SEOUL BIKE RENTALS

FDA

CONTENT

CONTENT

INTRODUCTION

Functional data analysis (FDA) is a statistical framework that enables us to convert discretely measured count of bike rented and weather conditions for the city of Seoul into functional curves in a manner that breaks away from the conventional multivariate statistic analysis.

A functional analysis would sever us to handle dynamic data by modeling the discrete measurements as continuous values over a period of time.

The focus of the project will be conducting functional data analysis of the bike rental count of the city Seoul in South Korea.

Data was collected on an hourly basis for an entire year with the first observation recorded from the first of December 2017 to the last day of November 2018.

We have three categorical variables and eight numerical variables including our response variable.

DATA DESCRIPTION- CONTINUOUS

Variable Name	Description	Data Type
date	year-month-day	date
count	Count of bikes rented collected at each hour	integer
temo	Temperature in Celsisus	numerical
humidity	Humidity in %	numerical
wind	Wind speed in m/s	numerical
visibility	Visibility in 10m	numerical
dew_temp	Dew temperature in Celsius	numerical
solar	Solar Radiation MJ/m2	numerical
rain	Numerical Rainfall in mm	numerical
snow	Snowfall in cm	numerical

DATA DESCRIPTION- CATEGORICAL

Variable Name	Description	Data Type
season	Winter; Spring; Summer; Autumn	categorical
holiday	Holiday or No Holiday	categorical
functional	NoFunc (No Functional Hours), Fun(Functional Hours)	categorical

PRELIMINARY

Converting our variables into their appropriate representations, where categorical variables are represented as factors and date used the appropriate format.

Checked for missing values and duplicates and none were found.

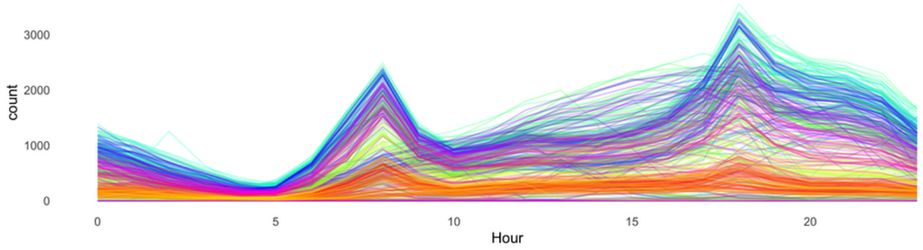
Excluded the days where the bike rental service was not functional, leaving us with 353 days to analyse.

Snowfall and rainfall had predominantly zero values thus we added a small positive number and took the log tranformation.

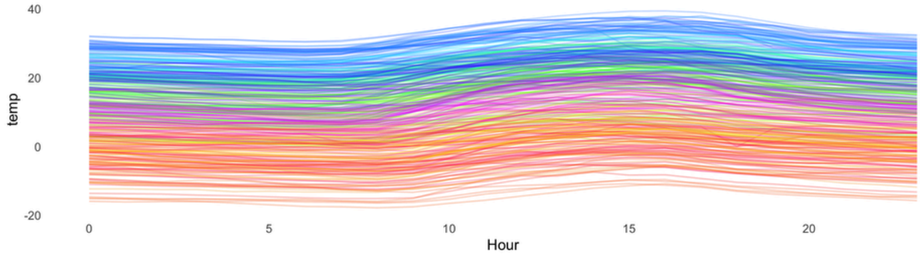
We visualised the discrete representation of our continuous variables to decide on the type of transformation we will be undertaking for our functional curves.

We inspected the bike rental count difference based on the differnet seasons and weekends.

