Group 2
P-MHD: Multivariate Methods

Aaryan Kaushik (2159244), Deo Byabazaire (2159254), Edmond Sacla Aide (2159278), Muhammad Bergas Nur Fayyad (2159123), Daniel R. Yildirim (2056569), Wouter Smeets (1849125)

May 23, 2022

1.Introduction

Precipitation and temperature are among the most important elements to describe the climate in a certain area. This study aims to find similarities and dissimilarities among Canadian cities regarding the precipitation regime.

1.1 Data

The precipitation is the main data available in the Canadian weather dataset. It contains the average daily rainfall (mm/day) for the 365 days in the year and for 35 Canadian cities. In addition, this dataset includes other information such as the regions, provinces and coordinates. We have added a new column indicating the season for later use.

1.2 Research question

The main objective of this project is to sort out the variability between Canadian cities in terms of the average daily rainfall. It addresses the questions of which cities have similar precipitation patterns, and those with dissimilar patterns? Also, how can the differences between the cities be described (i.e., in what sense do cities differ)?

2.Functional Data Analysis (FDA)

2.1 Introduction

The Multi-Dimensional Scaling (MDS) is one of the multivariate methods that aims to find a low-dimensional representation, say k dimensional space, of n data points such that the distances between the n points in the k-dimensional space are a good approximation of a given squared distance matrix, say D_x. Hence, the research question leads us to consider it an appropriate method to address the objective of this project.

2.2 Transformation to functions

The precipitation dataset used in this study consists of $n = 35$ rows (cities) and $p = 365$ columns (days). This is high dimensional data (as $p >> n$). This structure makes it easier to look at the raw data table. A

functional data analysis (FDA), which assumes that functions are considered as observations, has been used to analyze this data. For each city, the observations of the precipitation are in the function of days.

To take this approach, it is necessary to first convert the data entries for each cities to a single function. As the aim of Singular Value Decomposition (SVD) is to transform the dataset to a lower-dimensional dataset called the parameter space, it therefore gives a matrix where $q < p$ is less than the original data. Consequently, each city will have its set of q parameter estimates, and so an $(n \times q)$ data matrix can be constructed. To give a meaningful interpretation to the results at the end of our analysis, the solution needed to be back-transformed from the parameter space to the functional space.

2.3 Multidimensional Scaling of Functions

Denote $Y_i(t)$ the outcome of observation $i = 1, \ldots, n$ as the average daily precipitation for cities i at time t between January 1 and December 31. For observation i there are data on times t_{ij}, $j = 1, \ldots, p_i$.

Consider the non-linear model

$$Y_i(t_{ij}) = f_i(t_{ij}) + \varepsilon_{ij} = \sum_{k=0}^{m} \theta_{ik}\phi_k(t_{ij}) + \varepsilon_{ij} \approx \sum_{k=0}^{m} \theta_{ik}x_{ijk} + \varepsilon_{ij}; \;\; i = 1, \ldots, n; j = 1, \ldots, p_i; k = 0, \ldots, m$$

where $f_i(\cdot)$ is a smooth function and ε_{ij} i.i.d. with mean 0 and constant variance σ^2. The $\phi_k(\cdot)$ form a set of orthonormal basis functions.

The days were re-scaled to [0,1] interval to avoid numerical problems.

Next basis functions, with either a polynomial basis or Fourier basis were compared such that the one that gave a better fit to the dataset would be chosen. Thus, a small simulation was done to compare the goodness of fit of the linear regression using the polynomial basis with the one using Fourier basis. The comparison was based on the adjusted R-squared (adj-R^2) and the mean square error (MSE).

The Fourier basis seemed to have a better prediction power compared to the polynomial basis. Also from some previous studies by Tsai et al. (2016) & Adams et al. (2018) it was reported that the Fourier basis is often a better choice than the polynomial basis.

Next the degree m of the Fourier basis has been selected. For a given city, different values of m (1, 5, 10, 15, 16, 17, 18, 19, 20) were applied. It was observed that from $m = 17$ and above, there was no additional variation in the precipitation trend (see code). Hence, for the construction of the theta matrix, Fourier basis with degree 17 was used.

The estimation of the theta parameters for the chosen 35 cities was made.

Finally, the statistical model for country i in matrix notation is as follows:. $\boldsymbol{Y}_i = \boldsymbol{\theta}_i^t \boldsymbol{X}_i + \boldsymbol{\varepsilon}_i$ for $i = 1, ..., 35$

where $\boldsymbol{Y}_i$ is the vector with the outcomes of observation i (one for each day tij), $\boldsymbol{\theta}_i$ the vector with the θ_{ik} (one for each basis function k), $\boldsymbol{X}_i$ the matrix with the $xijk$ (days j in the rows, basis function index k in columns), and $\boldsymbol{\varepsilon}_i$ the vector with the i.i.d. error terms.

