matrix with off-diagonal elements equal to 0, meaning that there are no units that have been wrongly assigned. Both methods propose the same classification, so the two clusters of the two methodologies are made up of the same number of cities.
To compare the other two techniques, *FKM* and *CDPCA* I have computed the second confusion matrix in matrix shown in Table 12.

$$U'_{FKM}U_{CDPCA} = \begin{bmatrix} 34 & 0 \\ 0 & 26 \end{bmatrix}$$

Table 12: Confusion matrix for *FKM* and *CDPCA*.

Even in this case, there are no miss-classified units, so both methodologies propose a classification with the same number of clusters.
Therefore, in both cases the classifications seem good, in fact there are not miss-classified cities, and, hence, the summation of the diagonal is equal to the total number of units.
This is logical since the classification of units is almost the same for all the methodologies taken into account. This result means that cities are all well classified into their clusters, and hence they have similar value of the *Quality of Life Index*: one cluster gathers cities with a good life style, while the other gathers cities where is not guaranteed an high well-being.
In addition, I want to compare the technique that seems the best one based on dimensional reduction of variables, *RKM*, with the last technique which clusters variables, *DKM*. The comparison is conducted through the confusion matrix between the two membership matrices of the models in Table 13.

$$U'_{RKM}U_{DKM} = \begin{bmatrix} 32 & 2 \\ 1 & 25 \end{bmatrix}$$

Table 13: Confusion matrix for *RKM* and *DKM*.

In this case, there are some miss-classified units since the off-diagonal elements are different from 0. And hence, the two methods tend to classify 3 units in different ways: looking at the identified clusters, the miss-classified cities are Warsaw (24), Los Angeles (49) and Dublin (47), as I commented before.

Certainly, the most important value to take into account is the explained variance of the models, stored in Table 14. It is a measure of the variability of the data captured by the model, and it is computed as the ratio between the variance of the model and the total variance of the dataset.

Methodology	*PCA + KM*	*RKM*	*FKM*	*CDPCA*	*DKM*
Explained variance	25.614	31.682	31.682	31.658	28.287

Table 14: Values of explained variance of different techniques.

According to the explained variance values, the best methodologies are the *RKM* and the *FKM*, which returns the highest value.
The results are a bit disappointing, since I was hoping for at least 60% or 70% of explained variance. In fact, the two models capture only the 31.682% of the variability of the data, so that they are unable to explain a good part of the information.
Very low is the explained variance of the sequential procedure: it is logical, since the simultaneous procedures generally return the best solutions, and, hence, capture more information and variability.
Even the *DKM* has a low explained variance, in fact it is more useful for the dataset where a simultaneous

classification of units and variables is possible, such as customer by products or gene dataset. In fact, from the comparison with the *RKM*, the classification of cities is different, and maybe this lead to a penalization of the model.
According to the explained variance, the methodologies to prefer to analyze the dataset of quality life are the *RKM* and *FKM*, even if *CDPCA* has also a quite good value, very close to these two. In fact, especially *FKM*, these techniques work well with questionnaires and datasets with demographic and social variables, as happens in this case.
Finally, is clear that the classification in 2 clusters can distinguish well the richest cities, and hence is able to guarantee a good quality of life in terms of safety, healthcare, climate and economic well-being, such as Rotterdam (22) and Washington (13) from the poorer cities, which are less able to guarantee a good lifestyle, such as Colombo (60) and Kathmandu (29).
Similarly for variables, especially in *RKM* results, the 3 components identified in my computations are able to discriminate 3 different aspects: the economic item, the climate item and a general index of quality of life. Hence, I can say that *RKM* works quite well with the dataset of interest with respect to the other methodologies, which have some weaknesses.