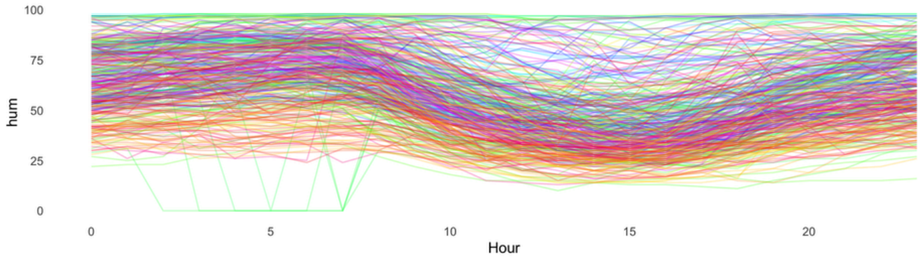
Hourly count for 365 Days



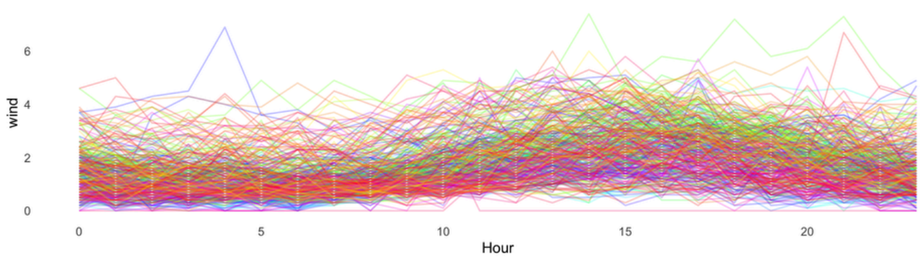
Hourly temp for 365 Days



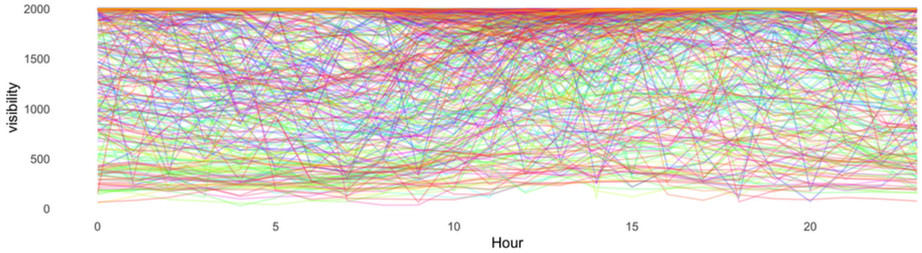
Hourly hum for 365 Days



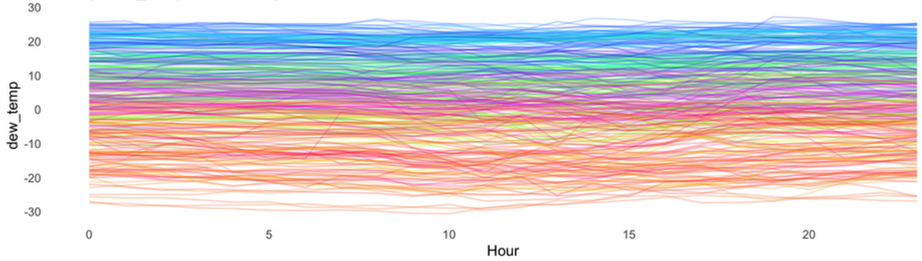
Hourly wind for 365 Days



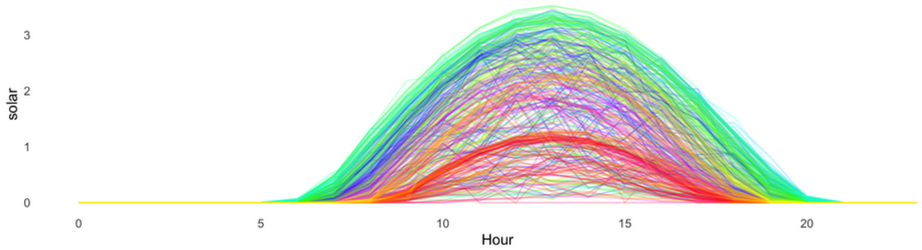
Hourly visibility for 365 Days



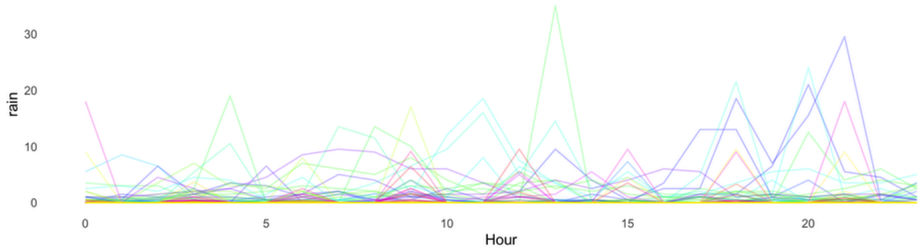
Hourly dew_temp for 365 Days



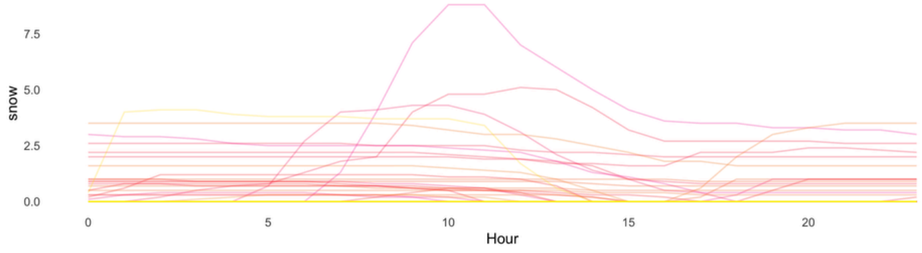
Hourly solar for 365 Days

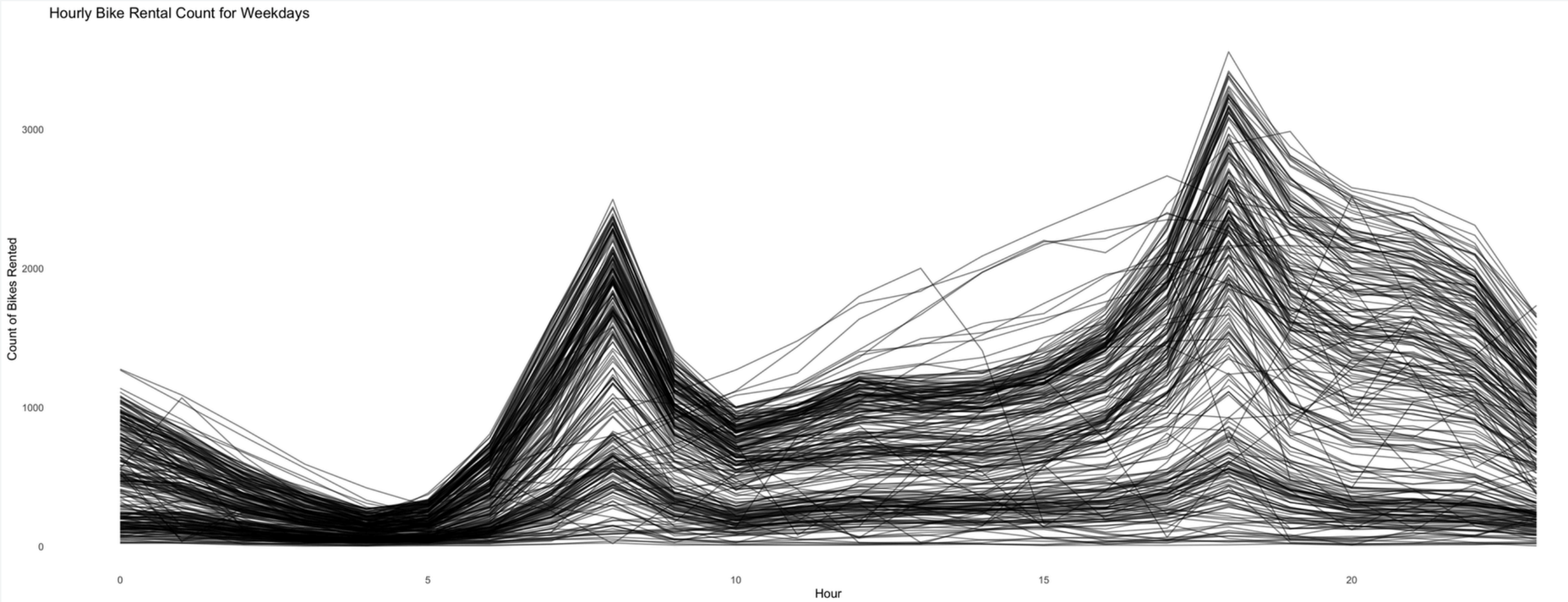


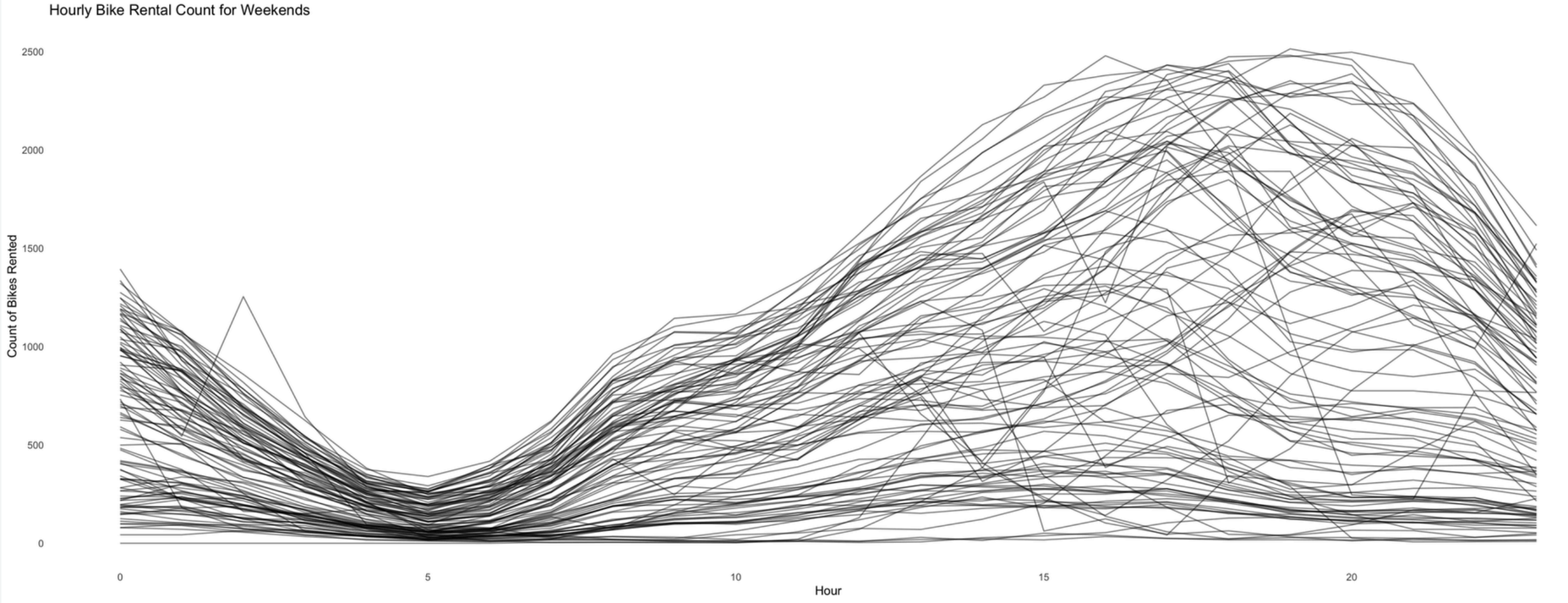
Hourly rain for 365 Days

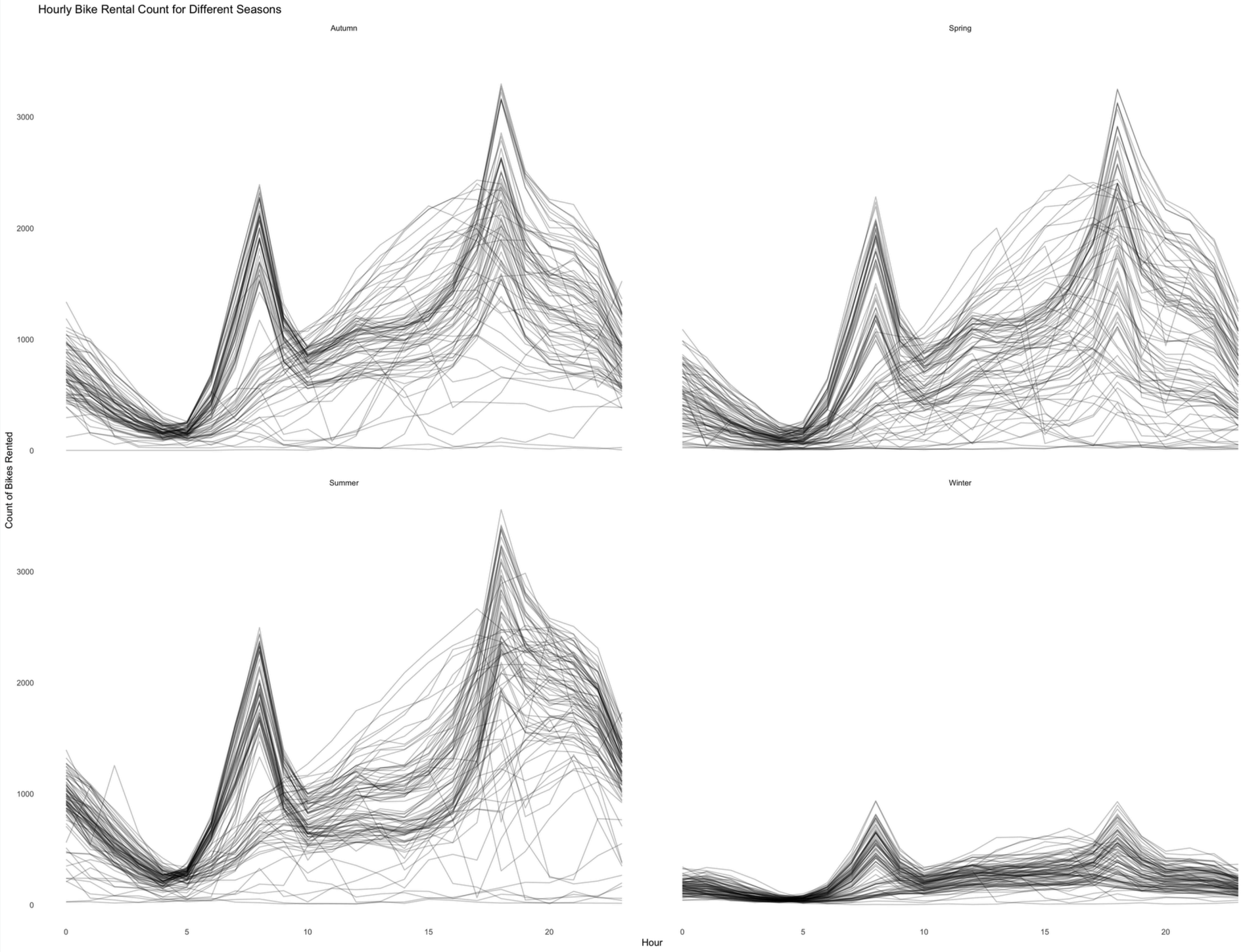


Hourly snow for 365 Days




Hourly Bike Rental Count for Weekdays
Count of Bikes Rented
3000
2000
1000
0
0
5
10
15
20
Hour


Hourly Bike Rental Count for Weekends
Count of Bikes Rented
2500
2000
1500
1000
500
0
0
5
10
15
20
Hour


Hourly Bike Rental Count for Different Seasons
Autumn
Spring
Summer
Winter
Count of Bikes Rented
0
1000
2000
3000
0
5
10
15
20
Hour

QUESTIONS?

Predicting the bike rental counts in the city of Seoul based on the weather condition.

Investigate the impact of the different pre-processing techniques, such as curves with smoothing roughness penalty and curves aligned with time warping and their performance in predicting bike count accurately.

The impact of seasonal variation in the bike rental patterns?

RESULTS OF PRE-PROCESSING

SMOOTHING

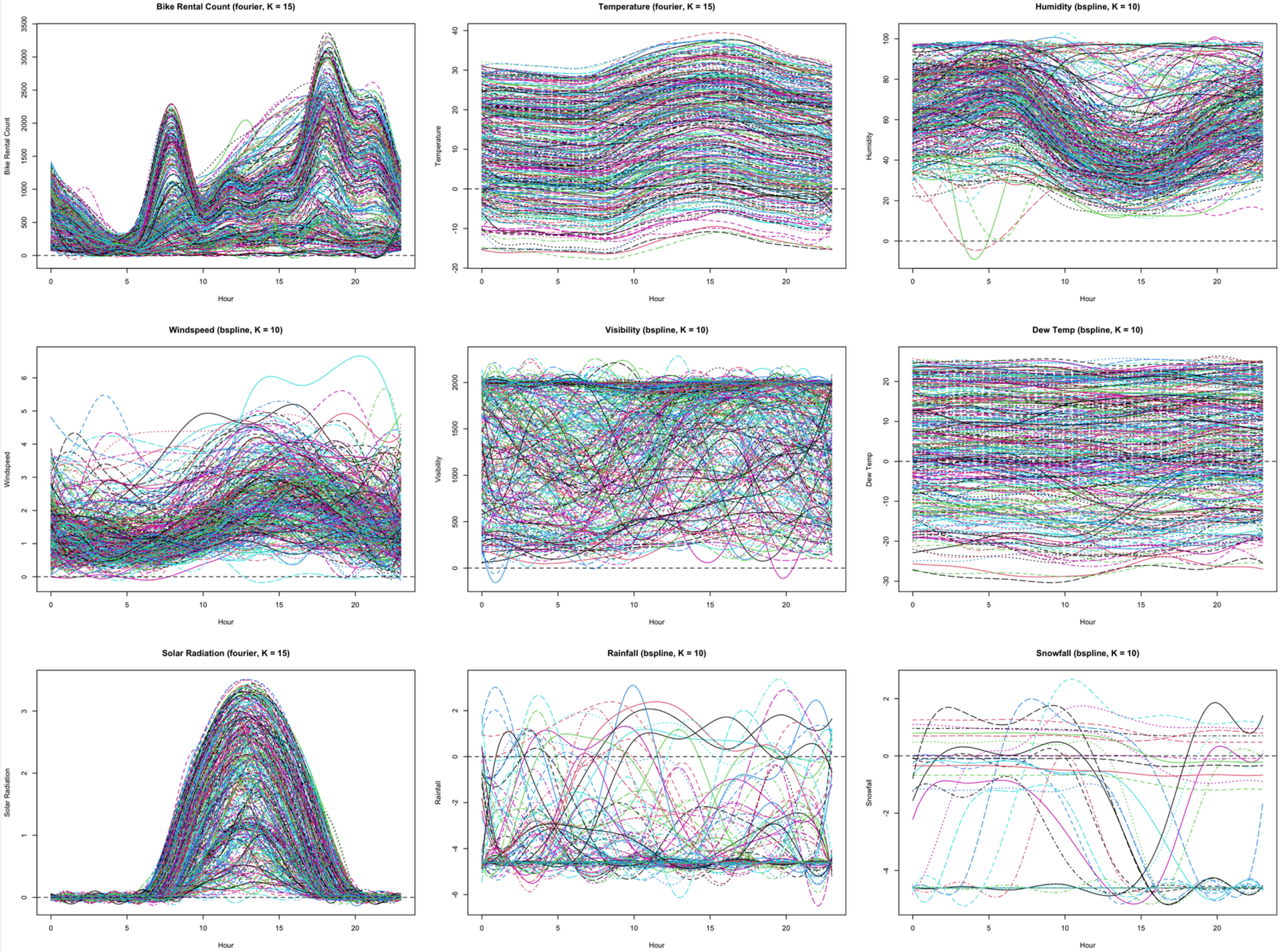
Each day is treated as a functional representation over the 24 hours. Overall we will have 353 functional curves that we will be relying for our analysis.

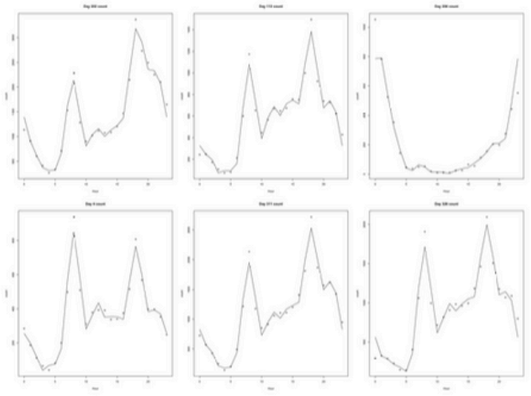
We converted some of the variables using the Fourier transformation and others using the B-spline. Initially we resorted to moderate level of basis.

Variables that will be relying on the Fourier transformation, namely, count, temperature and solar will take on a basis of K = 15.

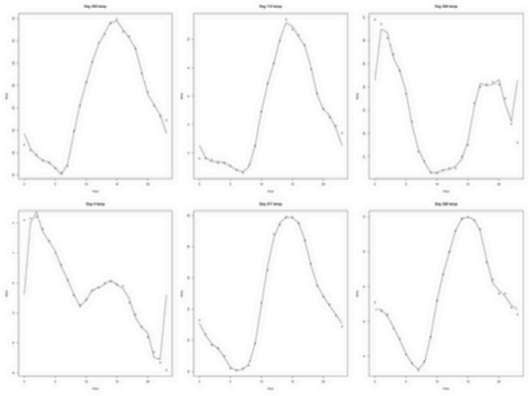
On the other hand variables that will be transformed using the B-spline will take on a basis K = 10.

In the next plot we will see the functional curves and also the sampled dyas for our smoothed curve.

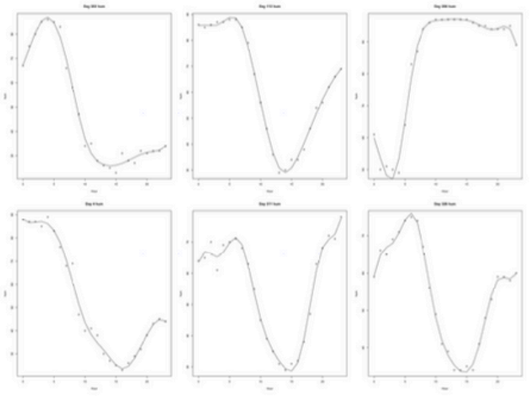

Bike Rental Count (fourier, K = 15)
Bike Rental Count
Hour
Temperature (fourier, K = 15)
Temperature
Hour
Humidity (bspline, K = 10)
Humidity
Hour
Windspeed (bspline, K = 10)
Windspeed
Hour
Visibility (bspline, K = 10)
Visibility
Hour
Dew Temp (bspline, K = 10)
Dew Temp
Hour
Solar Radiation (fourier, K = 15)
Solar Radiation
Hour
Rainfall (bspline, K = 10)
Rainfall
Hour
Snowfall (bspline, K = 10)
Snowfall
Hour



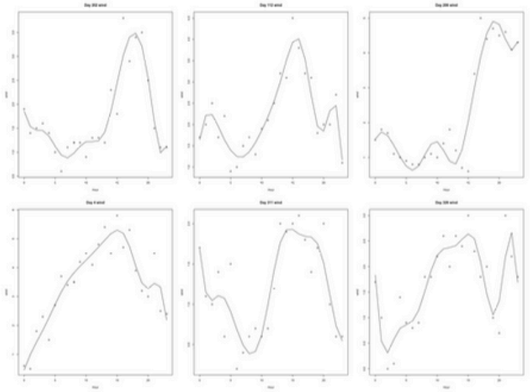
(a) Biked Rented Count



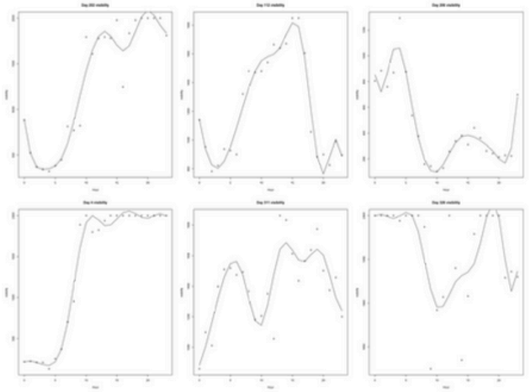
(b) Temperature



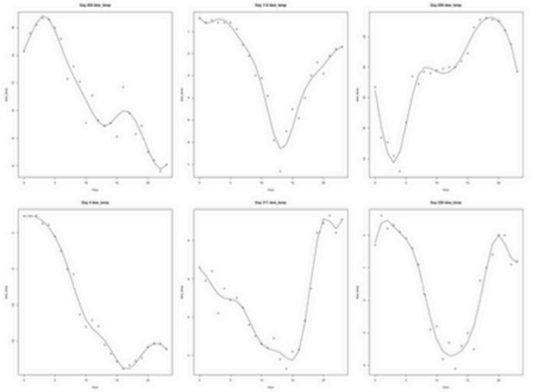
(c) Humidity



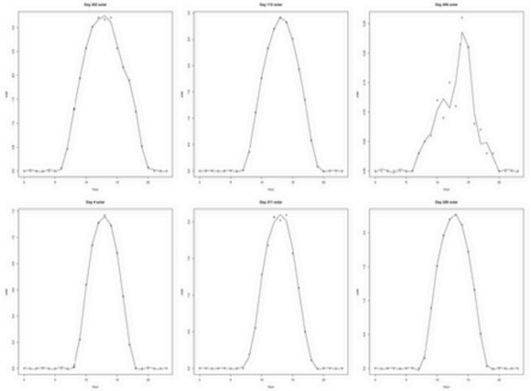
(d) Wind



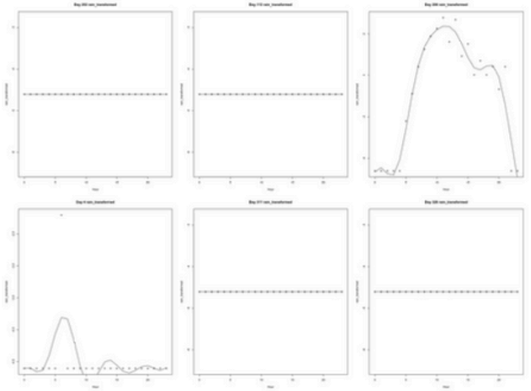
(e) Visibility



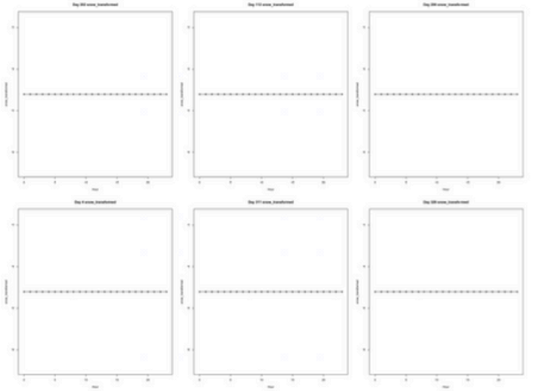
(f) Dew Temperature



(g) Solar



(h) Rain

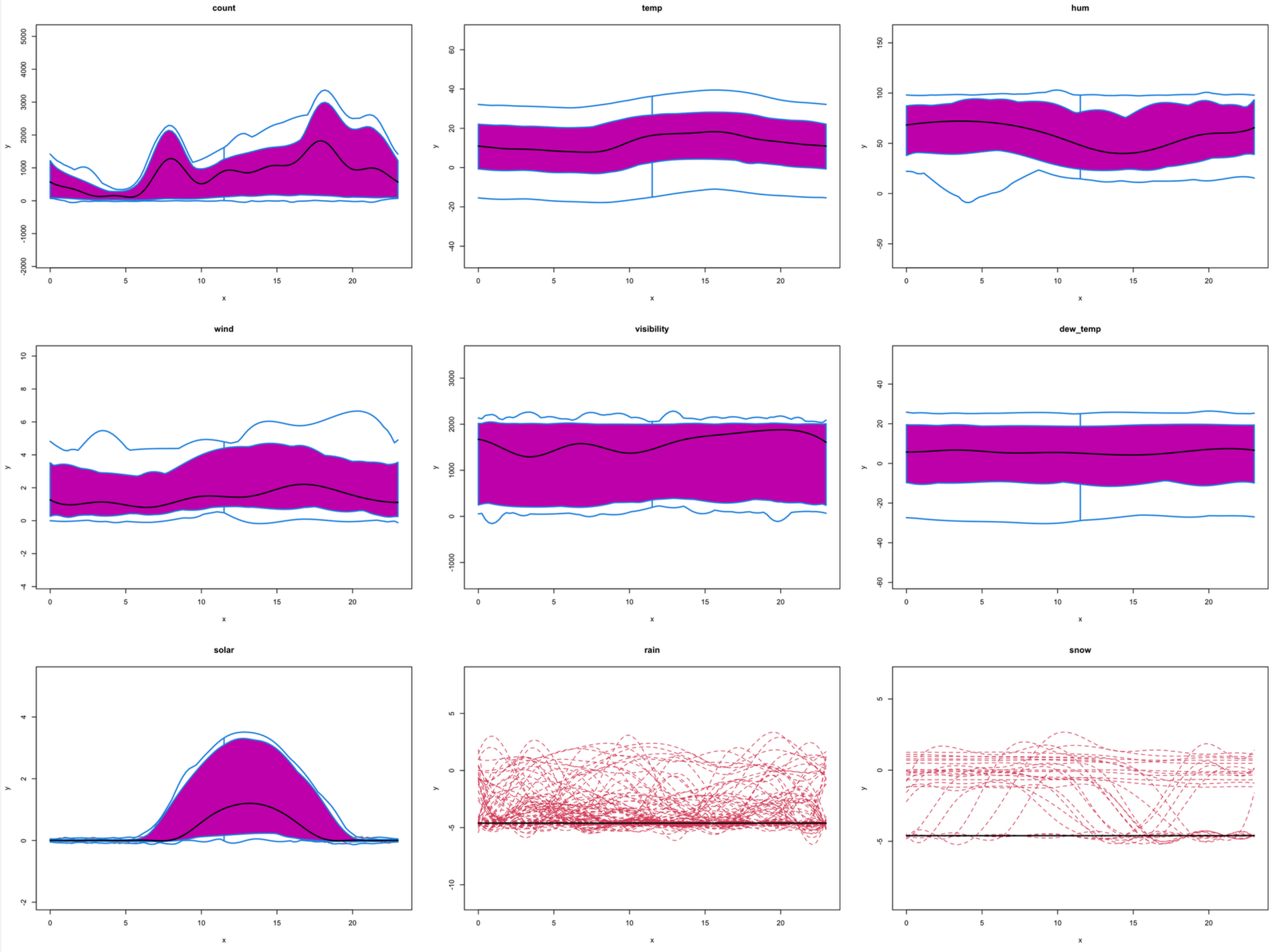


(i) Snow

Then we investigated if we have any outliers in our functional curves.