The parameters θ_{ik} can be estimated by means of least squares.

2.3 Multidimensional Scaling of Functions

Define $\hat{\boldsymbol{\theta}}_i$ as the vector with the parameter estimates. Then, the estimates for all cities can be collected into a single new $n \times (m + 1)$ data matrix $\boldsymbol{\Theta}$; where the i^{th} row of $\boldsymbol{\Theta}$ is $\hat{\boldsymbol{\theta}}_i^t$. Since $\boldsymbol{\Theta}$ has the structure of an ordinal data matrix, the MDS was applied to $\boldsymbol{\Theta}$, so that a 2-dimensional plot can be constructed with each point representing a Canadian's city. The distances between the points in the 2-dimensional MDS space are approximations of the distances between the rows of $\boldsymbol{\Theta}$. Hence, it can be interpreted as distances between the precipitation functions.

The MDS starts from the truncated SVD of $\boldsymbol{\Theta}$ (or after column-centering),

$$\boldsymbol{\Theta}_k = \boldsymbol{U}_k \boldsymbol{D}_k \boldsymbol{V}_k^t$$

(note that the index k now refers to the number of components in the truncated SVD and not to the index of the basis functions).

Then the column centering and SVD were applied on the theta matrix.

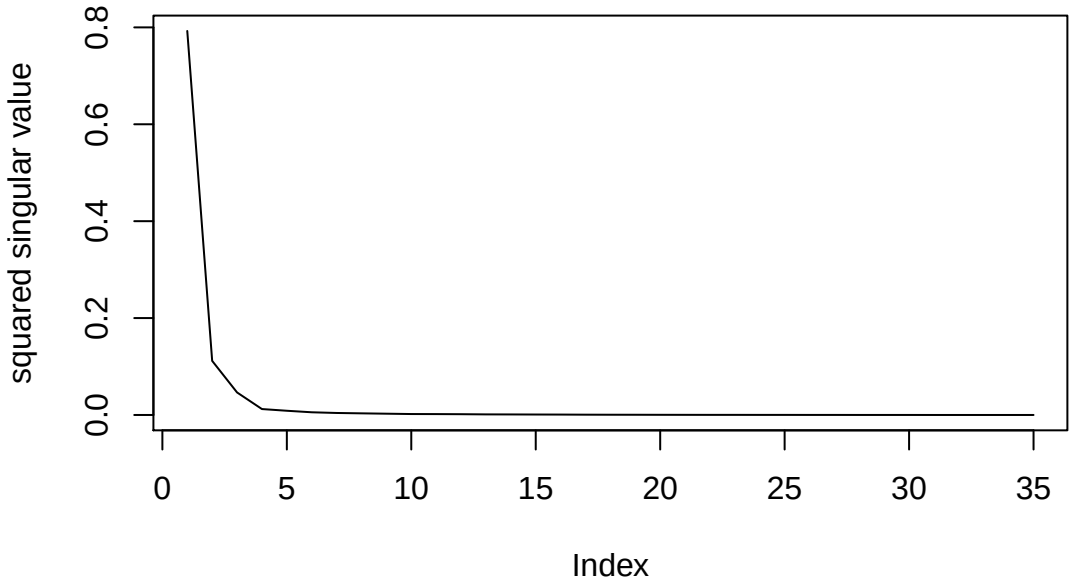

Figure 1: Squared Singular Value vs Index

Figure 1 indicates that the first few dimensions capture most of the information in the Θ matrix. We can see a rapid decrease of the singular values from the larger values. This implies the information in terms of variability contained in the squared singular values decreases rapidly and eventually tends to zero.

3.Functional biplot

The interpretation of the functional biplot is easier after transforming the SVD to the original function space.

The fitted model for all n cities may then be simultaneously written as

$$\hat{\boldsymbol{Y}} = \boldsymbol{\Theta} \boldsymbol{X}^t$$

with $\hat{\boldsymbol{Y}}$ the $n \times p$ matrix with ith row $\boldsymbol{Y}_i^t$, and $\boldsymbol{X}$ the $p \times (m+1)$ matrix $\boldsymbol{X}_i$ as defined before (note that all $\boldsymbol{X}_i$ are equal because for all cities i the measurements were obtained at the same p time points (years)).

After substituting $\boldsymbol{\Theta}$ with its truncated SVD (after k terms), the simplified model function can have the form:

$$\hat{Y}_{ki}(t) = \sum_{j=1}^{k} \sum_{r=0}^{m} z_{kij} v_{rj} \phi_r(t)$$

where z_{kij} is the (i,j)th element of $\boldsymbol{Z}_k$ and vrj is the (r,j)th element of $\boldsymbol{V}_k$

The following figure presents the functional-plot. The origin of the graph for both dimensions corresponding to the average precipitation function starts at (0,0). This is because $\boldsymbol{\Theta}$ is column-centered. Figure 2 suggests that both in the first and second dimension, some cities have negative score and some have positive score.

