Project Multivariate Methods Report Group 8

Group 8: Meynen Frederik, Kerga Meseret Assefa, Volkova Anastasia

23/5/2022

1 Introduction

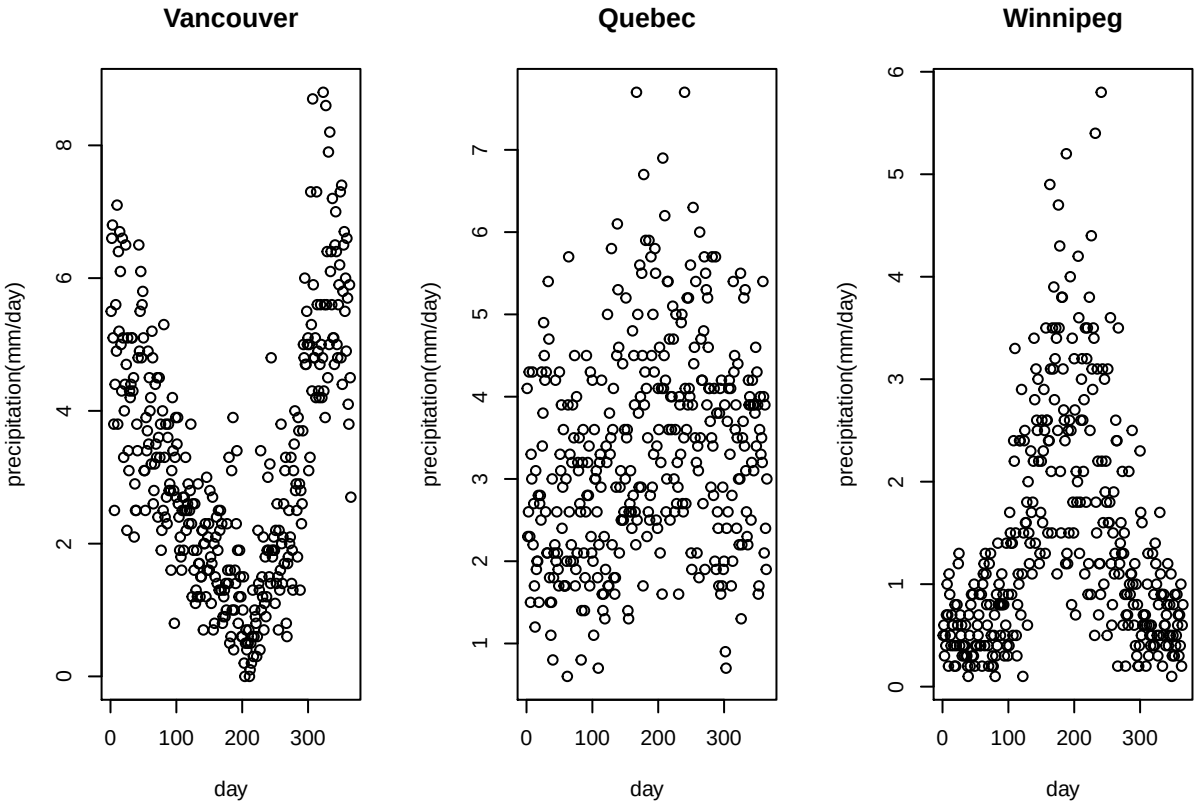
1.1 Data & Data exploration

The available dataset contains the precipitation data of the following 35 Canadian cities:

```
##  [1] "St. Johns"    "Halifax"      "Sydney"       "Yarmouth"     "Charlottvl"
##  [6] "Fredericton"  "Scheffervll"  "Arvida"       "Bagottville"  "Quebec"
## [11] "Sherbrooke"   "Montreal"     "Ottawa"       "Toronto"      "London"
## [16] "Thunder Bay"  "Winnipeg"     "The Pas"      "Churchill"    "Regina"
## [21] "Pr. Albert"   "Uranium City" "Edmonton"     "Calgary"      "Kamloops"
## [26] "Vancouver"    "Victoria"     "Pr. George"   "Pr. Rupert"   "Whitehorse"
## [31] "Dawson"       "Yellowknife"  "Iqaluit"      "Inuvik"       "Resolute"
```

For each of the 35 cities we have 365 measurements, one for each day of the year. In addition to these measurements we also have metadata to our disposal. This metadata contains the region, province and coordinates of the Canadian cities.

As a first exploration of the data we plot the precipitation data of 3 of the Canadian Cities



1.2 Research Question

The objective is to discover which cities have similar precipitation patterns, and which have dissimilar patterns. We need a 2-dimensional graph that shows each city as a point, such that cities with similar precipitation patterns are close to one another. We also want to understand the difference in rainfall patterns: in what sense do they differ? For example, in the example plots shown above Vancouver hade more precipitation in winter, Winnipeg had more in summer and Quebec had similar precipition all year round.

1.3 Functional Data Analysis

1.3.1 Introduction

In this homework we will make use of the functional data analysis (FDA) approach.

In FDA we consider functions as observations. For example, for each of the 35 cities we have $p = 365$ observations on a precipitation function. To make this approach work we first have to transform the 365 observations to a single function. This function will contain fewer parameter estimates than the original number of observations, say $q < p = 365$. Thus each city will have its set of q parameter estimates, and thus an nÖq data matrix can be constructed. These parameter estimates form now the input for the Multidimensional Scaling (MDS). To give a meaningful interpretation to the results, at the end we will back-transform our solution from the parameter space to the function space.

1.3.2 Transformation to functions

In this homework we will make use of the $poly()$ function in R to transform a vector with the days at which measurements are available, to a matrix. For a given city i, the number of rows of the matrix equals the

number of days, and each column corresponds to a basis function. The (j,k)th element of the matrix equals the kth basis function evaluated in the jth day (t_{ij}). Let $x_{ijk} = \phi_k(t_{ij})$ denote this element.

This means we can write a statistical model for the measurements of city i at time j: $Y_i(t_{ij})$,

$$Y_i(t_{ij}) = \sum_{k=0}^{m} \theta_{ik} x_{ijk} + \epsilon_{ij}$$

.

For a given city i, this has the structure of a linear regression model with outcomes $Y_i(t_{ij}), with j = 1, \ldots, n$, and $q = m + 1$ regressors x_{ijk}.