Not to our surprise we observe that the variables snowfall and rainfall do have outliers.

This plausible because we know that rainfall and snowfall are sporadic events in the city of Seoul. Also when investigating the summary statistics of the variables we had found that the variables were predominantly zero for a majority of the days.

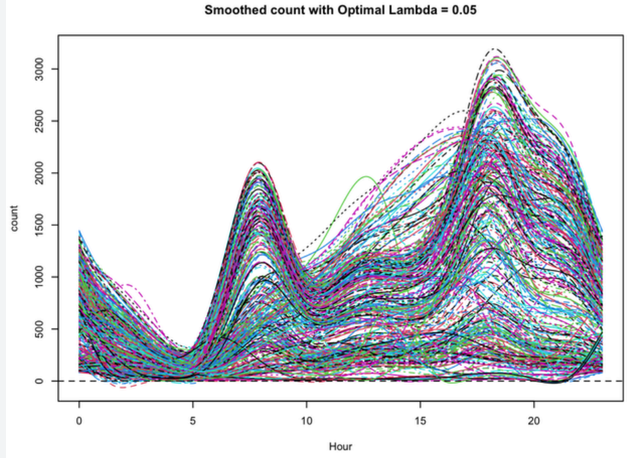

count
temp
hum
wind
visibility
dew_temp
solar
rain
snow
x
y

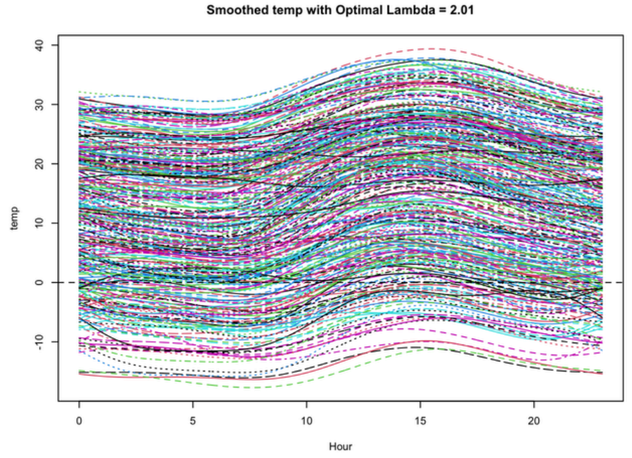
SMOOTHING WITH ROUGHNESS PENALTY

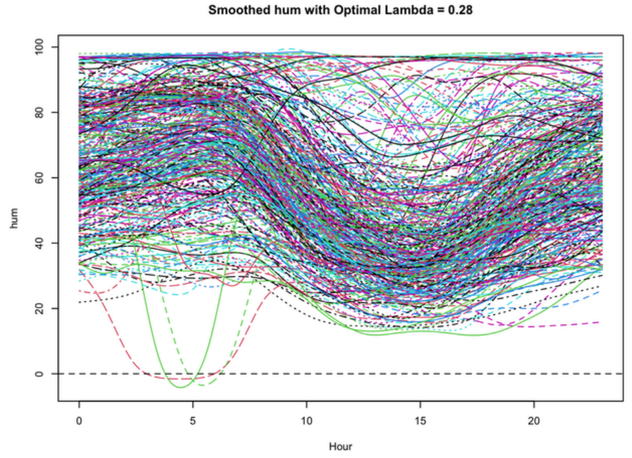
In this section we have applied roughness penalty in order to capture the underlying pattern of the observed.

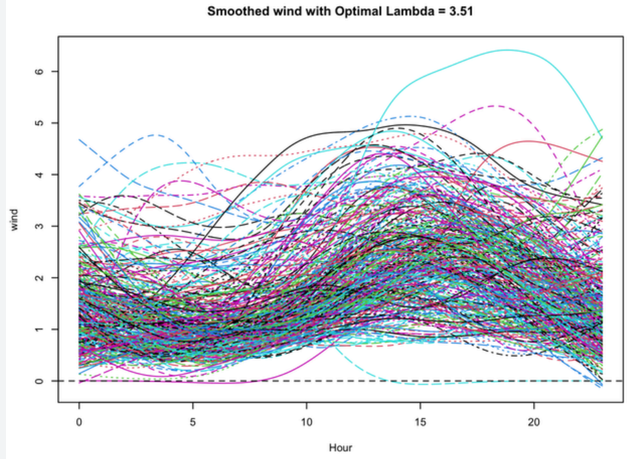
However instead of relying on the moderate basis we have selected an oversaturated basis of K = 23.

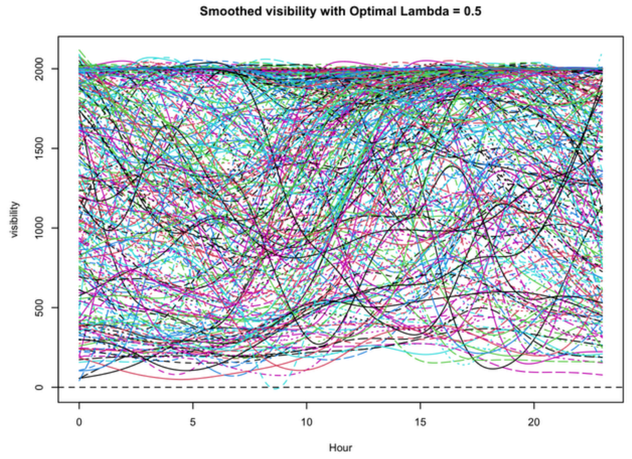
This has enabled us select an optimal lambda by minimising over the generalzied cross validation (GCV).