Appendix: Matlab Code

Below is the MATLAB code used for the computation of all the methodologies and all the plots in this report.

```
QLCit24;
rng(231202);
rowp = randperm(178);
Xp = Q(rowp, :);
X = Xp(1:60,:);
u60 = ones(60,1);
mean(X);
min(X);
max(X);
% EXERISE 1
% 1. STANDARDIZE DATA
Jc = eye(60)-(1/60)*ones(60);
Sc = 1/60*X'*Jc*X;
D = diag(diag(Sc).^0.5);
Z = Jc * X * D^-1;
% variance and covarianze matrix of Z or correlation matrix of X
Sz = 1/60*Z'*Z;
% heatmap of variance and covariance matrix
heatmap(Sz);
% 2. COMPUTE THE EUCLIDIAN DISTANCE BETWEEN UNITS
D = squareform(pdist(Z, 'euclidean'));
% heatmap of the matrix of distances
heatmap(D);
% 3. COMPUTE THE WSPP (3 clusters)
[U3,b3, a3, f3,iter3]=WSPP(D, 3, 50);
[pf3,Dw3,Db3] = psF(Z,U3);
% heatmap of membership matrix U3
heatmap(U3)
% how many cities in each cluster
unitCounts1 = sum(U3);
% which are the cities for each cluster
find(U3(:,1));
find(U3(:,2));
find(U3(:,3));
% silhouette plot
[~, clusterAssignments1] = max(U3, [], 2);
unitCounts1 = histcounts(clusterAssignments1, 1:(3+1));
silhouette(D,clusterAssignments1);
title('Silhouett plot');
xlabel('Silhouette index');
ylabel('Cluster');
mean(silhouette(Z,clusterAssignments1));
% 4. COMPUTE THE WSP (3 clusters)
[U3,Db3, Dw3, f3,iter3]=WSP(D, 3, 50);
[pf3,Dw3f,Db3f] = psF(Z,U3);
% heatmap of membership matrix U3
heatmap(U3)
```

```
% how many cities in each cluster
unitCounts1 = sum(U3);
% silhouette plot
[~, clusterAssignments1] = max(U3, [], 2);
unitCounts1 = histcounts(clusterAssignments1, 1:(3+1));
silhouette(D,clusterAssignments1);
title('Silhouett plot');
xlabel('Silhouette index');
ylabel('Cluster');
mean(silhouette(Z,clusterAssignments1));
% 5. COMPUTE THE PD  (3 clusters)
[U3pd,Db3pd, Dw3pd, f3pd,iter3pd]=PD(D, 3, 50);
[pf3pd,Dw3pd,Db3pd] = psF(Z,U3pd);
% heatmap of membership matrix U3
heatmap(U3pd)
% how many cities in each cluster
unitCounts1 = sum(U3pd);
% silhouette plot
[~, clusterAssignments1] = max(U3pd, [], 2);
unitCounts1 = histcounts(clusterAssignments1, 1:(3+1));
silhouette(D,clusterAssignments1);
title('Silhouett plot');
xlabel('Silhouette index');
ylabel('Cluster');
mean(silhouette(Z,clusterAssignments1));
% verify ultrametricity
Q3 = U3pd*Db3pd*U3pd'+U3pd*Dw3pd*U3pd'-diag(diag(U3pd*Dw3pd*U3pd'));
Verultrametrica(Q3);

% EXERCISE 2
% 2. COMPUTE PCA TO DETERMINE THE NUMBER OF PRINCIPAL COMPONENTS
[A, L] = eigs(Sz, 9);
diag(L);
A3 = A(:,1:3);
A3r = rotatefactors(A3);
Y = Z * A3r;
% scatter plot
% Component 1 vs Component 2
subplot(1,3,1);
scatter(Y(:,1), Y(:,2), 50, 'filled');
xlabel('Component 1');
ylabel('Component 2');
title('Component 1 vs Component 2');
grid on;
% Component 1 vs Component 3
subplot(1,3,2);
scatter(Y(:,1), Y(:,3), 50, 'filled');
xlabel('Component 1');
ylabel('Component 3');
title('Component 1 vs Component 3');
grid on;
% Component 2 vs Component 3
subplot(1,3,3);
scatter(Y(:,2), Y(:,3), 50, 'filled');
xlabel('Component 2');
ylabel('Component 3');
title('Component 2 vs Component 3');
grid on;
% 3. COMPUTE K-MEANS ON THE NUMBER OF COMPONENT IDENTIFIED IN THE STEP 1
[loopOtt,UOtt,fOtt,iterOtt] = kmeansVICHI(Y,2,50);
[pf,Dw,Db] = psF(Z,UOtt);
% silhouette plot
[~, clusterAssignments1] = max(UOtt, [], 2);
unitCounts1 = histcounts(clusterAssignments1, 1:(2+1));
silhouette(Z,clusterAssignments1);
title('Silhouett plot');
xlabel('Silhouette index');
ylabel('Cluster');
mean(silhouette(Z,clusterAssignments1));
% pseudoF comparison
K = 10;
pfKm = zeros(K,3);
for k = 2:K
```

```
    [~,UOtt,~,~]=kmeansVICHI(Y,k,50);
    [pF,Dw,Db]=[psF,Dw,Db];
    pfKm(k,:)=[psF,Dw,Db];
end
c = [pfOtt2, pfOtt3, pfOtt4, pfOtt5, pfOtt6, pfOtt7, pfOtt8, pfOtt9, pfOtt10];
% heatmap fo 2 clusters
heatmap(UOtt2);
sum(UOtt2);
find(UOtt2(:,1));
find(UOtt2(:,2));
% 5. COMPUTE REDUCED K-MEAN (k=2)
[Urkm,Arkm, Yrkm,frkm,inrkm]=REDKM(Z, 2, 3, 50);
pFrkm = psF(Yrkm,Urkm);
% heatmap for U
heatmap(Urkm);
% 6. COMPUTE FACTORIAL K-MEAN;
[Ufkm,Afkm, Yfkm,ffkm,infkm]=FKM(Z, 2, 3, 20);
pFfkm = psF(Yfkm,Ufkm);
heatmap(Ufkm);
% find cities in clusters
sum(Ufkm);
find(Ufkm(:,1))
find(Ufkm(:,2))
% rotation of factors
rotatefactors(Afkm);
% 7. COMPUTE CDPCA
[Vcdpca,Ucdpca,Acdpca, Ycdpca,fcdpca,incdpca]=CDPCA(Z, 2, 3, 50);
heatmap(Ucdpca);
heatmap(Vcdpca);
% 8. COMPUTE DKM
% choice of q
[loopd,UOttd,fOttd,iterOttd]=kmeansVICHI(Z',2,50);
[pfOttd,Dwd,Dbd]=psF(Z',UOttd);
% application of DKM
[Vdkm,Udkm,Ymdkm,fdkm,indkm] = DKM(Z, 2, 2, 50);
heatmap(Udkm);
heatmap(Vdkm);
% 9. COMPARE THE RESULTS BY USING THE CONFUSION MATRIX (CONTINGENCY TABLE)
% confusion matrix 4 and 5
UOtt*Urkm;
% confusion matrix 6 and 7
Ufkm'*Ucdpca;
% confusion matrix 5 and 8
Urkm'*Udkm
% 10. COMPUTE THE EXPLAINED VARIANCE FROM 4 TO 8
Db/(sum(var(Y)));
```

Listing 1: MATLAB code