This model can also be written in matrix notation:

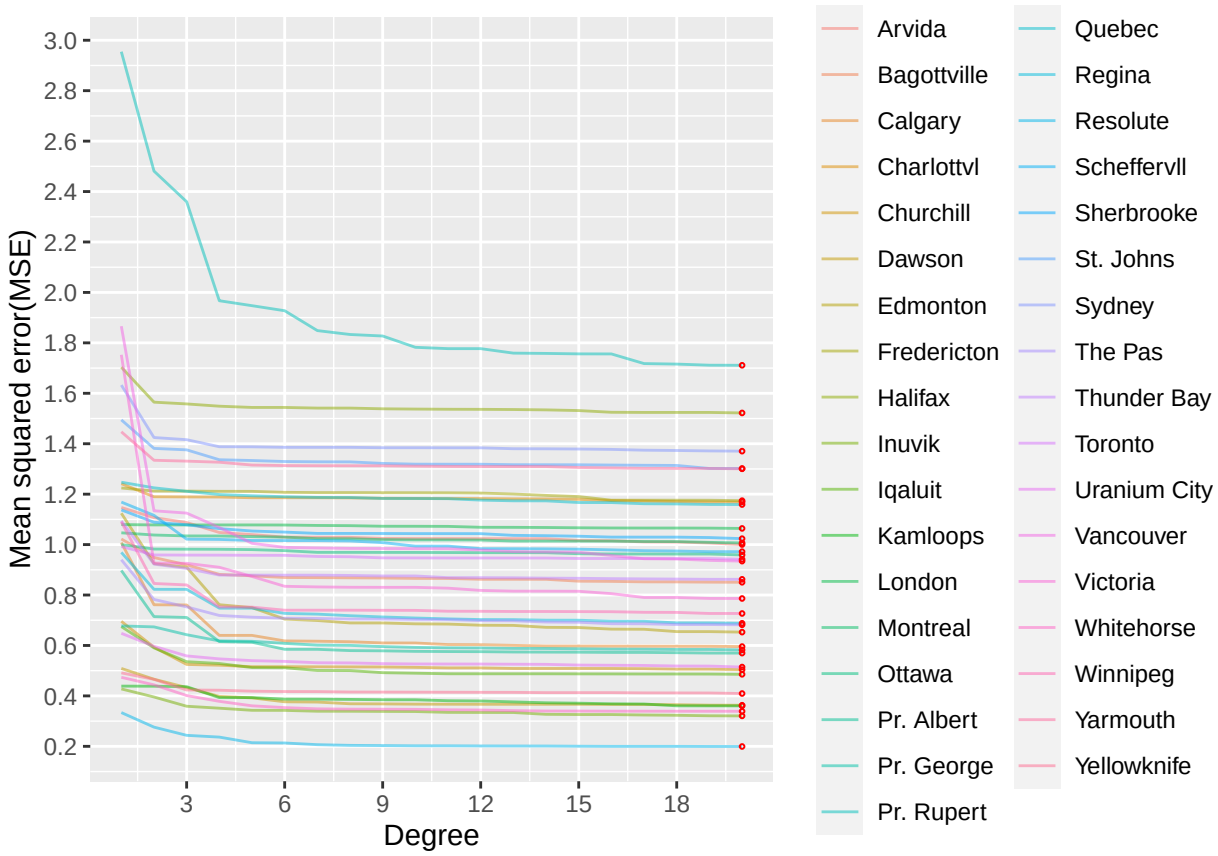
$$Y_i = \theta_i^t X_i + \epsilon_i$$

Thus, first we must choose an appropriate number of basis functions m. To identify this number we look at the amount of basis functions that would give a low mean square error. To calculate the MSE of a single city we use the following code:

```
days <- 1:365
days <- (days-min(days))/(diff(range(days))) ## rescaling


# selecting m degree using Mean square error(MSE)
df_mse = data.frame()  # mse storage created for each city and degree
ncity = colnames(da)

for (d in 1:20){ # Go through all possible values for m
  phi<-poly(days,degree=d) #  polynomial degree d
  for (c in ncity){
    m.c<-lm(da[,c]~phi) #  parameter estimation using linear regression
    pred.c = predict(m.c, phi) # prediction made for each city and for a given degree

    mse<-sqrt(mean((da[,c] - pred.c )^2)) #  mse calculated for each city and for a given degree
    tmp_df = data.frame(city= c, d=d, mse=mse) # store values of city, d and mse as data frame
    df_mse = rbind(df_mse, tmp_df) # store the value in df_mse
  }
}

min_mse<- df_mse %>%
dplyr::select(city, mse, d) %>%
dplyr::group_by(city) %>%
dplyr::summarise(mse= min(mse))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
ggplot(df_mse,aes(x= d, y= mse, color = city))+
  geom_line(alpha=0.5)+
  labs(y ="Mean squared error(MSE)", x = "Degree")+
  geom_point(data = min_mse, color = "red",
  shape = 1, size = 0.5)+
  scale_y_continuous(breaks = seq(0, 3, by = 0.2))+
  scale_x_continuous(breaks = seq(0, 20, by = 3))
```



This graph shows us that the minimum MSE for all cities occurs at $m = 20$ basis functions, which is not surprising, we expect the MSE to go down for higher numbers of m. But there is barely any change in MSE from $m = 15$ basis functions onwards. This is why we will continue with $m = 15$ basis functions.