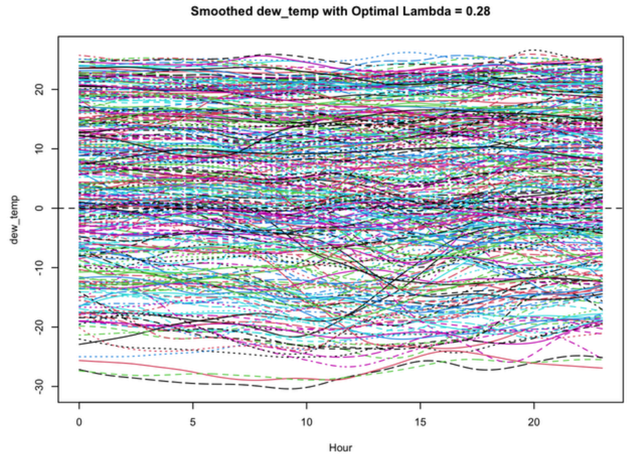

Smoothed count with Optimal Lambda = 0.05
count
Hour

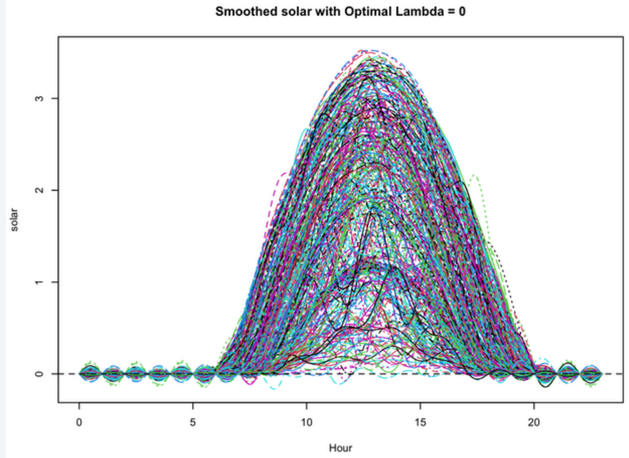

Smoothed temp with Optimal Lambda = 2.01
temp
Hour

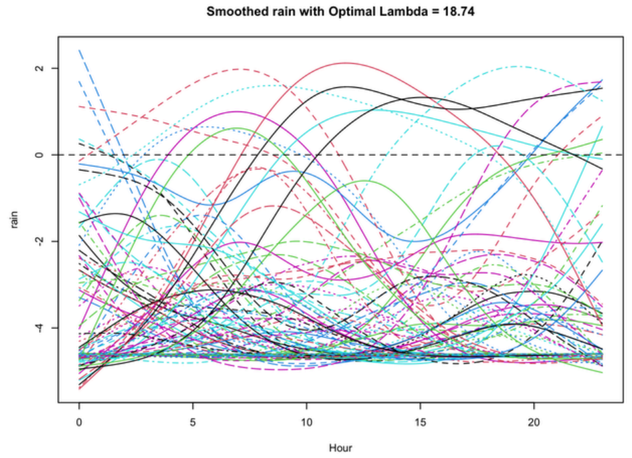

Smoothed hum with Optimal Lambda = 0.28
hum
Hour

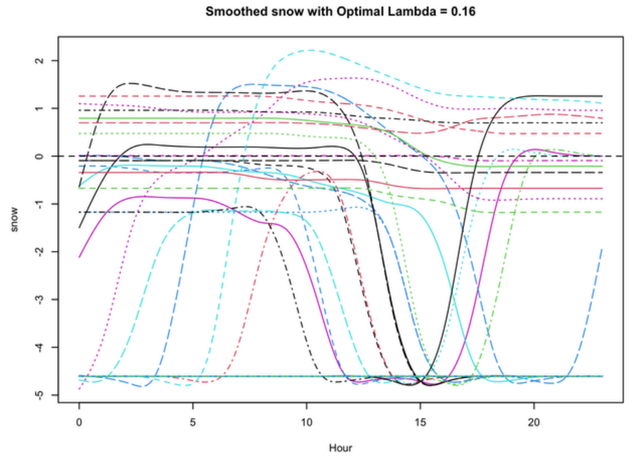

Smoothed wind with Optimal Lambda = 3.51
wind
Hour

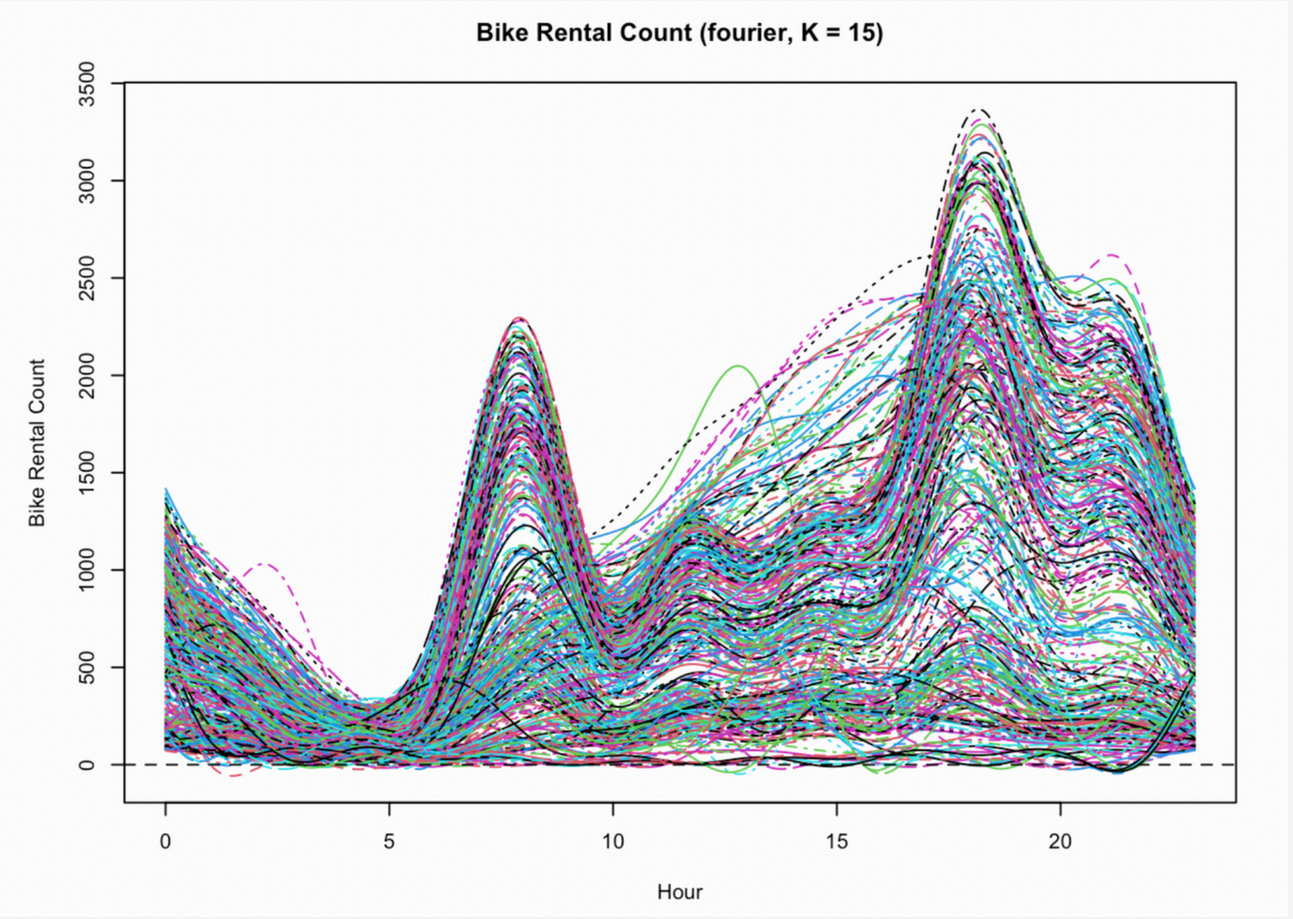

Smoothed visibility with Optimal Lambda = 0.5
visibility
Hour

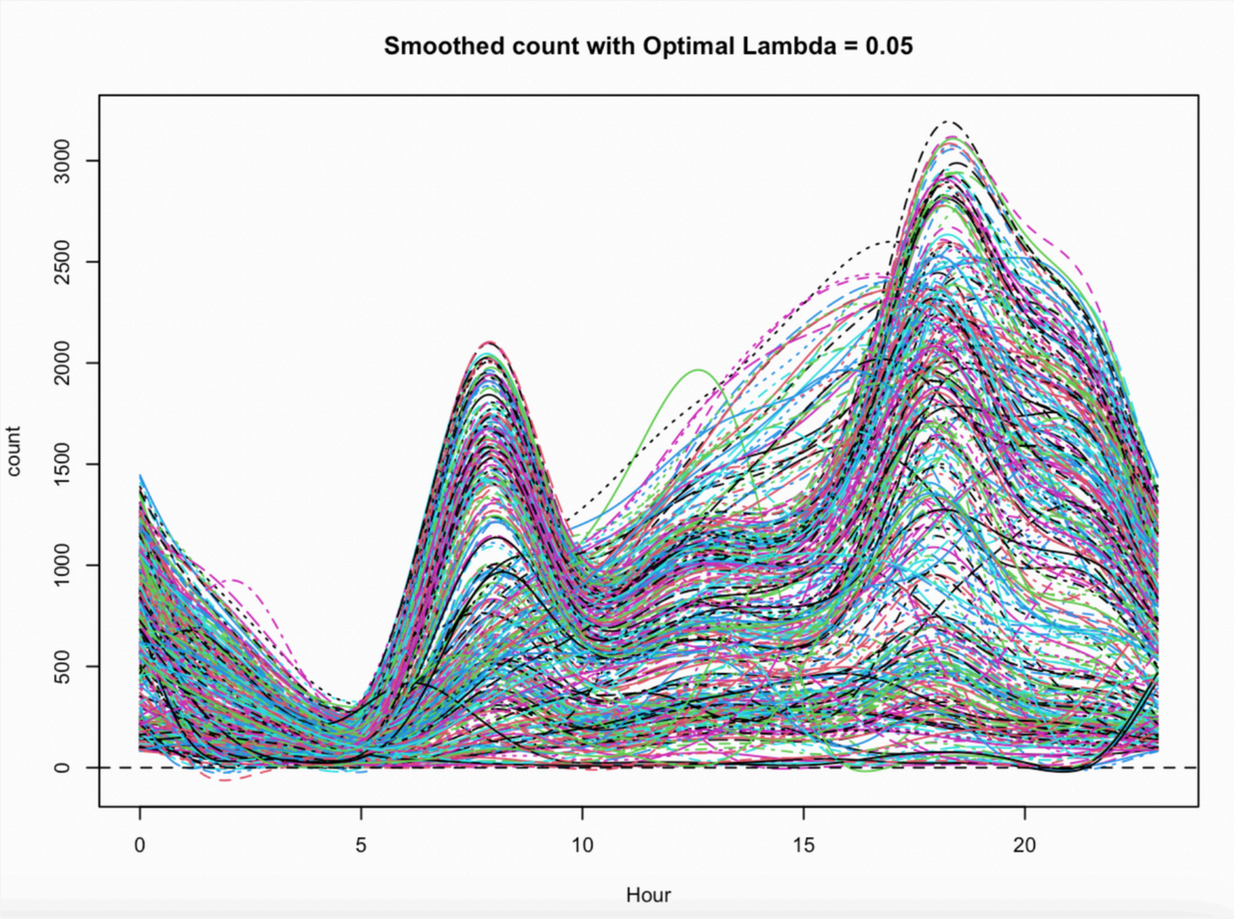

Smoothed dew_temp with Optimal Lambda = 0.28
dew_temp
Hour

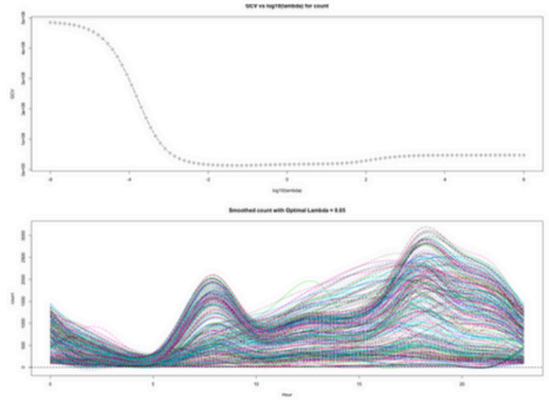

Smoothed solar with Optimal Lambda = 0
solar
Hour


Smoothed rain with Optimal Lambda = 18.74
rain
Hour

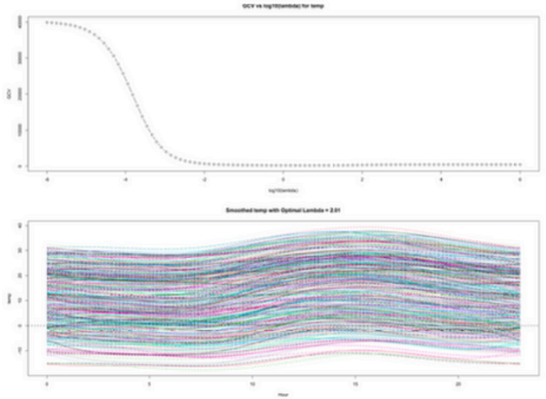

Smoothed snow with Optimal Lambda = 0.16
snow
Hour


Bike Rental Count (fourier, K = 15)
Bike Rental Count
0
500
1000
1500
2000
2500
3000
3500
0
5
10
15
20
Hour

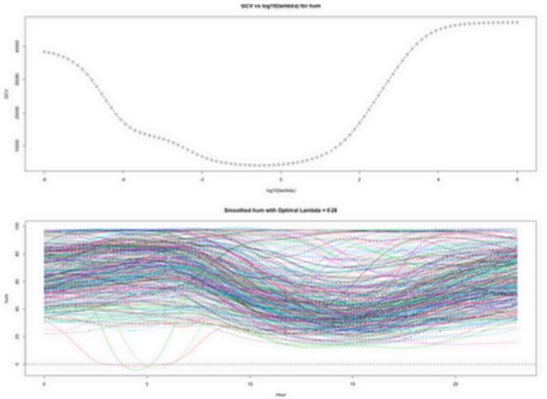

Smoothed count with Optimal Lambda = 0.05
count
0
500
1000
1500
2000
2500
3000
0
5
10
15
20
Hour



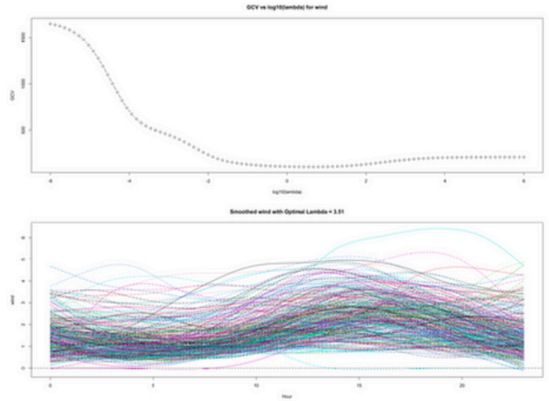
(a) Count



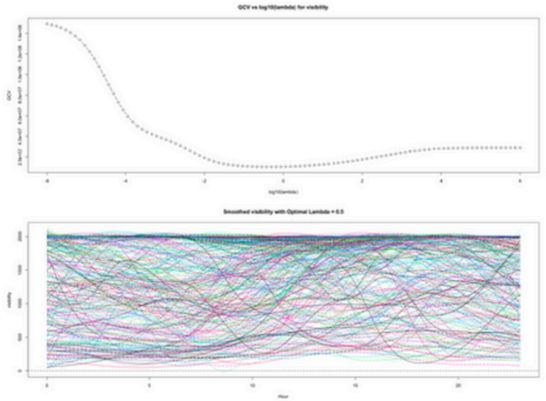
(b) Temperature



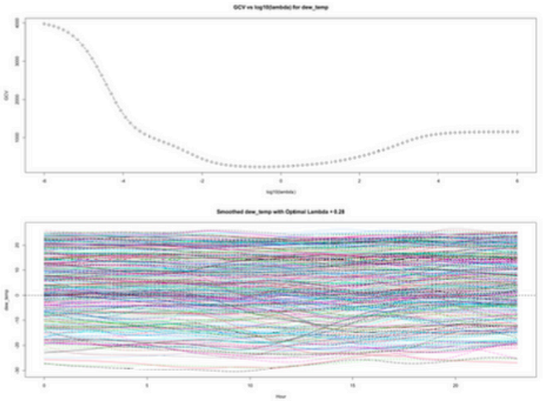
(c) Humidity



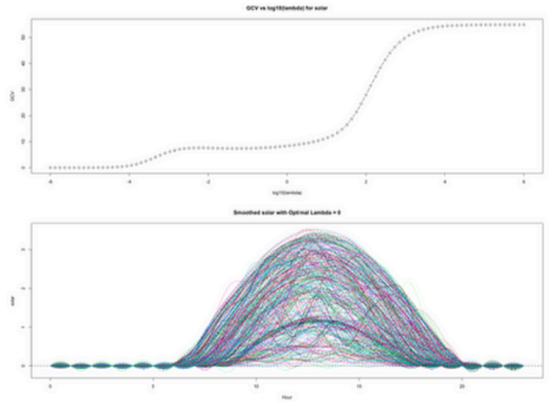
(d) Wind



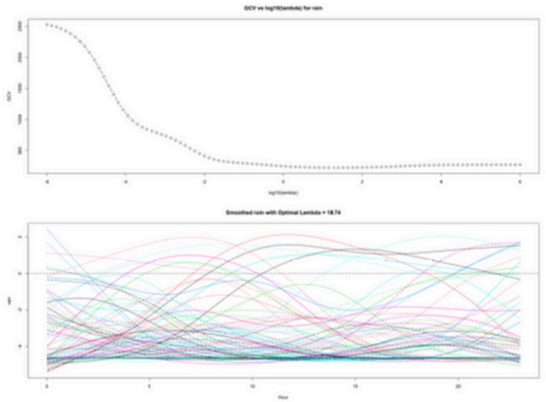
(e) Visibility



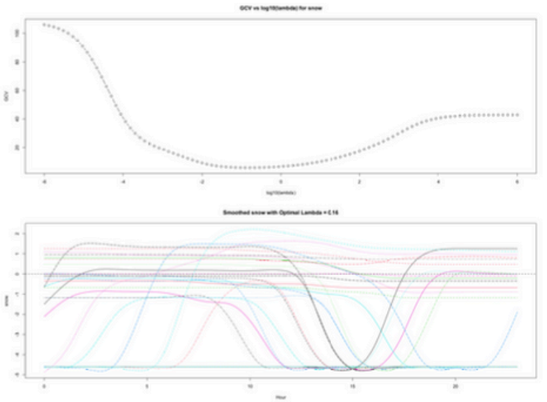
(f) Dew Temperature



(g) Solar



(h) Rain



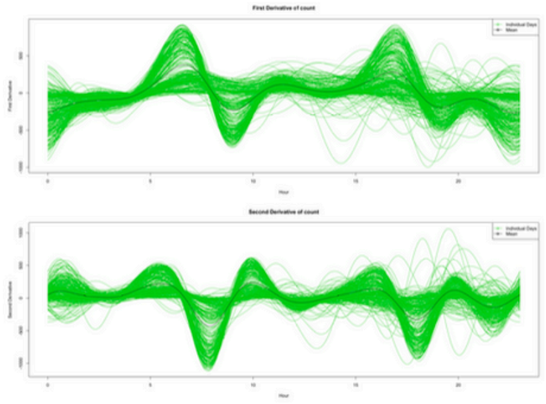
(i) Snow

ALIGNEMENT: PHASE VARIATION USING WARPING METHOD

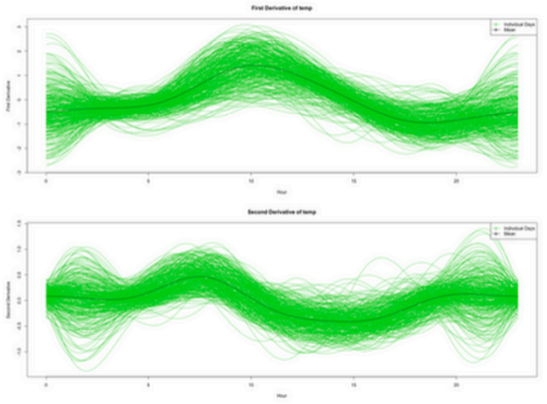
Let us briefly assess the first and second derivatives of our functional curves. This will enable us to inspect the rate of change that occurs for each variable.

Inspecting the second derivatives will give us a way of investigating the rate of rate of change.

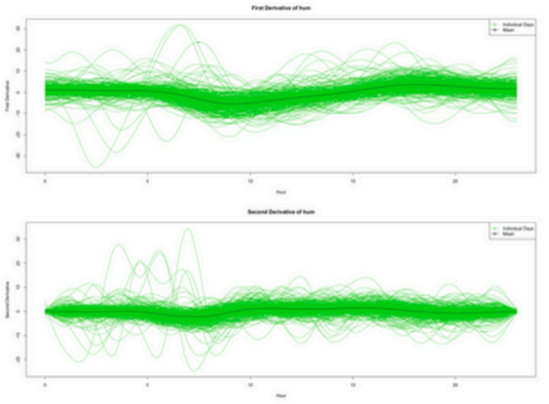
Even though it is more of interest to study the derivatives of the functional data, where we use them as objects of analysis in our instance we have mainly used to visualise the patterns for our alignment procedures.



