Coursework for MATH96067: Introduction to Statistical Learning
CID: 01379199

1) Introduction:

Tunisia is a small country in North Africa with a population of over 10 million habitants. It is considered as a developing country; Tunisia's Gross National Income per capital is relatively low and the current political and economic situation is unstable. The country is also suffering from major disparities between regions that led the government to rethink its current road infrastructure. This coursework is focused on applying the distance-based method to evaluate which areas need to be the centre of attention. We follow the methodology of multidimensional scaling outlined in the course lecture notes. We discover that the estimated areas that need major changes are mostly aligned with our conjectures. Our analysis has been performed with R.

2) Data:

In our analysis, we will use a dataset that contains the travelling time between each of the 24 Tunisia's governorate. This is a symmetric matrix with zeroes on the diagonal. This dataset has been collected by myself. I used google-maps to estimate the travelling time between each governorate. Throughout my research, the time difference between going from A to B and B to A was relatively small (around 2-3 minutes). We neglected this difference as it will allow us to perform a classical scaling where the distance matrix is symmetrised. I extracted the data during late-night hours to make sure that it is not affected by external factors like traffic. Tunisia is a relatively flat country therefore we can neglect effects of roads that might be situated in mountains.

3) Methodology:

With the distance-based method, we will be able to analyse the information that we have on travelling time between governorates to construct a configuration that represents the map of Tunisia. Since this data does not represent geographic distances, the map will be entirely based on travelling time. Therefore, we will understand the main discrepancies of Tunisia's road infrastructure (no presence of highway in the south, developed infrastructure in the North...).
The method consists of getting our configuration from the distance matrix. This analysis is carried out using the cmdscale function in R. The cmdscale function allows us to get our configuration. However, we must choose its dimensions ie. number of dimensions that we think are most relevant to represent our configuration. We initialise by setting this number to equal to 10. After careful analysis of the matrix eigenvalues, we will make a better choice based on the results that we find. Finally, we plot our configuration and compare it to the real one.

4) Results:

As a first step for our analysis, we decide which dimension we should set for our configuration. The plot in figure 1 shows us the value of the eigenvalues against their respective number. Based on the mathematical model, we expected to get a specific number of eigenvalues that are particularly large. This number will be used for our dimensions. The other eigenvalues must be equal to zero. However, in our application, we are not confronted with this case. We can observe from figure 1 that there are, indeed, two eigenvalues that are significantly larger, however the other ones are not all equal to zero and some of them are negative. This contradicts our assumption of positive definiteness. This can be explained by Euclidean and Non-Euclidean errors. But overall, we choose two as dimensions for our configuration which is in line with what we expected from our model (2D map).

In figure 2, we have plotted the configuration obtained from the data and it looks coherent with Tunisia's map: maritime cities at the north and east borders, sahara cities at the centre and south of the graph.

The most interesting part originates from figure 4 where we can notice the differences between the real and simulated configurations: the time dissimilarities between the "developed cities" like Tunis, Sfax, Sousse are coherent with the real figure. However, the configuration is less accurate when it comes down to cities like Sidi Bouzid, Kasserine, Tozeur, Gafsa, Siliana, Jendouba and Kef. From our simulated configuration, these cities seem to be farther from Tunis for example than in a real scenario. Travelling to those cities is not coherent with their geographic distances. This can be explained by poor infrastructures. In the past 10 years, Tunisia's government spend significant efforts on developing a new highway that connects the capital Tunis with touristic cities like

Mahdia, Sousse, Nabeul, Medenine and Tozeur, which might explain the coherence of the configuration regarding the east region. We can conclude from that figure that there are clear disparities between east regions and west regions in terms of accessibility, like suggested.

5) Conclusions:

The discrepancies observed in the figures reinforce our thoughts on the different inequalities between the main regions in Tunisia. Travelling from one city to another in the west seems to be less convenient than in the east. This is mainly due to the lack of road infrastructure in the concerned regions (no highways and roads that are poorly maintained which deter a constant cruising speed for vehicles). The government should definitely focus its resources on the infrastructure development of the south and west regions, especially for cities like Jendouba, Kasserine or Gafsa. These cities are located in rural regions that suffer from high unemployment rates. Most of the job market is concentrated in coast cities that benefit from warm weather and tourism activities. Developing inter-cities exchanges will allow a decentralisation of power. It will allow economically disadvantaged regions to have better accessibility to opportunities that will promote growth. Also, Tunisia's economy is mainly based on tourism and those regions have the potential to attract a significant number of tourists, only if access is facilitated.

Distance-based method highlighted, when considering the travelling-time dissimilarity, the main disparities between road infrastructures in Tunisia's governorate ecosystem. However, the dataset is small; therefore, our results must be interpreted cautiously. Having access to a more accurate dataset, with travelling-time from each specific sub-city for example, would have allowed us to get a better understanding of the main issues that the current infrastructure department face, especially when we assumed that no other external factor than road infrastructure have an impact on the travel time.