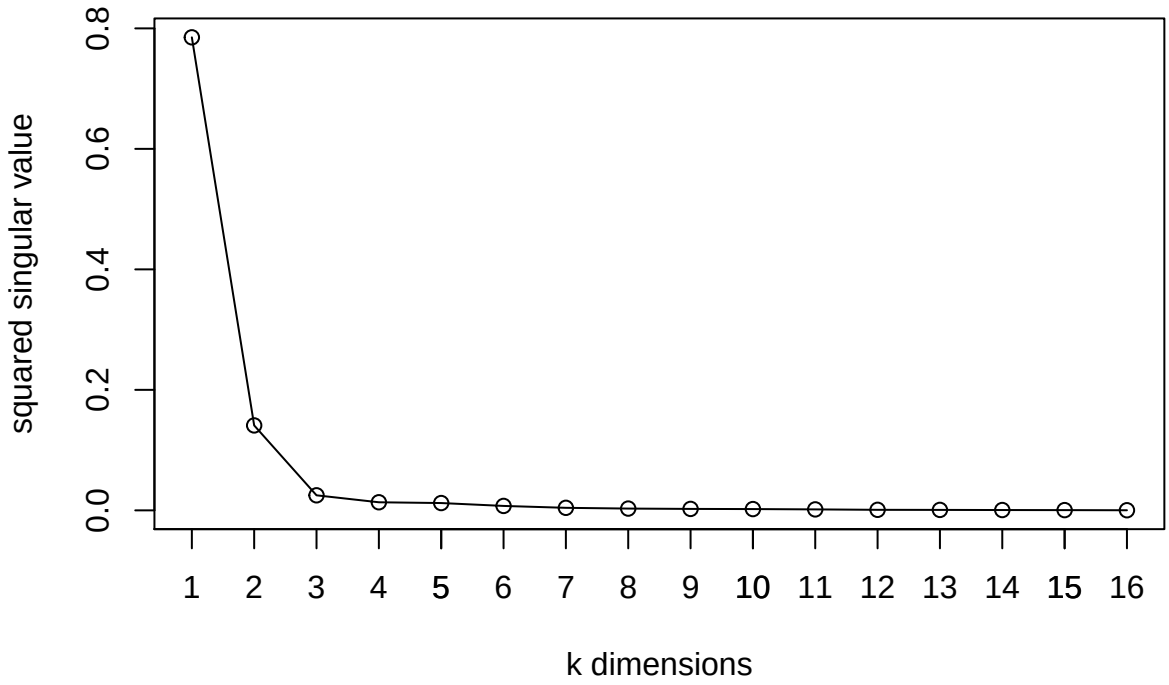
Now we can perform the actual transformation to functions.

The resulting Θ matrix contains the θ parameter estimates for all 35 cities and all basis functions ($16 = m+1$ because we add a column for the intercept). The i-th row of Θ is the vector $\hat{\theta}_i^t$ that contains the parameter estimates. This Θ matrix is now considered to be our data matrix and we can now apply MDS.

1.3.3 Multidimensional Scaling of Functions

First we will column center the Θ matrix, after which we start the actual single value decomposition on the centered Θ Matrix

To check how many dimensions we would need to adequately approximate the data matrix we plot a scree plot:




```
## [1] 0.9260936
```

This scree plot shows that most (>92%) of the variability is retained with just 2 dimensions. Because of this we will use only $k = 2$ dimensions.

This means that in the fitted model:

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

the *Theta* matrix can be substituted by its truncated SVD.

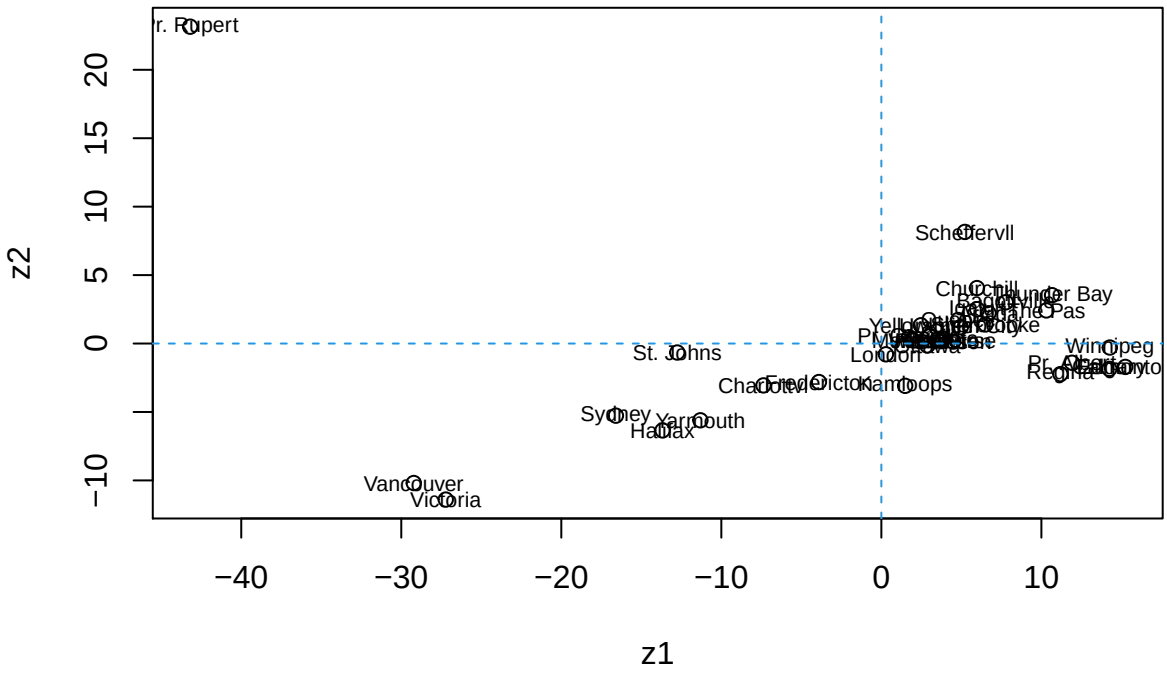
$$\hat{Y}_2 = \Theta_2 X^t = U_2 D_2 V_2^t X^t = Z_2 V_2^t X^t$$

2 Results

2.1 Functional biplot

We started by using the truncated SVD ($U_k and D_k$) to calculate the scores (Z_k)