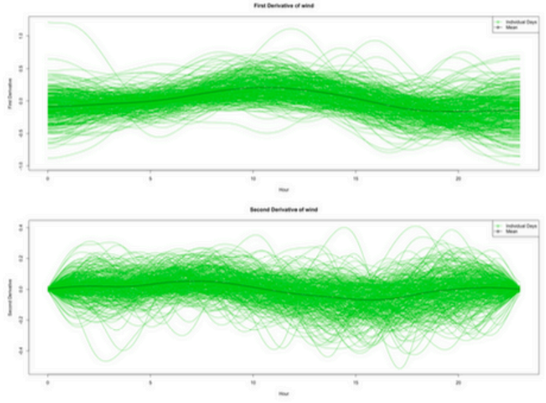
(a) Count Derivatives



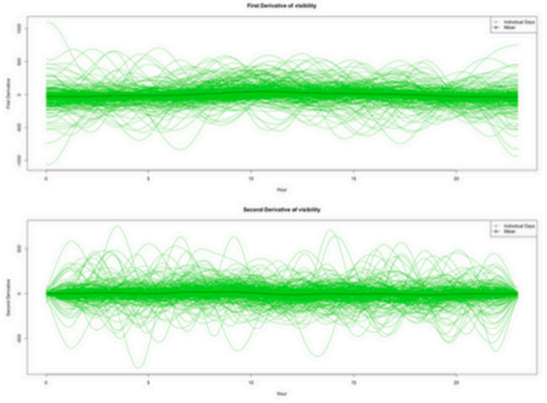
(b) Temperature Derivatives



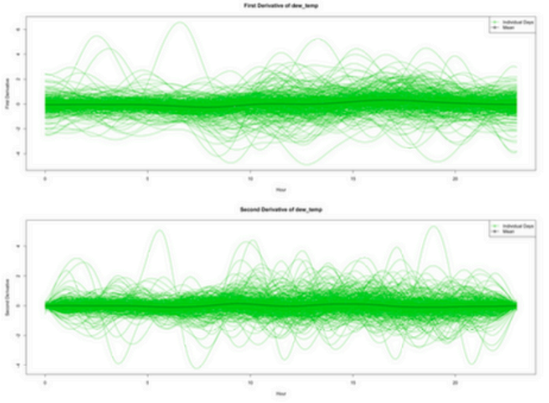
(c) Humidity Derivatives



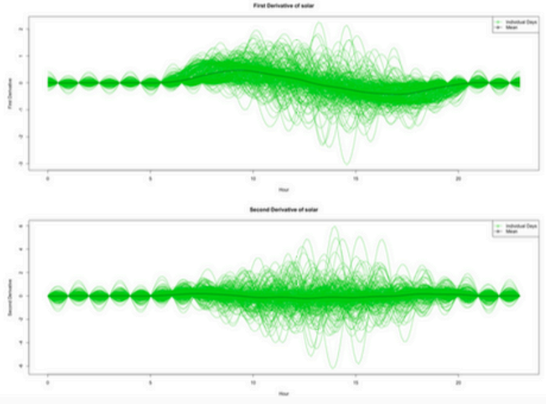
(d) Wind Derivatives



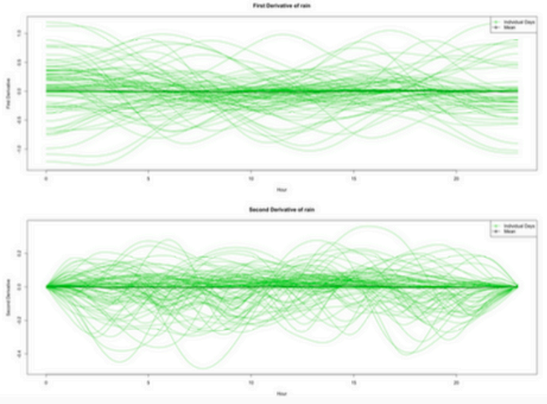
(e) Visibility Derivatives



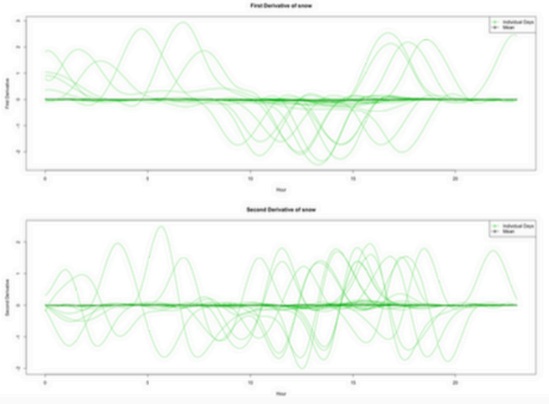
(f) Dew Temperature Derivatives



(g) Solar Derivatives



(h) Rain Derivatives



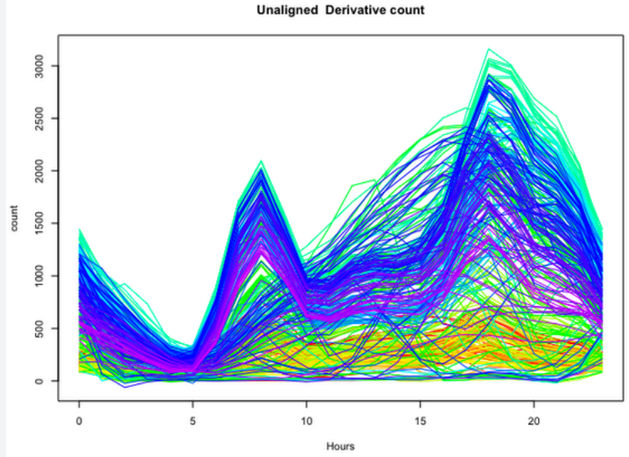
(i) Snow Derivatives

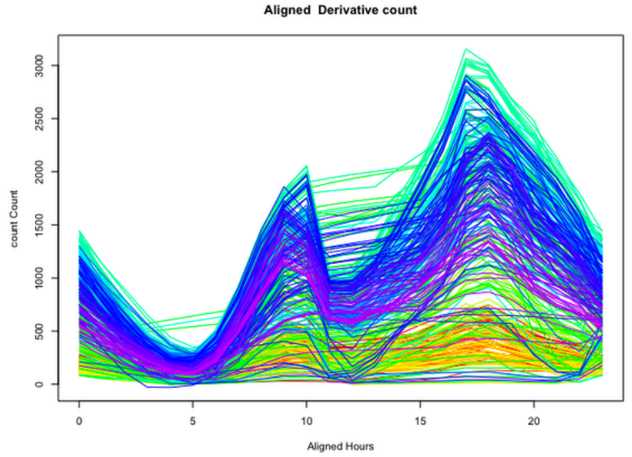
There are two main techniques that are used in the data registration process for the alignment procedure, which are landmark registration, where we would inspect the derivatives for the minima, maxima and zero crossing or other corresponding values of the derivatives that may be of interest to our analysis.

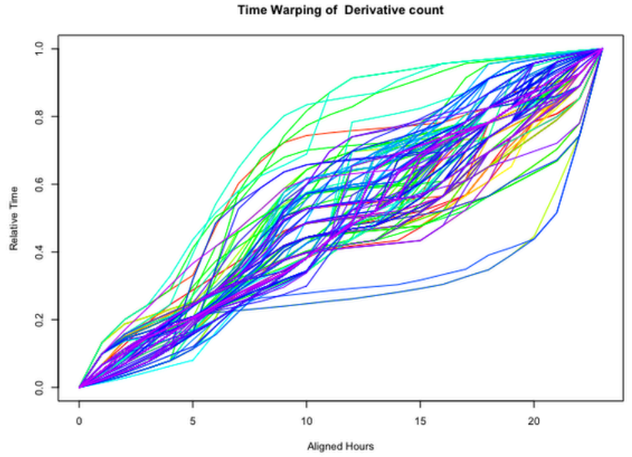
Conversely, we have relied on the warping approach. This means it will align curves that have similar shapes but they occur at different times.

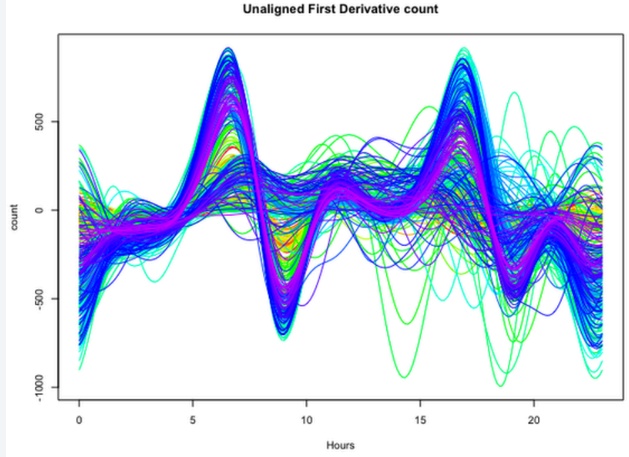
Investigating the time warping plots for the different variables we can clearly observe that there is noticeable phase distortion for all of them.

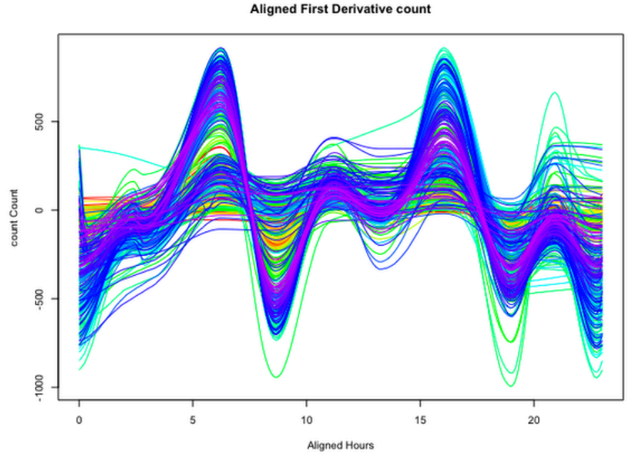
We have exploited this avenue to inspect if our bike demand prediction could be improved using the technique.