In the first dimension, Pr Rupert, St Johns, Sydney, Halifax, Yarmouth and Varcouver have higher negative score. These cities are opposed to Resolute, Inuvik, Whitehorse, Regina, Pr Albert and Calgary that present positive score.

In the second dimension, Victoria and Vancouver have negative score whereas Quebec, Sherbrook, Thundi and Winnipeg seem to have significant positive score.

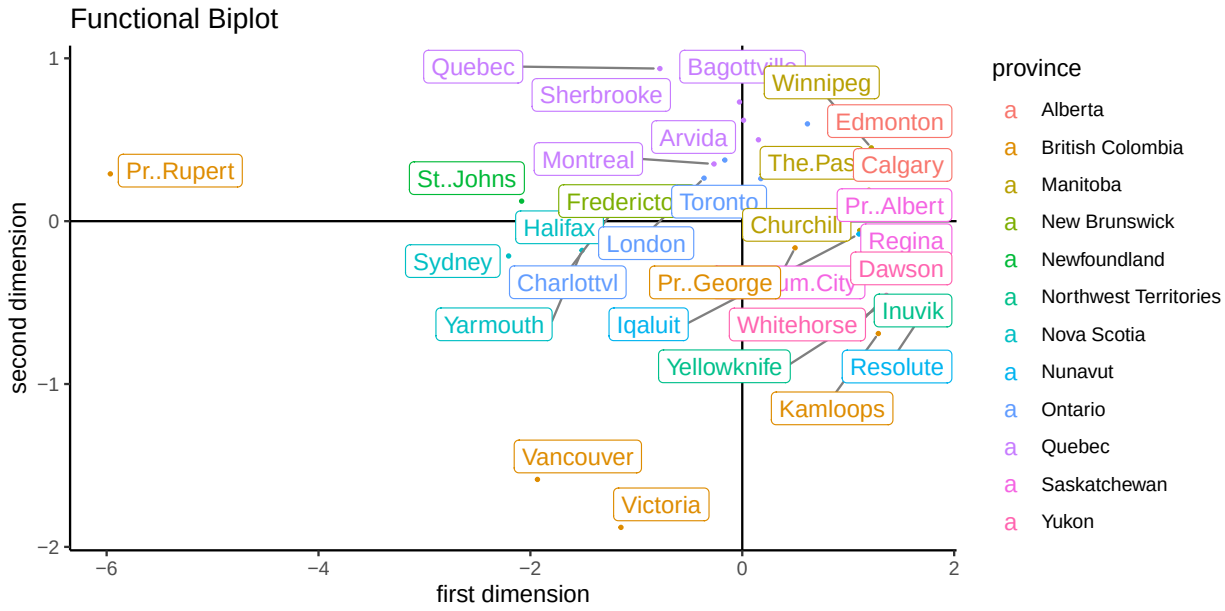

Figure 2: Functional biplot of Precipitation of Canadian Cities

From the functional biplot (Figure 2), although there were no other important geographic and topological features of cities that influence precipitation included, such as latitude, proximity to large bodies of water, and location relative to mountain ranges, we see it can be seen that most cities in the same provinces have a similar rainfall regime.

For a further interpretation of the functional biplot, the SVD was transformed back to the original function space.

Based on Figure 3, it was possible to conclude that cities with a large negative score (red line) have high and decreasing precipitation average from late winter to the middle of summer and increasing precipitation average from mid summer until the end of autumn. Then cities with negative scores have high precipitation compared to the average. The average precipitation for cities with large positive scores (blue line) increases from late winter until mid summer and decreases until early winter. These cities with positive scores have low precipitation compared to the average. Whereas, the black line represents average rainfall for all cities and shows similar rainfall patterns for all the seasons and there are no clear visible peaks.

Figure 4 shows that cities with negative scores (red line) in the second dimension have a low precipitation average compared to the average from spring to mid autumn. In contrast, cities with a positive score in that dimension have higher precipitation than the average. Moreover, we observe an opposite trend for the rest of the year (mid autumn to winter).