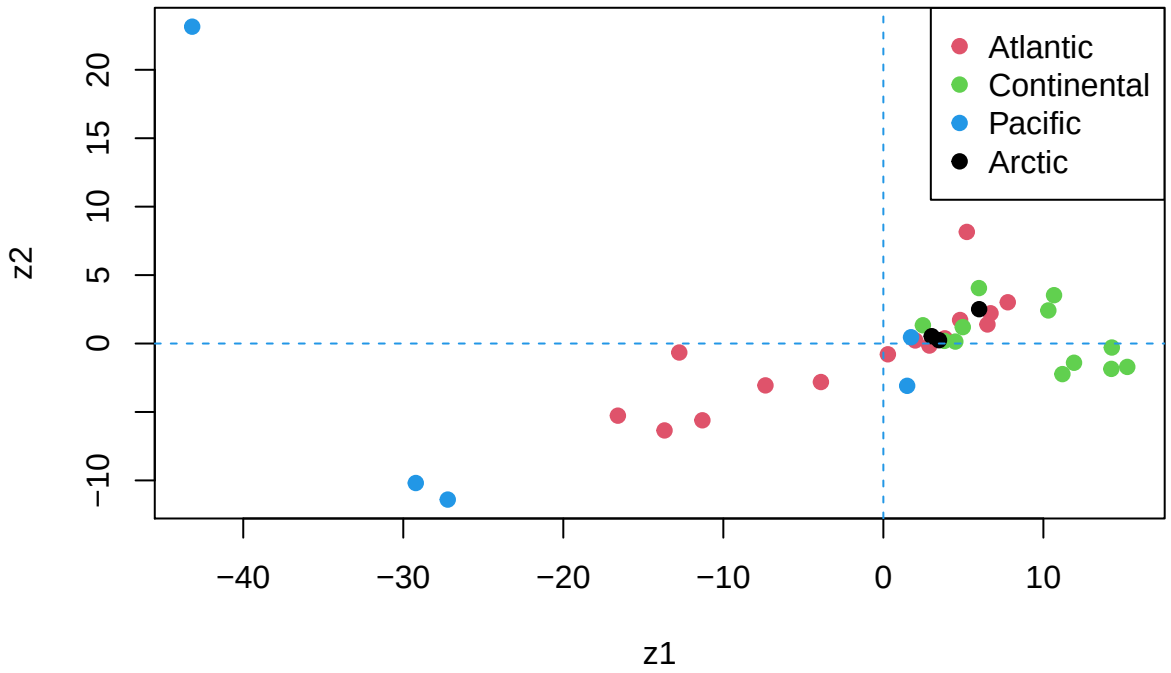
Next we plotted the scores.



Since we column-centered Θ the origin corresponds with the average precipitation function. As we can see there are some cities with large negative values along the first dimension (Pr. Rupert, Vancouver and Victoria) and some with large positive (Pr. Rupert) and large negative (Vancouver, Victoria) scores along the second dimension.

Note that there is also a very clear outlier (Pr. Rupert). The precipitation pattern in this city is thus very different from the other cities. It is possible that this outlier influences the result because of its relative large distance to all other points in the original space. This might also explain the slope seen in the other data points. To make sure that this city didn't influence our results we performed the SVD again without Pr. Rupert. At a first glance excluding this outlier doesn't seem to change our interpretation much. More information can be found in the appendix.

To better understand the position of the different cities we will use the metadata to differentiate between regions.



This graph shows that cities in the Contintental region have similar precipitation patterns (they all have a positive Z1 and a Z2 close to the average). The same is true for cities in the Arctic region. Cities in the Atlantic region have an interesting distribution on this biplot. When Z1 is negative, Z2 is negative as well. When Z1 is positive, Z2 tends to be positive as well. Cities in the Pacific region show the most erratic pattern. Three of the cities in the Pacific region (namely Pr. Rupert, Vancouver and Victoria) have the lowest Z1 and the most extreme Z2 values while the 2 other cities (Pr. George, Kamloops) are very close to the functional mean.

2.2 Backtransformation

To beter understand the meaning behind the values of the Z1 and Z2 scores we will backtransform the SVD to the original function space. We do this because constructing a biplot from the SVD of the Θ matrix would create a hard to interpret graph, since the arrows would point to the different basis functions.

Note that from the previous paragraph we had that the approximate model fit was

$$\hat{Y}_2 = \Theta_2 X^t = U_2 D_2 V_2^t X^t = Z_2 V_2^t X^t$$

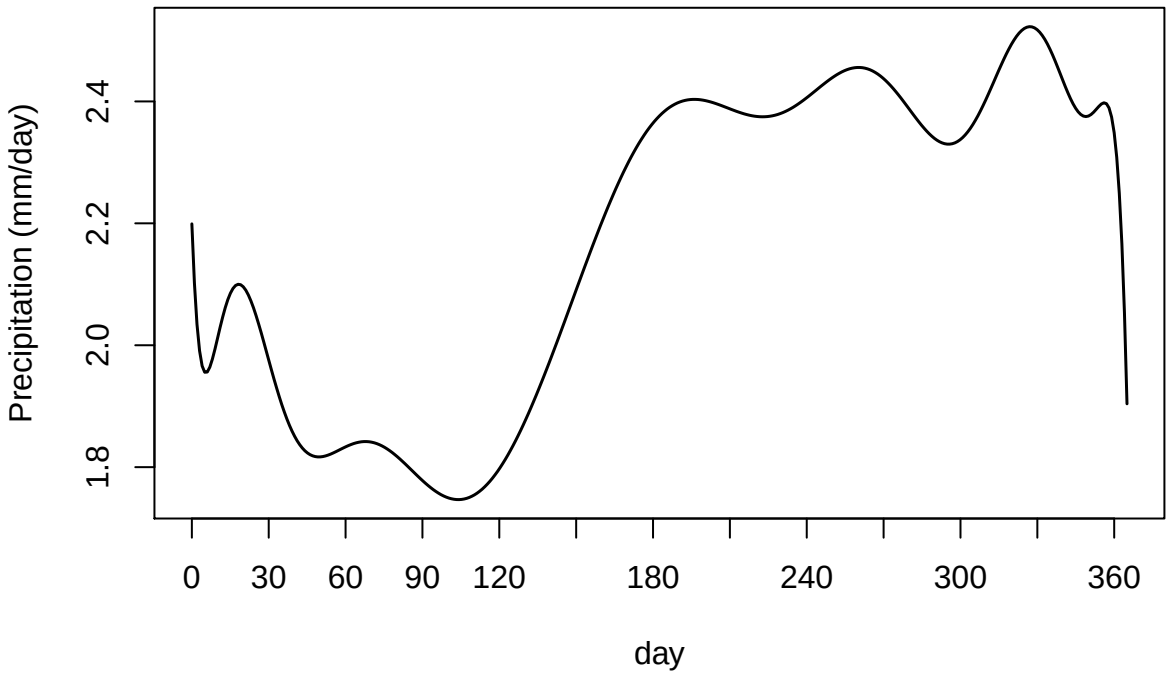
The matrix X is the matrix that connects the basis functions to the /*theta*-parameters. This model fit can thus also be written as:

$$\hat{Y}_{2ij}(t) = \sum_{d=1}^{2} \sum_{r=0}^{15} z_{2id} v_{rd} \phi_r(t)$$

Since we column-centered the Θ we took out the average precipitation function, so it is better written as:

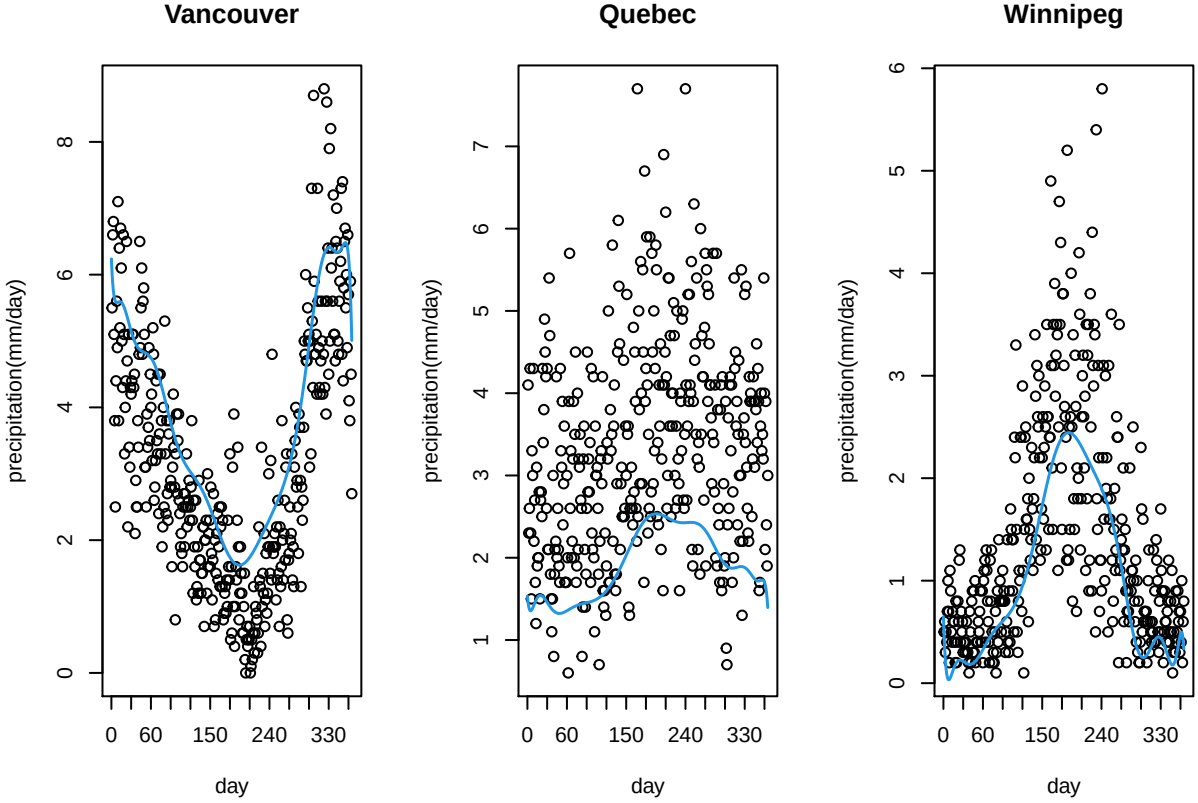
$$\hat{Y}_{2ij}(t) = \bar{Y}(t) + \sum_{d=1}^{2} \sum_{r=0}^{15} z_{2id} v_{rd} \phi_r(t)$$

This formula now gives us the ability to look at the effect of the scores.

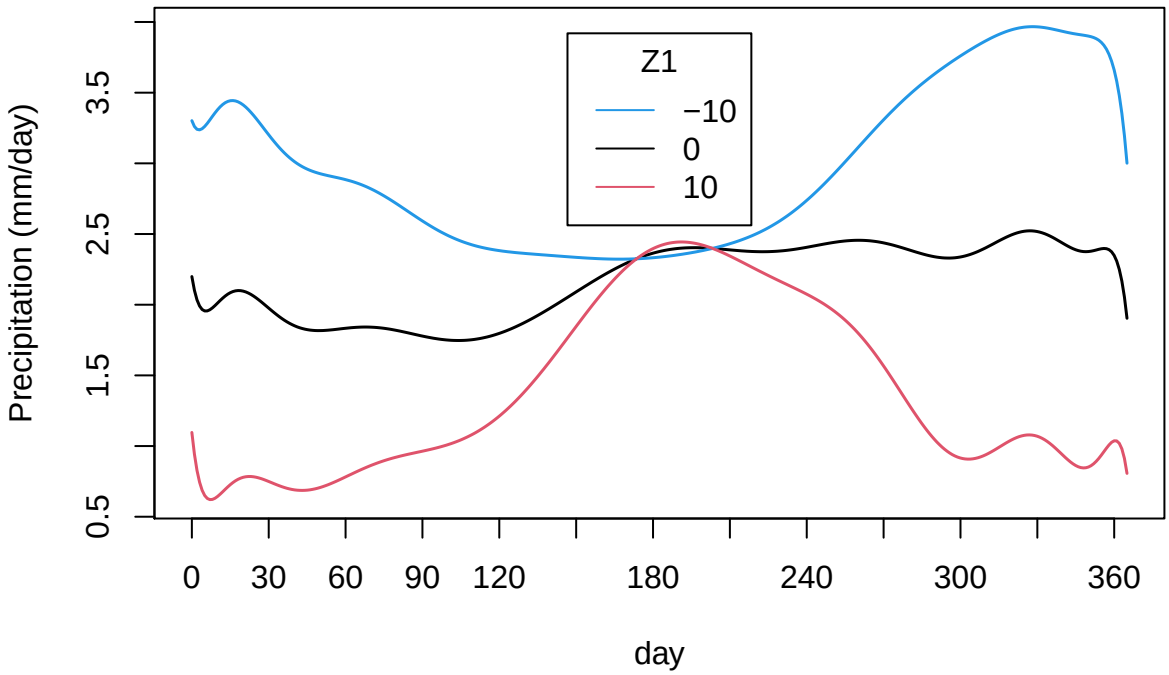


The mean curve shows that most of the precipation for the average city (close to the origin) occurs in the second half of the year (days 150-onwards).

To check the backtransformation we will reprise the three example plots of paragraph 1.1 with $\hat{Y}_{2i}(t)$ added to it.