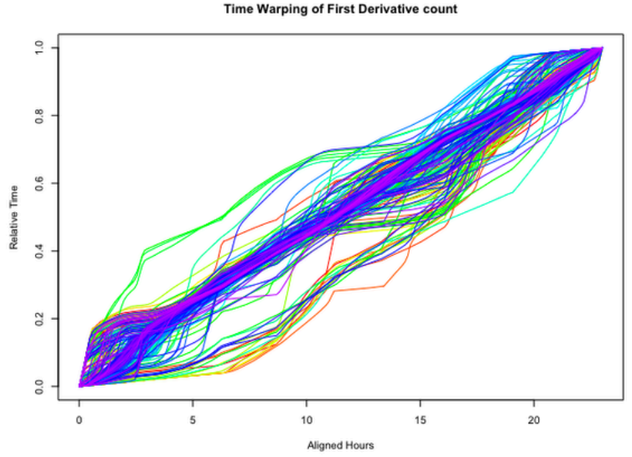

Unaligned Derivative count
count
Hours

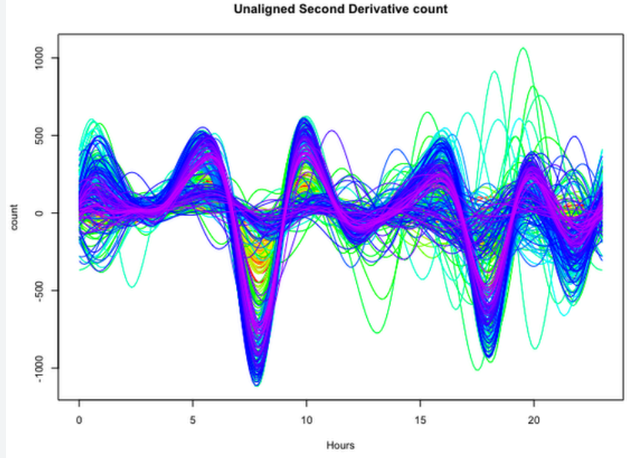

Aligned Derivative count
count Count
Aligned Hours

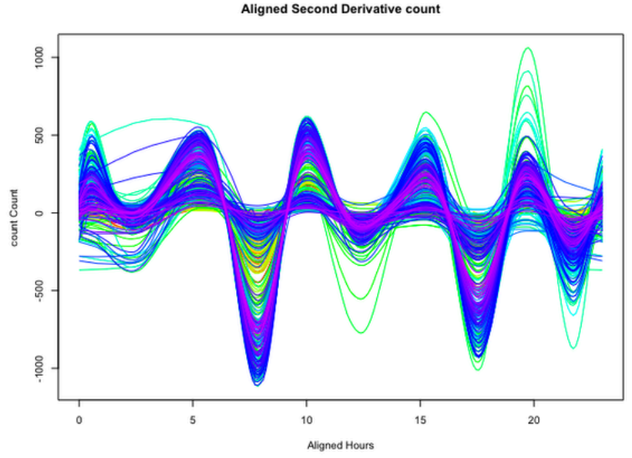

Time Warping of Derivative count
Relative Time
Aligned Hours

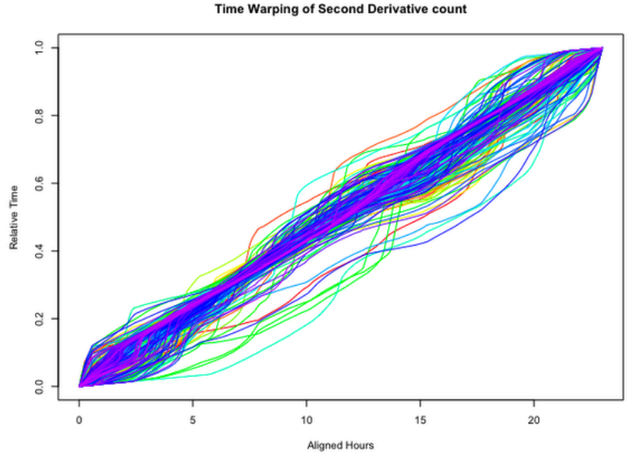

Unaligned First Derivative count
count
Hours

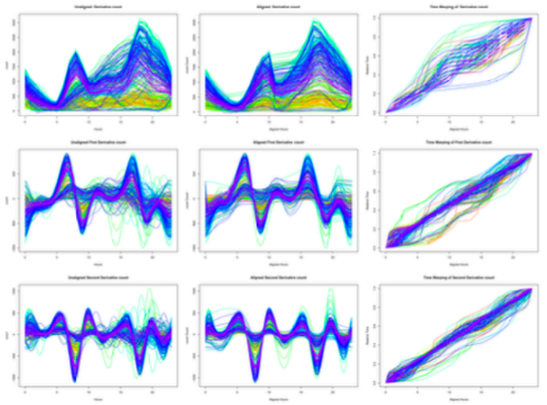

Aligned First Derivative count
count Count
Aligned Hours


Time Warping of First Derivative count
Relative Time
Aligned Hours

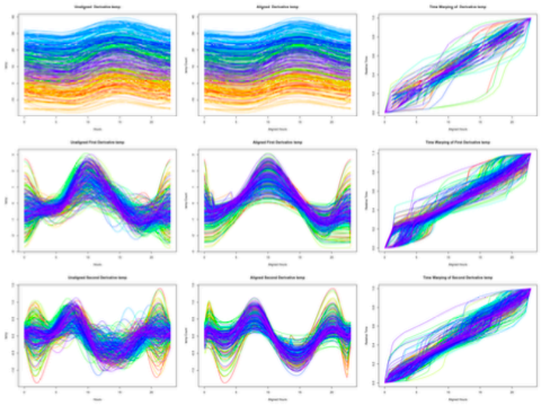

Unaligned Second Derivative count
count
Hours


Aligned Second Derivative count
count Count
Aligned Hours

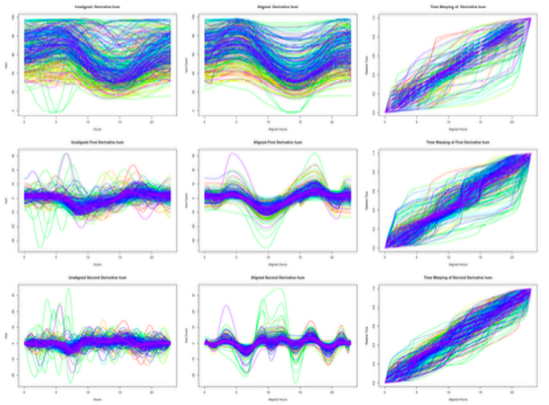

Time Warping of Second Derivative count
Relative Time
Aligned Hours



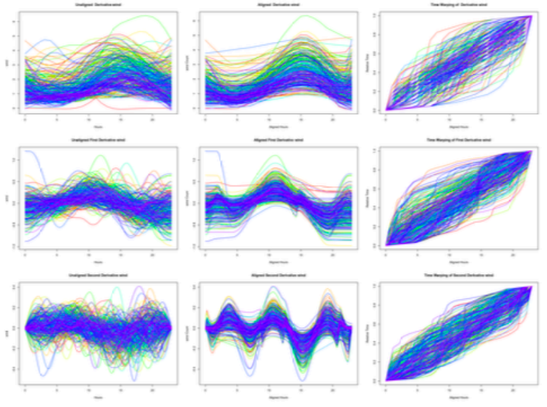
(a) Count Warping



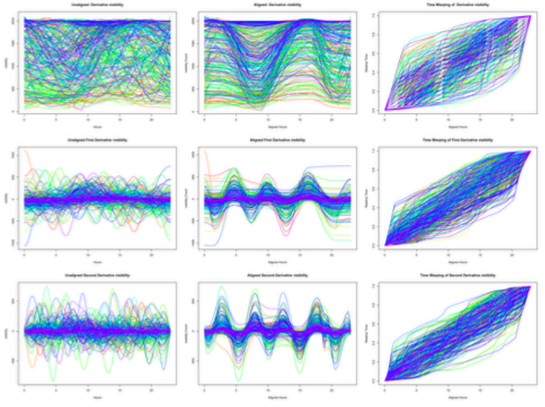
(b) Temperature Warping



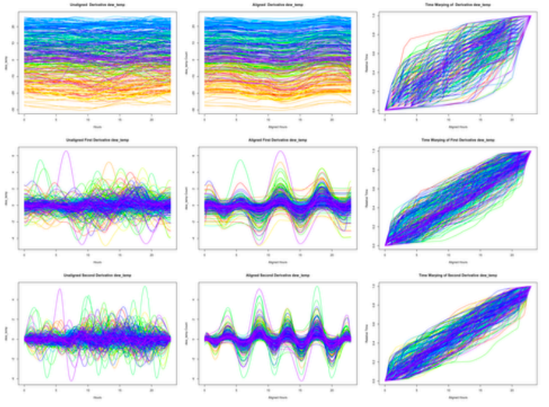
(c) Humidity Warping



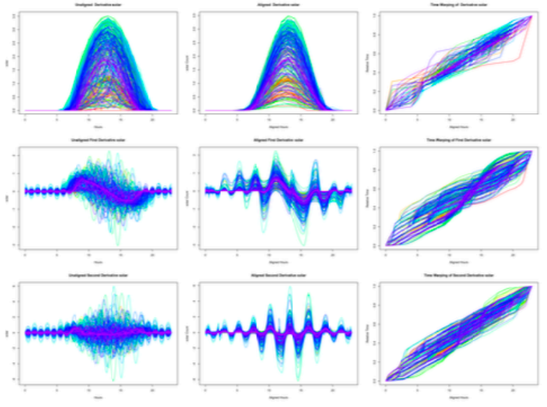
(d) Wind Warping



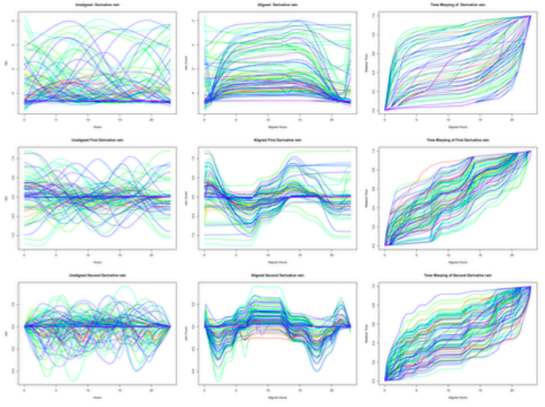
(e) Visibility Warping



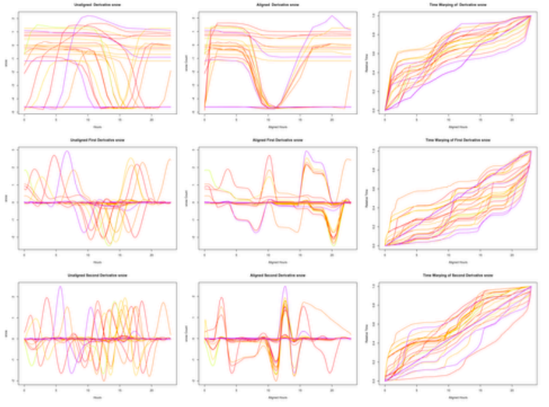
(f) Dew Temperature Warping



(g) Solar Warping



(h) Rain Warping

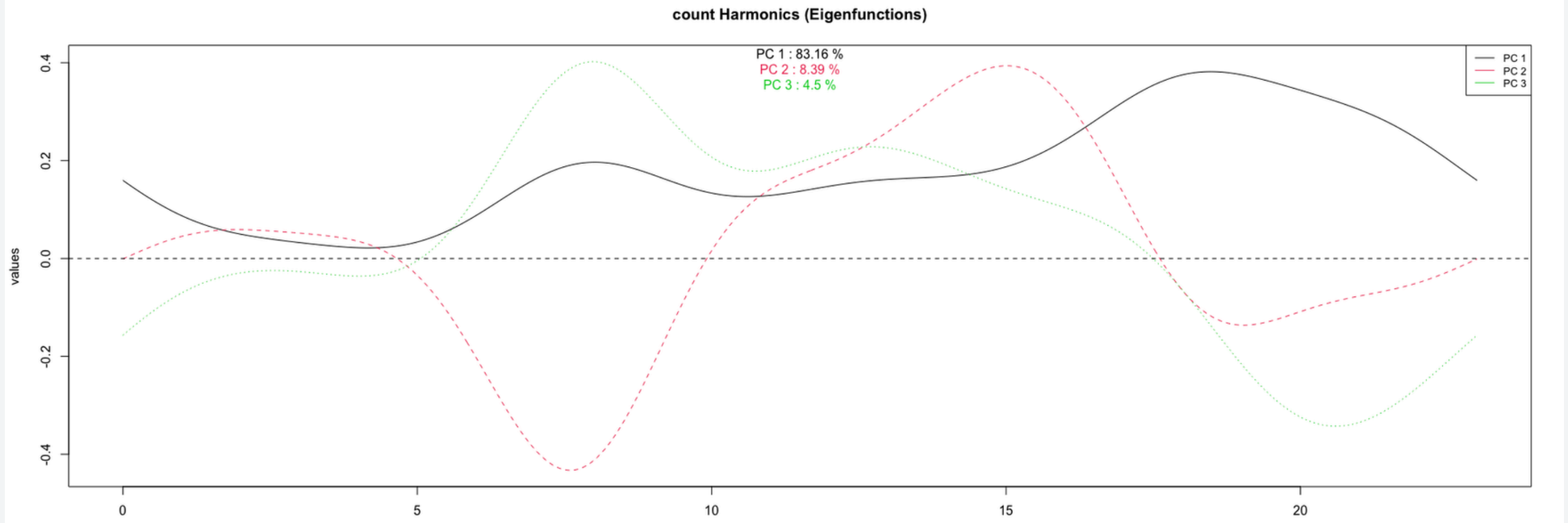


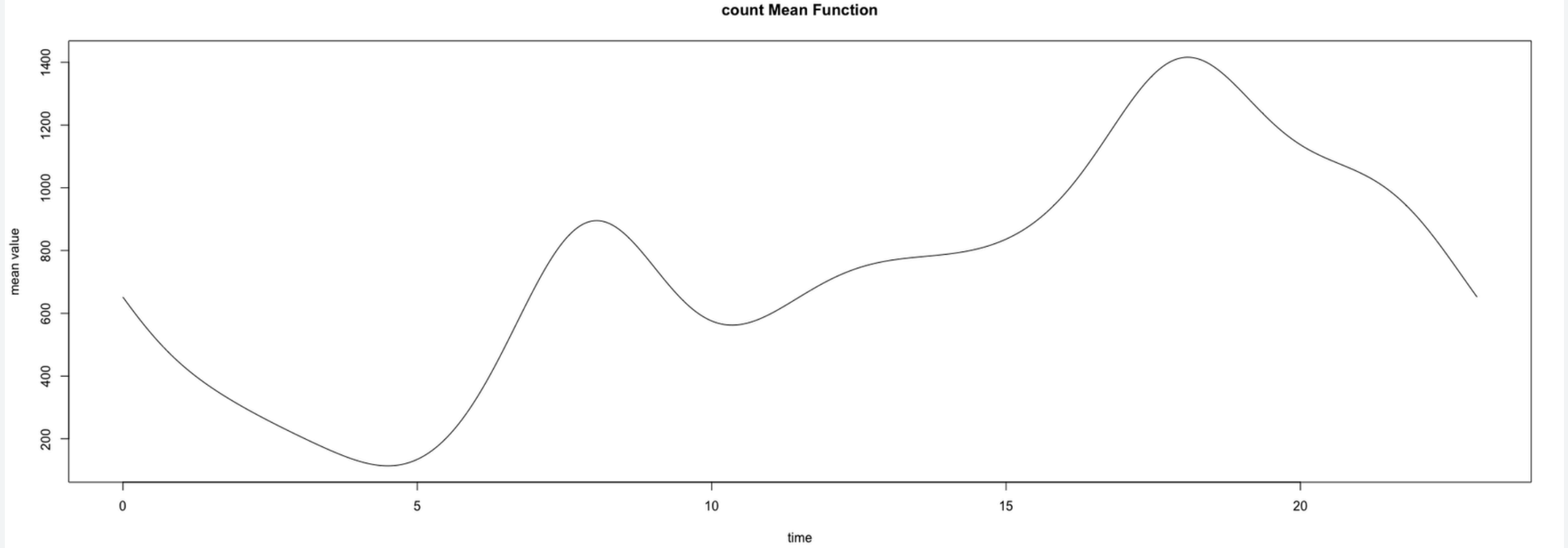
(i) Snow Warping

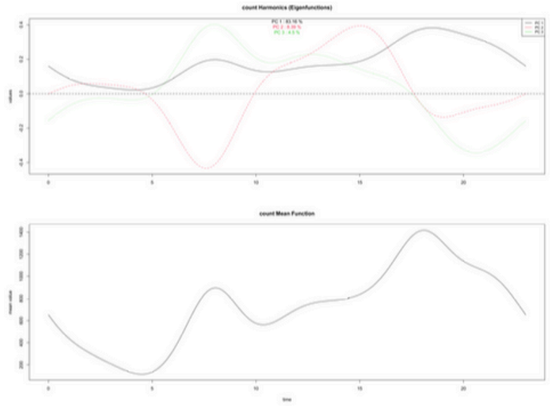
FPCA: DIMENSIONALITY REDUCTION

One of the most important concepts is Functional Principal Component Analysis (FPCA), which allows us to conduct dimensionality reduction to a manageable finite vectors of FPC scores.

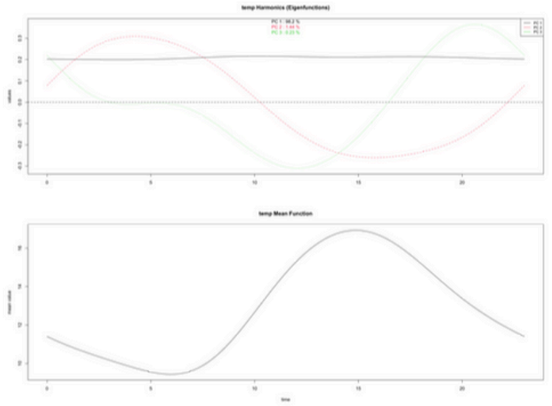
In this manner we can represent the principal component scores, which enable use to get a truncated representation our functional curves. With respect to the FPCA we will be relying on the plots of the harmonics and the mean to understand the variation and average patterns of our observations.


count Harmonics (Eigenfunctions)
PC 1 : 83.16 %
PC 2 : 8.39 %
PC 3 : 4.5 %
PC 1
PC 2
PC 3
values
0.4
0.2
0.0
-0.2
-0.4
0
5
10
15
20