Appendix:

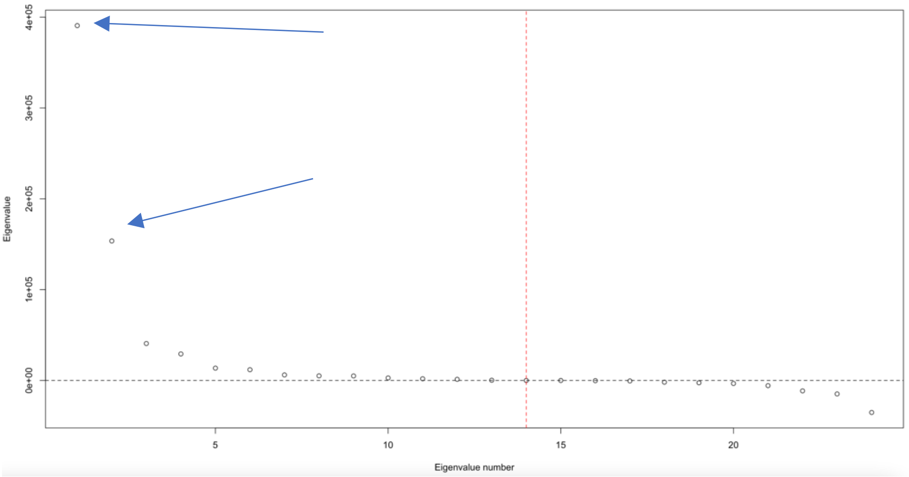

Figure 1: Plot of eigenvalue number against its value for time-dissimilarities matrix.

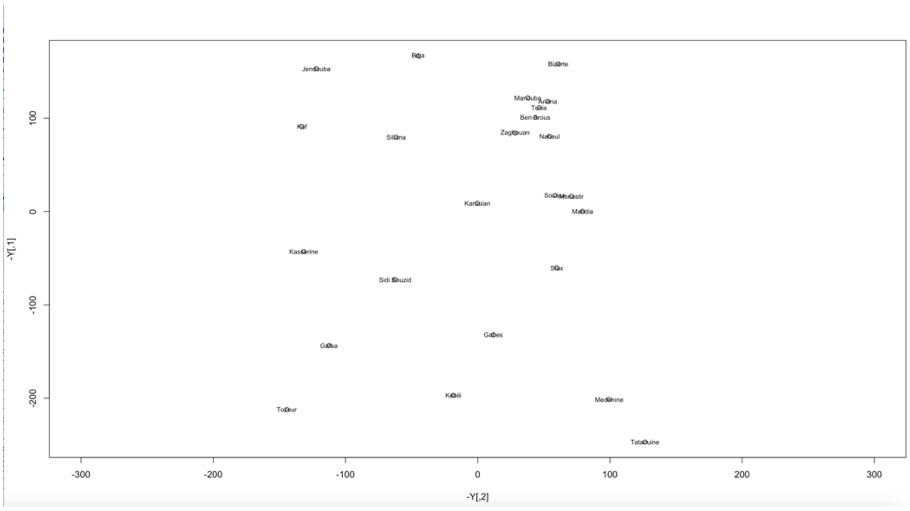

Figure 2: Plot of configuration obtained from time-dissimilarities matrix.

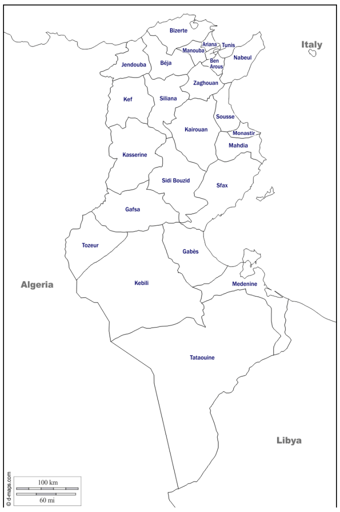

Figure 3: Tunisia map by governorate.

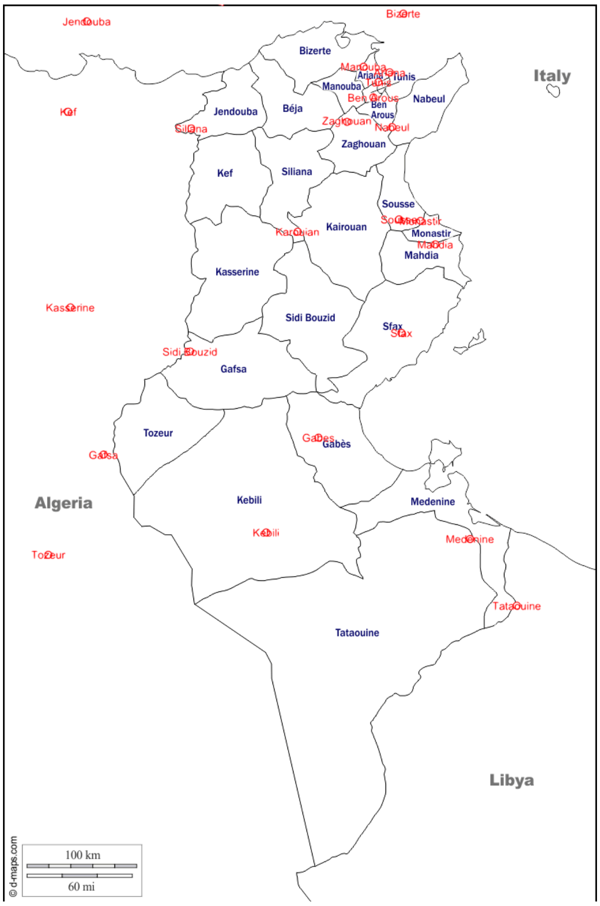

Figure 4: Real map (Blue) against configuration (Red) obtained from time-dissimilarities matrix