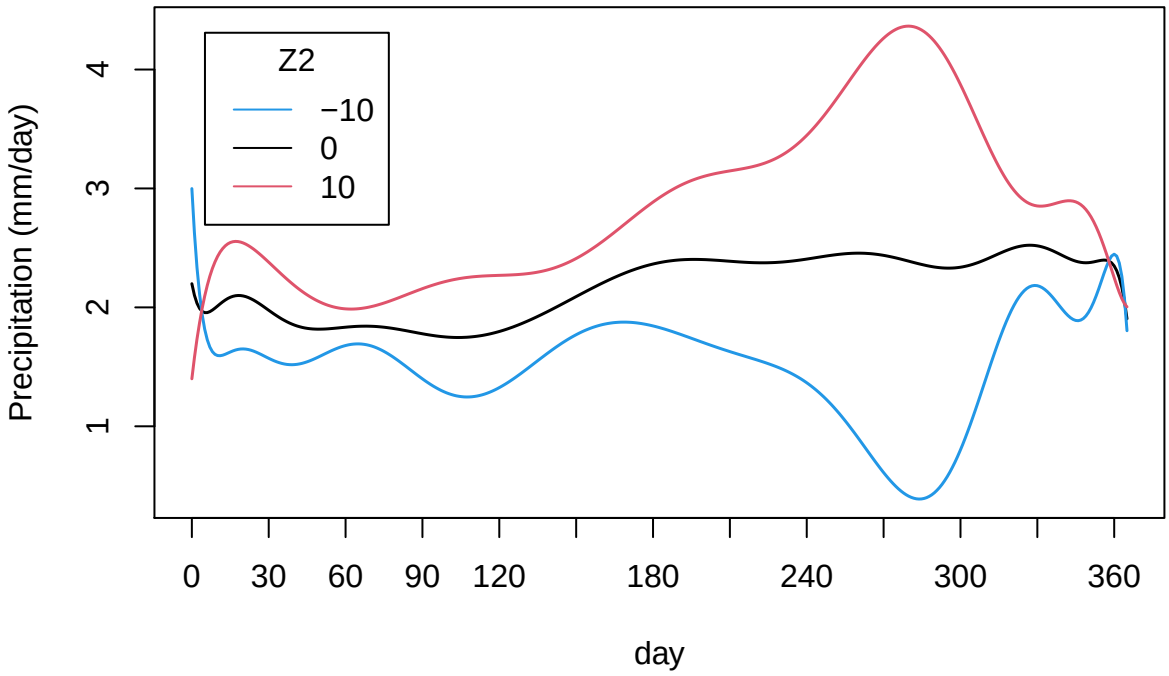


To interpret the Z1 scores we will plot $\hat{Y}_{2i}(t)$ with z_{2i1} variating between -10, 0 and 10 and z_{2i2} fixed at 0.



This graph shows that cities with a higher score on Z1 have, in comparison to the mean, low precipitation in the beginning (days 0-100) and end (days 300-365) of the year. Example cities are Calgary, Winnipeg and Edmonton. Cities with a negative Z1 score follow an inverted pattern with high precipitation in the beginning and end of the year. Examples of such cities are Pr. Rupert, Vancouver and Victoria.

To interpret the Z2 scores we will produce a similar plot of $\hat{Y}_{2i}(t)$ with z_{2i2} variating between -10, 0 and 10 and z_{2i1} fixed at 0.



This graph shows that cities with a high score on the Z2 have higher precipitation overall compared to the mean but especially during autumn (between days 240 and 320). Examples of such cities are Pr. Rupert, Scheffervll and Churchill. Negative scores follow an inverted pattern with lower precipiation overall compared to the mean but especially lowered during autumn. Example of such cities are Victoria, Vancouver

and Halifax (The Continental and Arctic cities in general).

3 Conclusion

Our analysis of the Canadian weather data shows that precipitation in cities differs mainly along 2 lines. The main difference is the precipitation in the beginning and end of the year. The second difference is precipitation in autumn. On average there is more precipitation in the second half of the year.

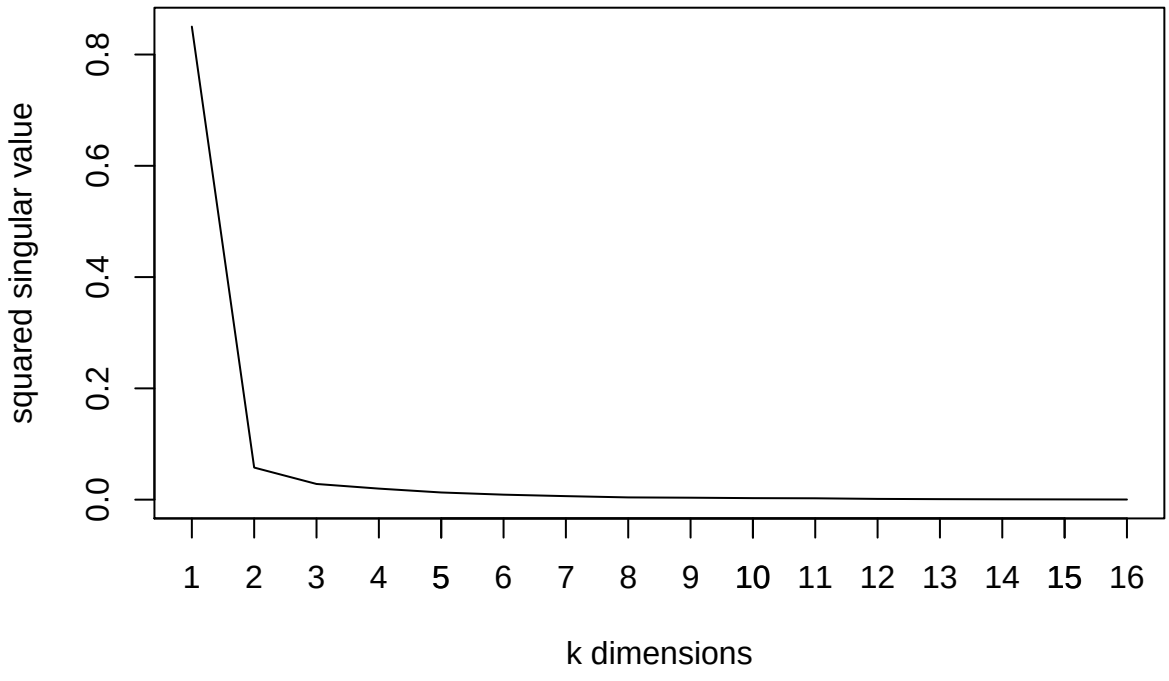
We saw that cities in the Continental region have similar precipitation patterns, with low precipitation in the beginning and end of the year, and close to average precipitation in autumn. Arctic cities follow a similar pattern.

Cities in the Atlantic region have either high precipitation in the beginning and end of the year and low in autumn, or the other way around. The precipitation patterns of the Pacific Region were erratic.