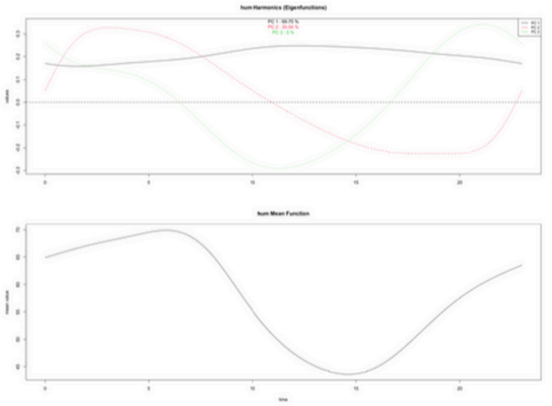

count Mean Function
mean value
1400
1200
1000
800
600
400
200
0
5
10
15
20
time



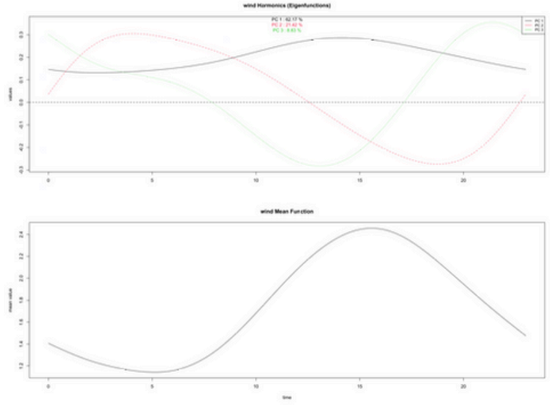
(a) Count FPCA



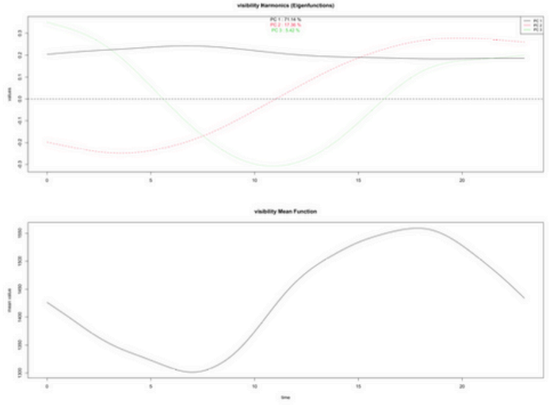
(b) Temperature FPCA



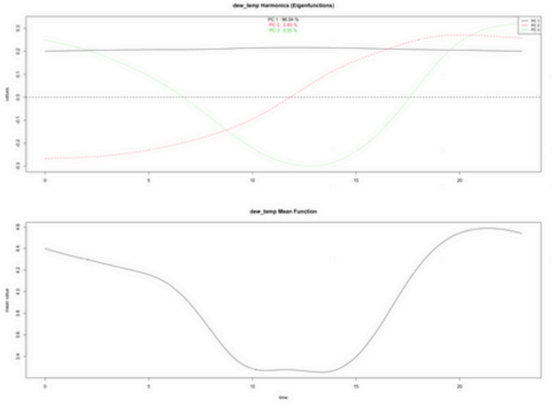
(c) Humidity FPCA



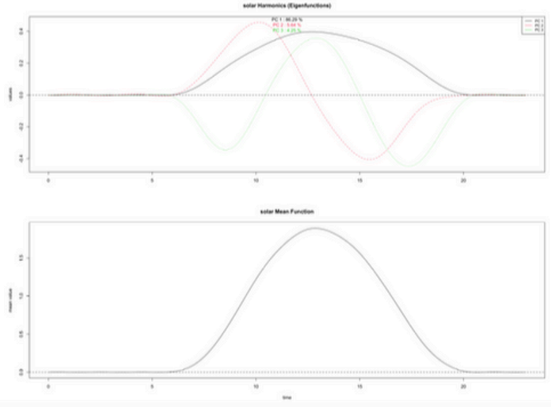
(d) Wind FPCA



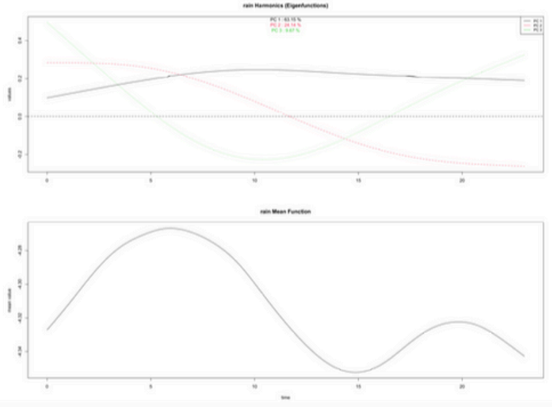
(e) Visibility FPCA



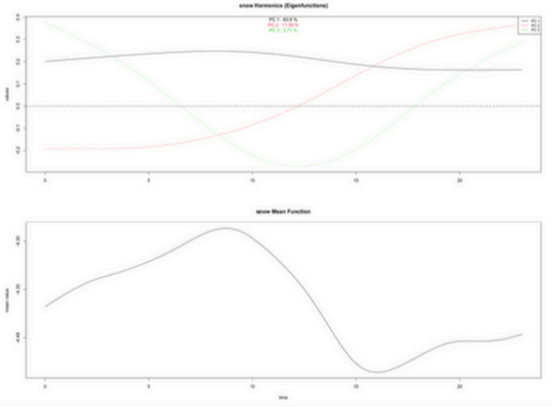
(f) Dew Temperature FPCA



(g) Solar FPCA

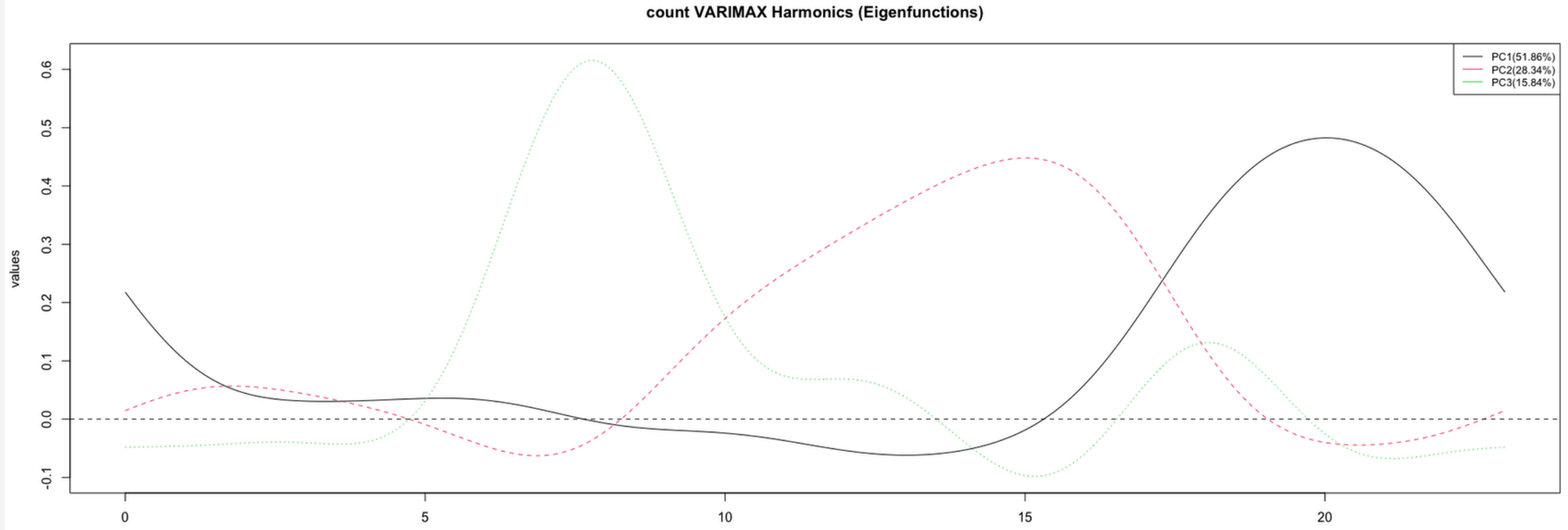


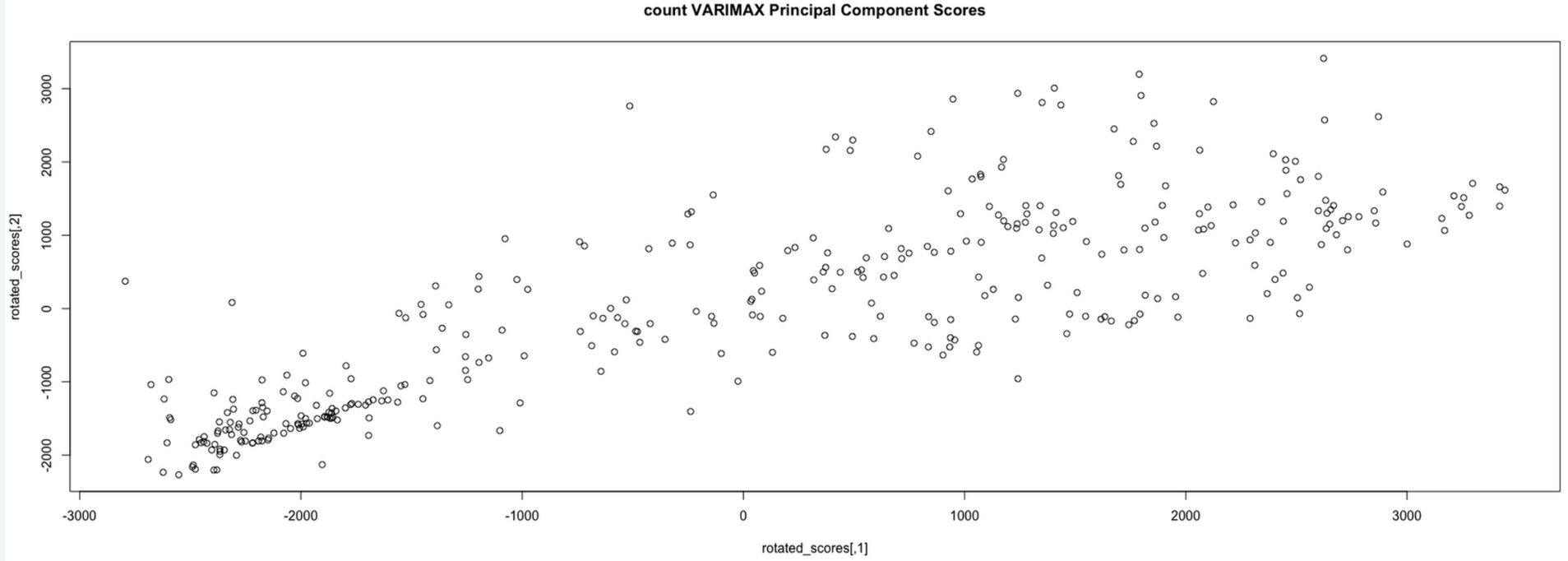
(h) Rain FPCA

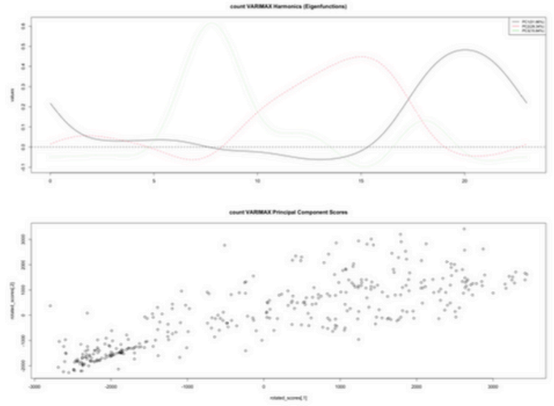


(i) Snow FPCA

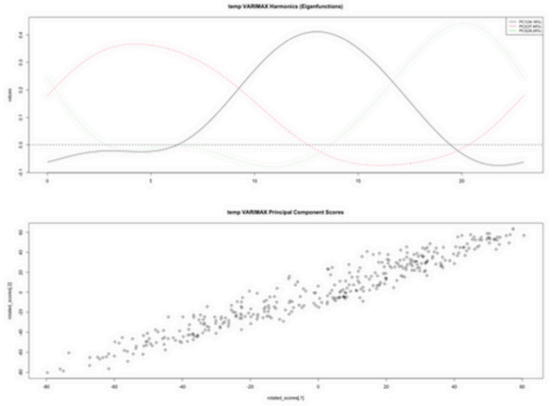
Taking the concept of FPCA one step further we also inspect the VARIMAX rotations. We are able to gain more insights and interpretation to the structure of our functional curves by maximising the variance squared of the loading of each of the components that are gained when conducting the FPCA. We also will be observing the results of the rotated harmonics and the observed values on the first principal component and the second principal component.


count VARIMAX Harmonics (Eigenfunctions)
PC1(51.86%)
PC2(28.34%)
PC3(15.84%)
values
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-0.1
0
5
10
15
20

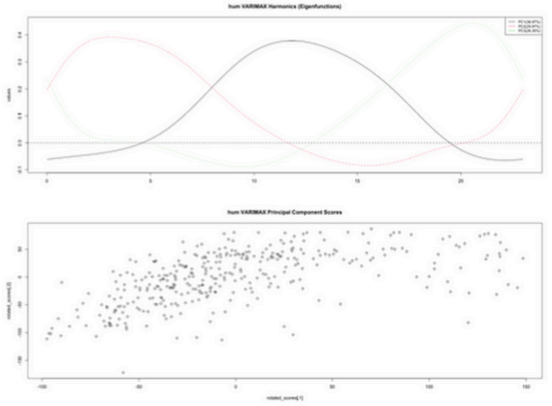

count VARIMAX Principal Component Scores
rotated_scores[,2]
3000
2000
1000
0
-1000
-2000
-3000
-2000
-1000
0
1000
2000
3000
rotated_scores[,1]



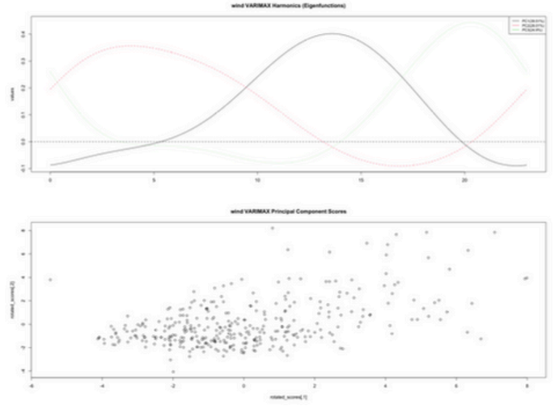
(a) Count Varimax



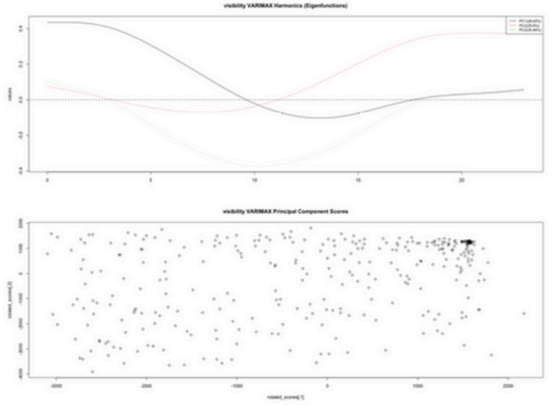
(b) Temperature Varimax



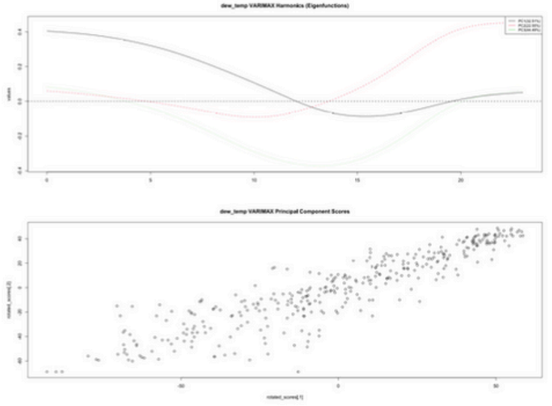
(c) Humidity Varimax



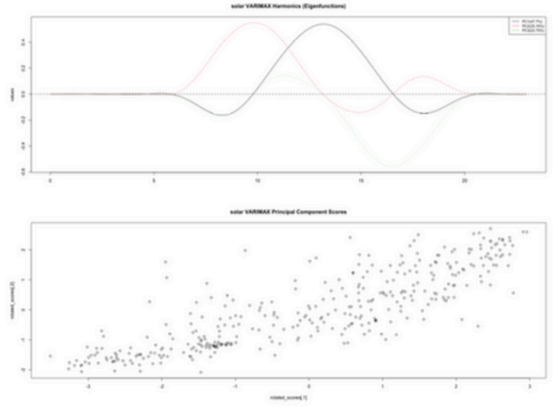
(d) Wind Varimax



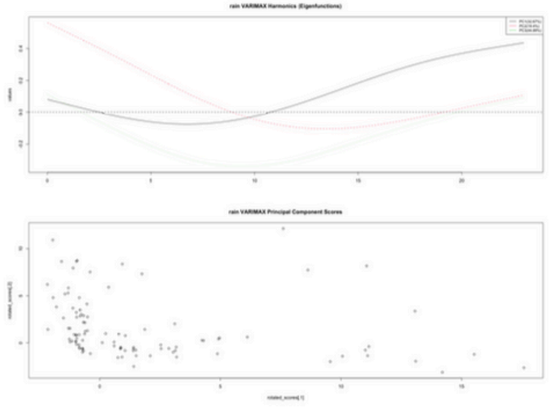
(e) Visibility Varimax



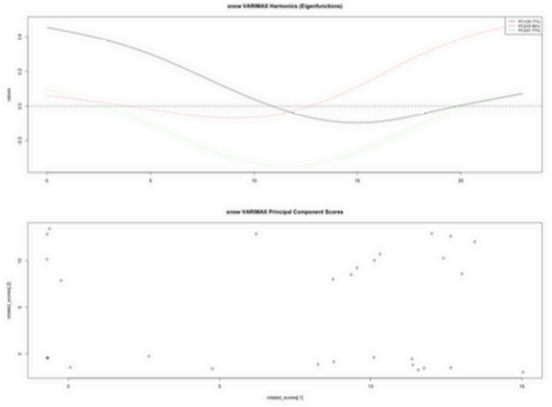
(f) Dew Temperature Varimax



(g) Solar Varimax



(h) Rain Varimax



(i) Snow Varimax

RESULTS OF FUNCTIONAL REGGRESSION

SCALAR ON FUNCTION REGRESSION

We have conducted all of the pre proccessing for our functional curves.

In this section we will be reporting the results from our scalar on functional regression.

We have used a comparative approach with a uniform basis of K = 5 for all of the covariates, which will be temperature, humidity, wind, visibility, dew temperature, solar, rain and snow.

As our response variables we have used mean of bike rented for the day.

Then we conducted scalar on function regression on the unaligned, which is the functional curves that are smoothed using the optimal lambda as roughness penalty again using the generalized cross validation technique.

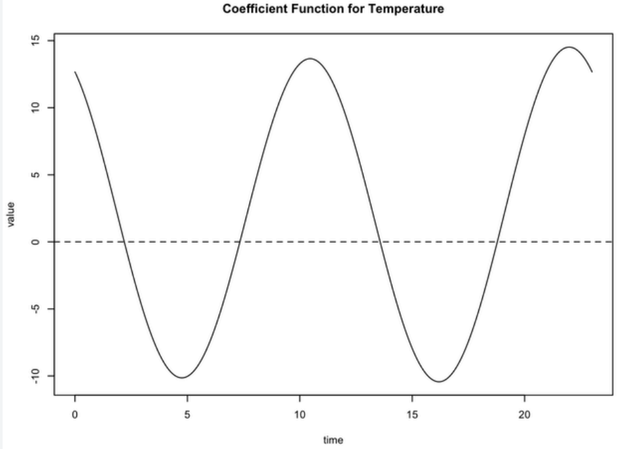
We have evaluated the models based on the RMSE, R-squared and F-ratio.

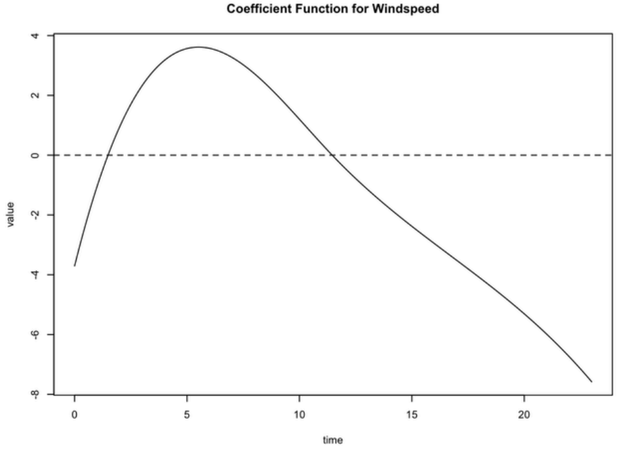
Models	RMSE	R^2	F-ratio
Aligned without Constant	185.74	0.80	31
Aligned with Constant	186.8	0.79	35.38
Unaligned without Constant	169.96	0.83	38.48
Unaligned with Constant	180.14	0.81	39.83

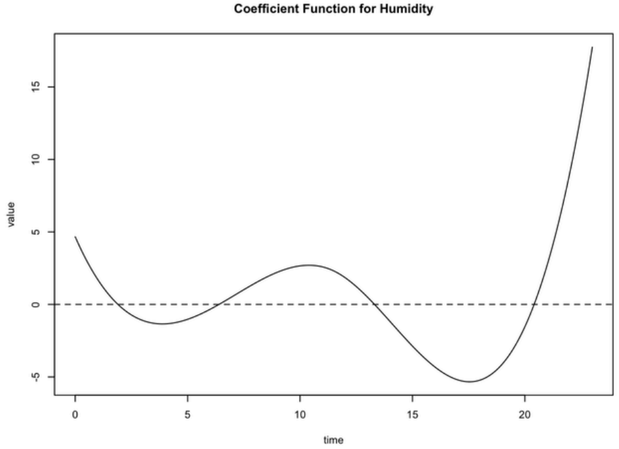
Results indicate that the models with constants have a relatively higher F-ratio.

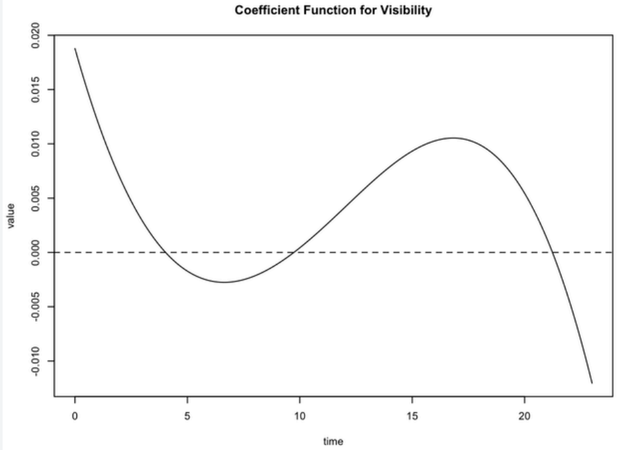
Another key takeaway is that the unaligned model perform much better than the models that have been aligned for their phase variation through the warping method.