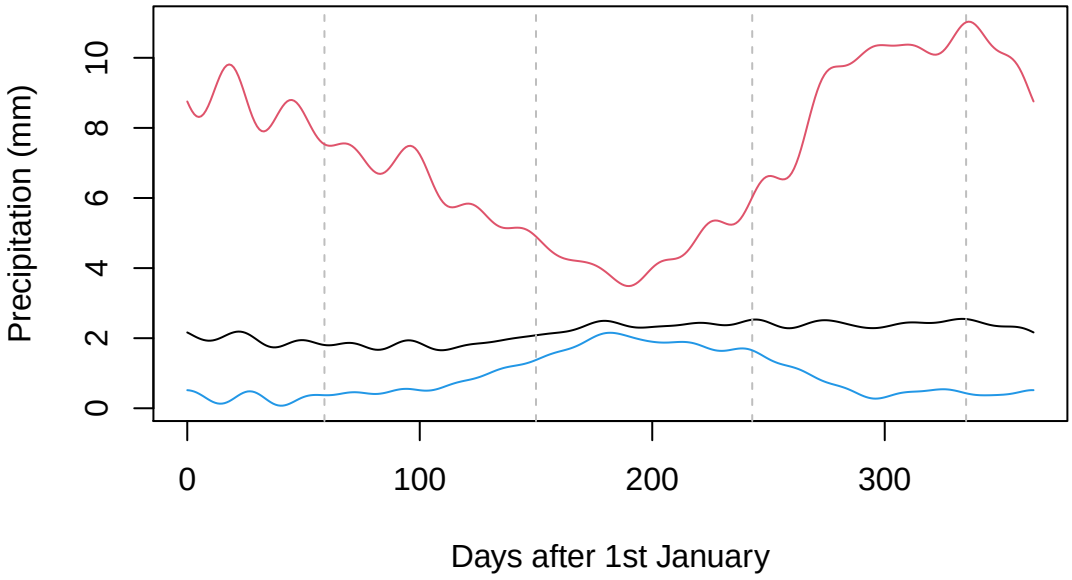

Figure 3: Back-transformed to the original function space in the First Dimension

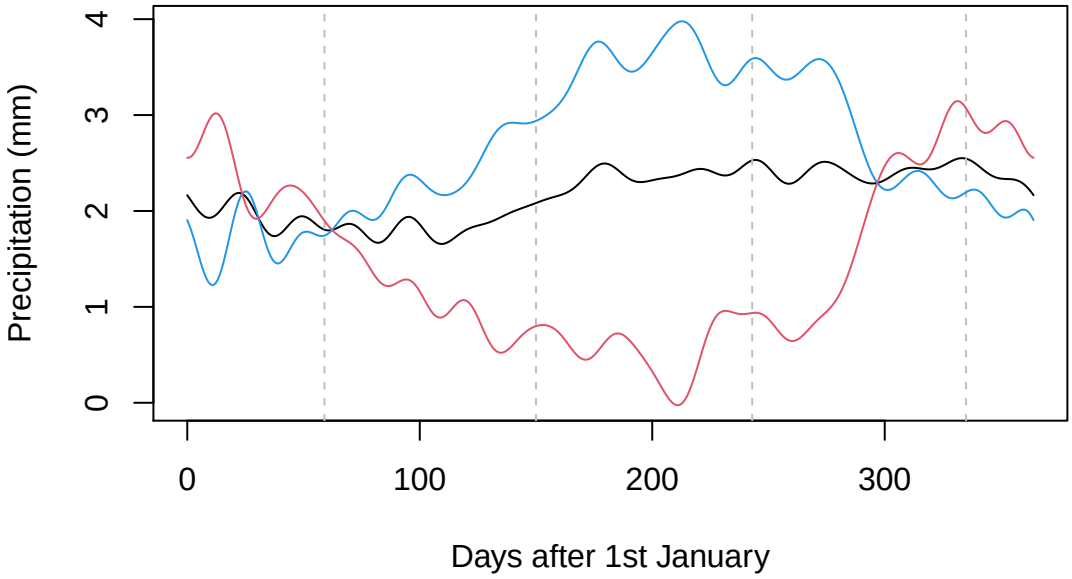

Figure 4: Back-transformed to the original function space in the Second dimension

Conclusion

Considering the information from the score plot and the trend plots i.e first dimension plot that includes Pr Rupert, St Johns, Sydney, Halifax, Yarmouth and Varcouver have a daily precipitation higher than the average during the whole year. But this precipitation decreases from late winter to the middle mid summer and increases the rest of the year. In Resolute, Inuvik, Whitehorse, Regina, Pr Albert, and Calgary, the daily precipitation is lower than the average. It increases from late winter to mid summer and decreases the rest of the year. In Victoria, the average precipitation is lower than the average from spring to mid autumn but higher than the average for the rest of the year. Furthermore, in Quebec, Sherbrook, Thundi and Winnipeg, the daily precipitation is higher than the average from spring to middle mid autumn.

References

Tsai, Cho-Liang, Wei Tong Chen, and Chin-Shiang Chang. "Polynomial-Fourier series model for analyzing and predicting electricity consumption in buildings." Energy and Buildings 127 (2016): 301-312.

2022. Cs.Princeton.Edu. https://www.cs.princeton.edu/courses/archive/fall18/cos324/files/basis-functions.pdf.