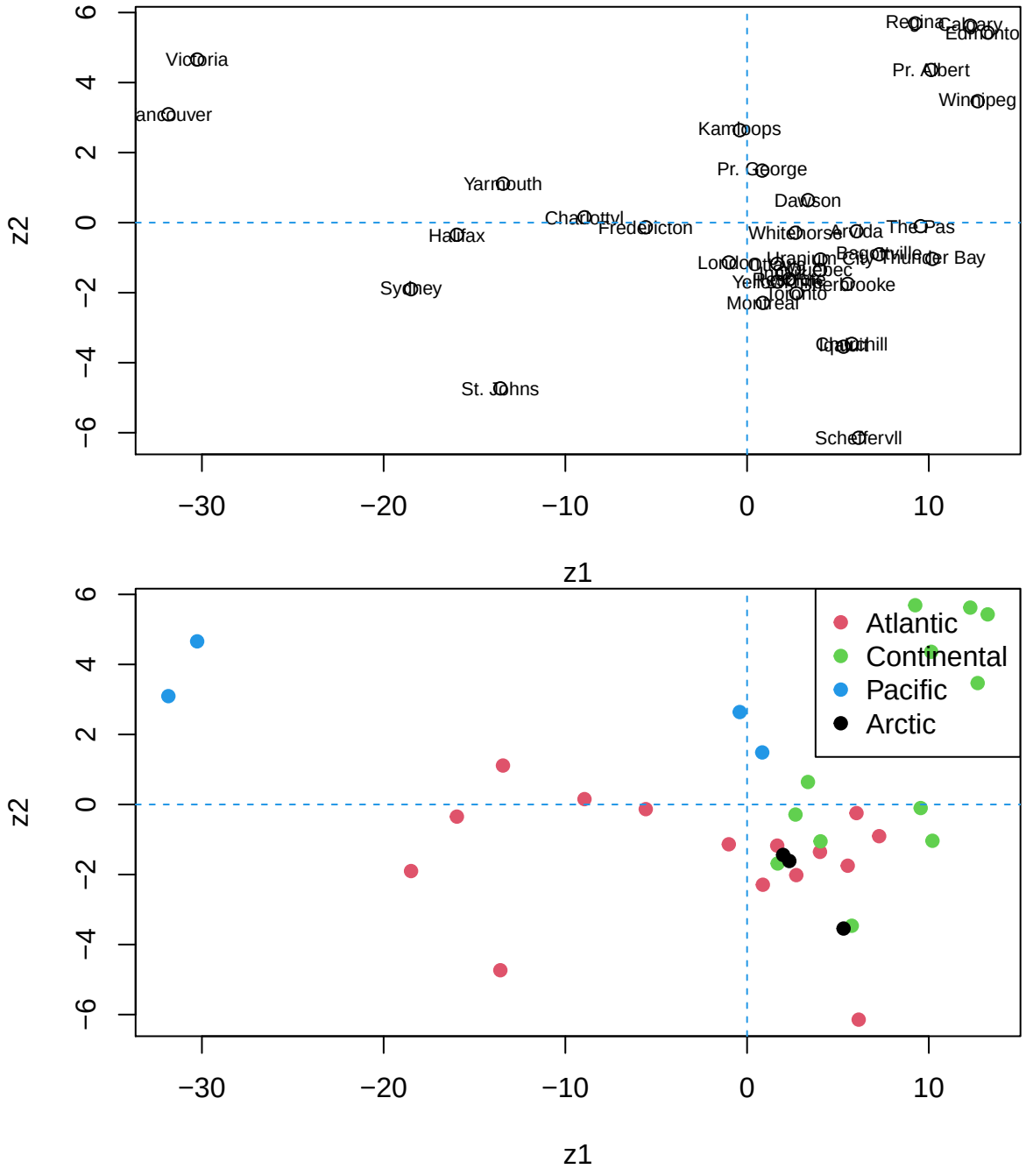
4 Appendix

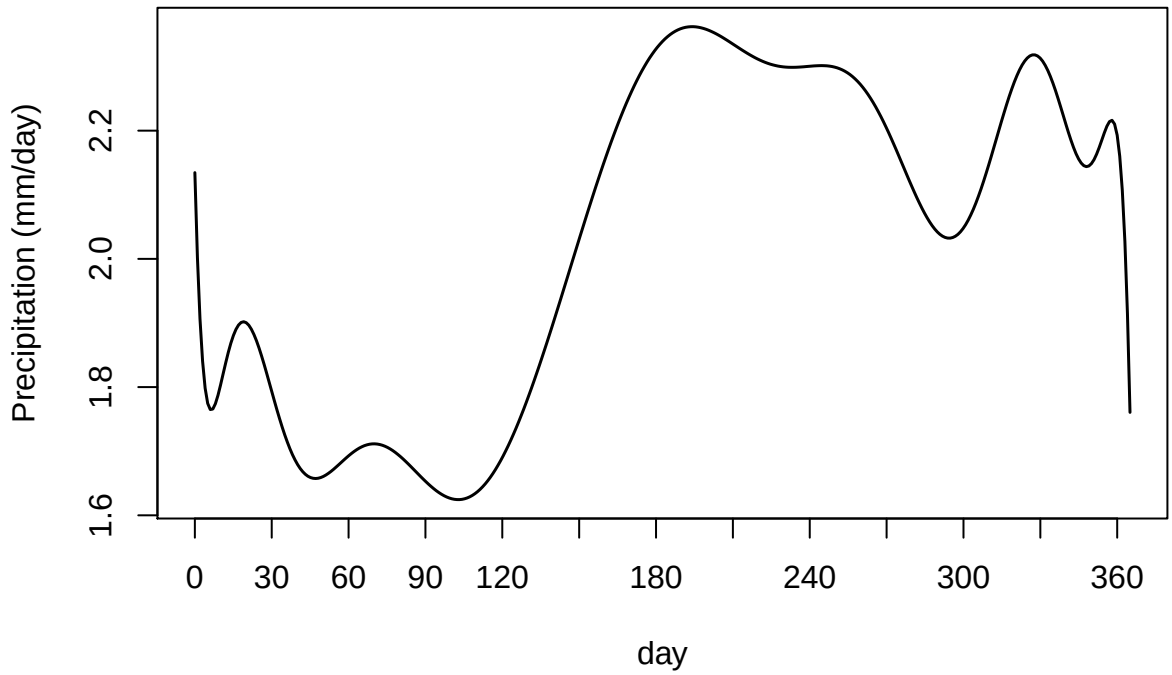
As discussed in section 2.1 we have repeated the SVD with exclusion of the outlier (Pr. Rupert). We briefly discuss these results here.

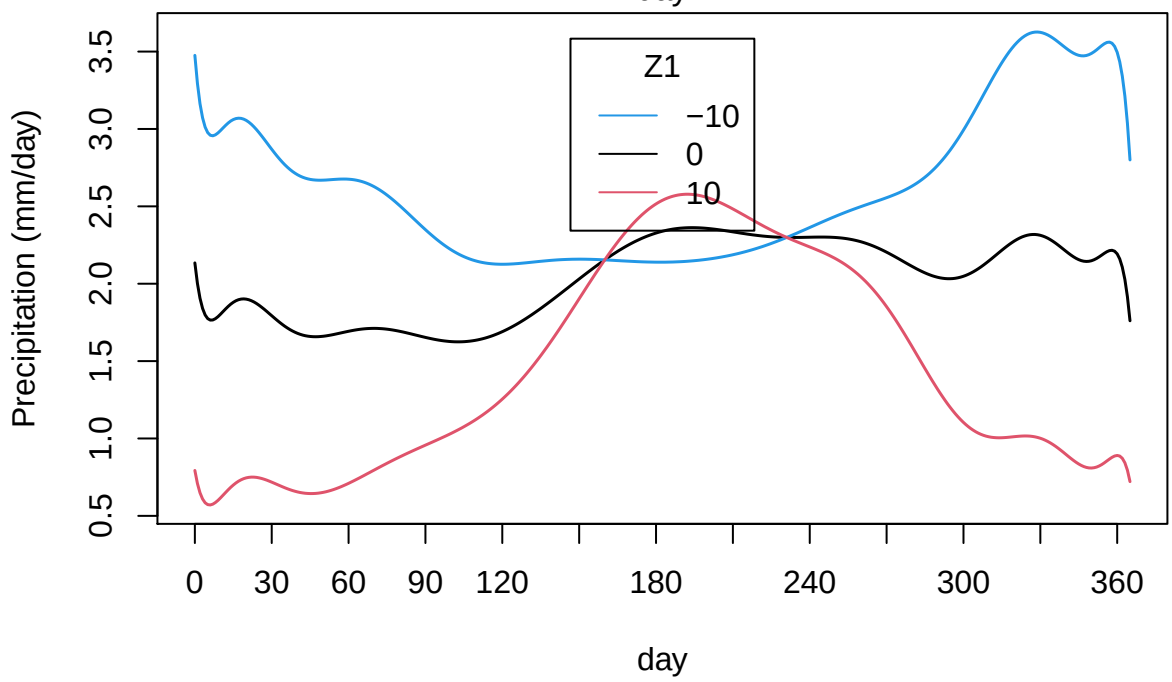
- There is no longer a 'slanted' pattern in the biplot
- The biplot still groups cities in Continental and Arctic region together
- The average precipitation is very similar as in the main analysis
- Z1 has a similar interpretation
- Z2 however has a different interpretation now. It does no longer only include precipitation in autumn, but also earlier in the year (days 30-120)

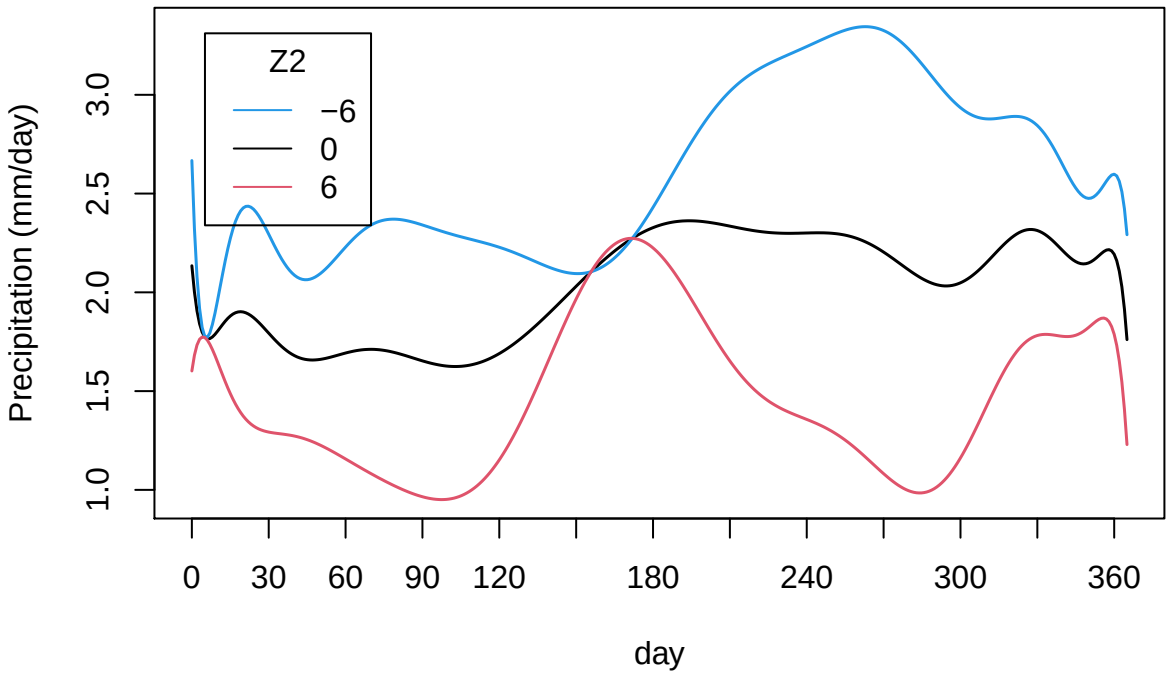



```
## [1] 0.9078298
```


z2
z1
Atlantic
Continental
Pacific
Arctic


Precipitation (mm/day)
1.6
1.8
2.0
2.2
0
30
60
90
120
180
240
300
360
day


Precipitation (mm/day)
0.5
1.0
1.5
2.0
2.5
3.0
3.5
Z1
−10
0
10
0
30
60
90
120
180
240
300
360
day


Z2
−6
0
6
Precipitation (mm/day)
1.0
1.5
2.0
2.5
3.0
0
30
60
90
120
180
240
300
360
day