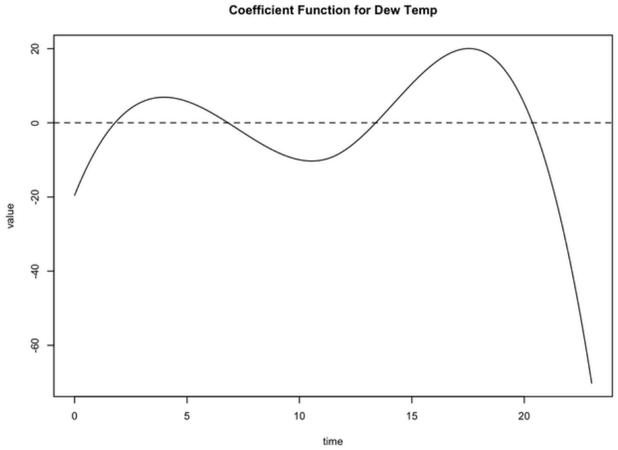
The best performing model is the unlaigned model without a contsant with 83% of the variability in the bike rental count explained by the model and average predicted value having an error of about 170 units.

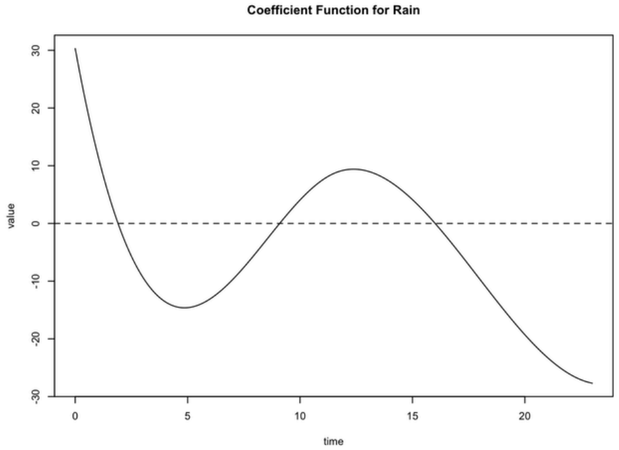

Coefficient Function for Temperature
value
time

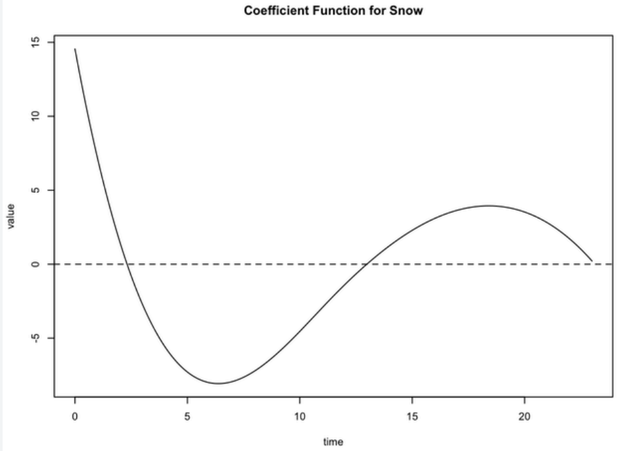

Coefficient Function for Windspeed
value
time

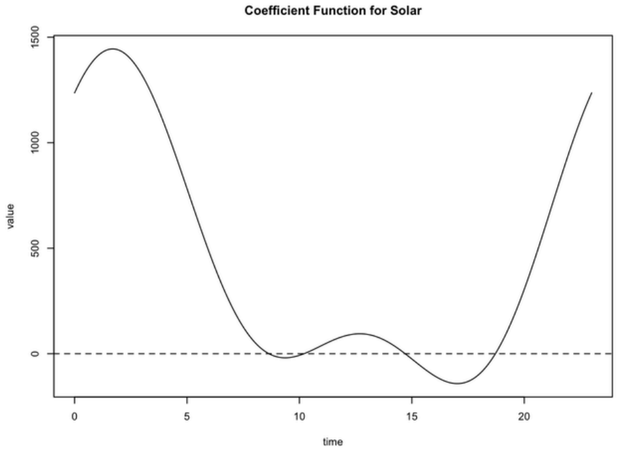

Coefficient Function for Humidity
value
time

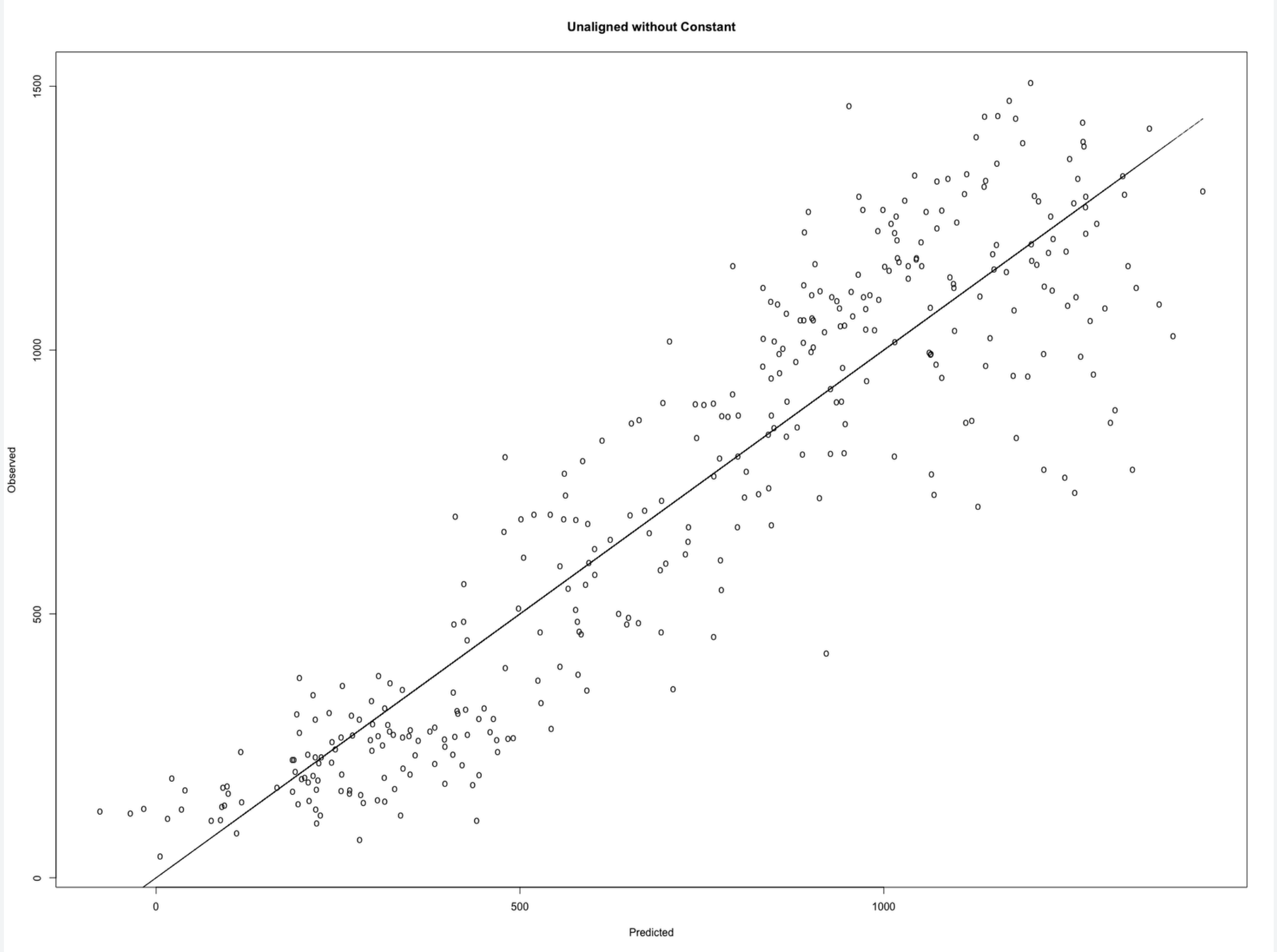

Coefficient Function for Visibility
value
time

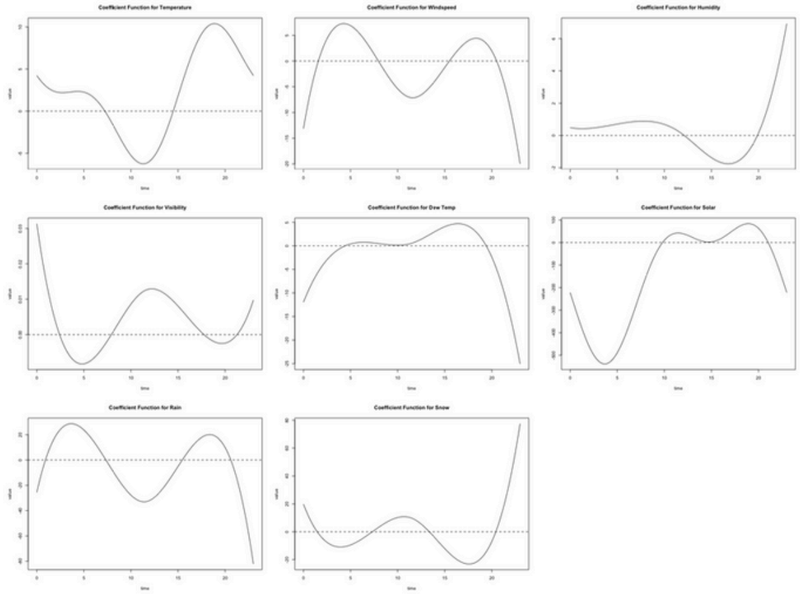

Coefficient Function for Dew Temp
value
time


Coefficient Function for Rain
value
time

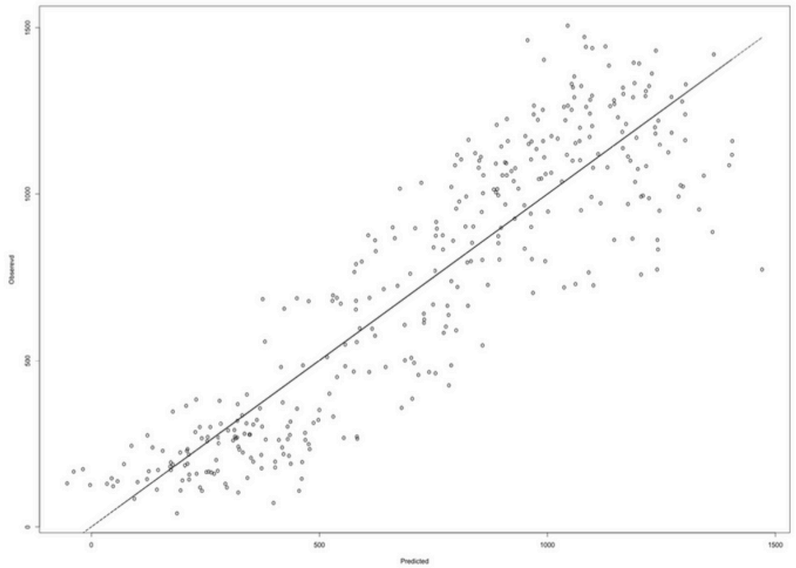

Coefficient Function for Snow
value
time


Coefficient Function for Solar
value
time

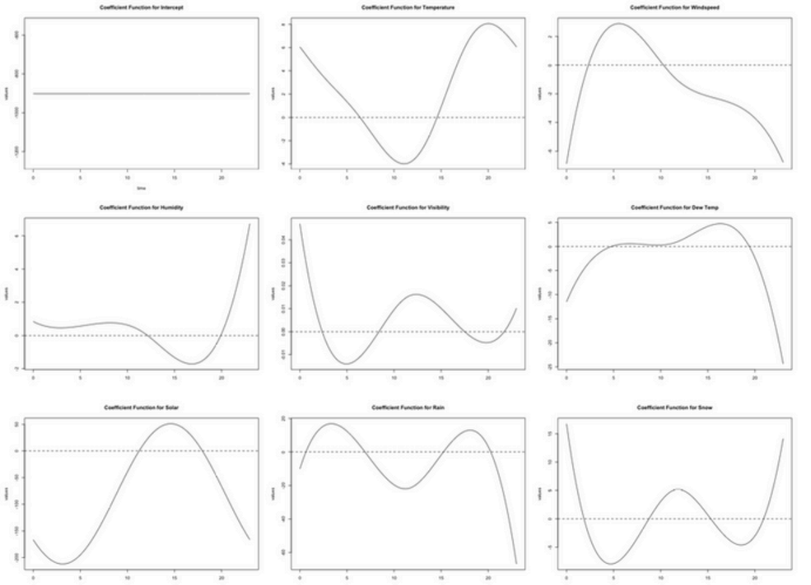

Unaligned without Constant
Observed
Predicted
0
500
1000
1500
0
500
1000



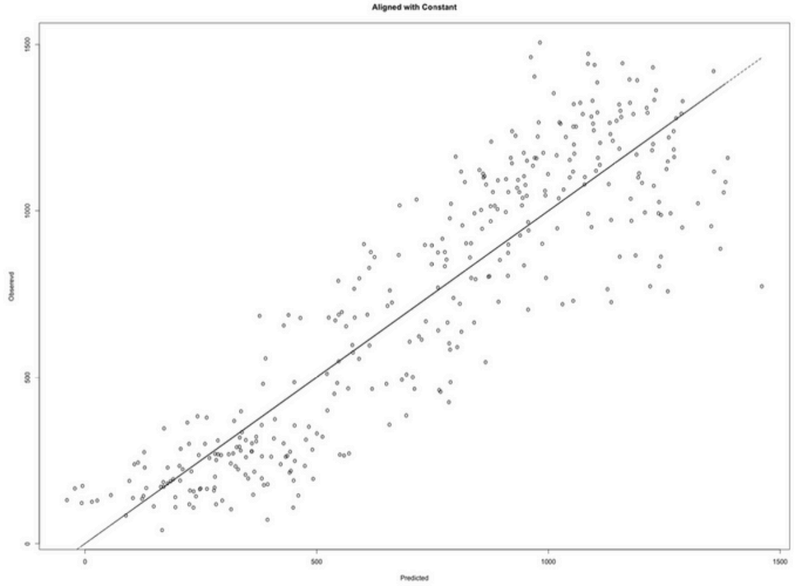
(a) Aligned Without a Constant Coefficients



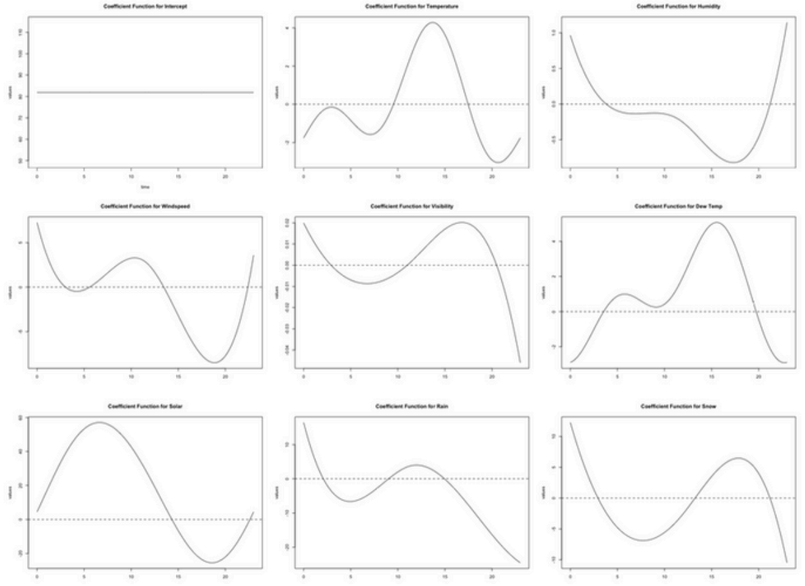
(b) Aligned without Constant Observed vs Predicted



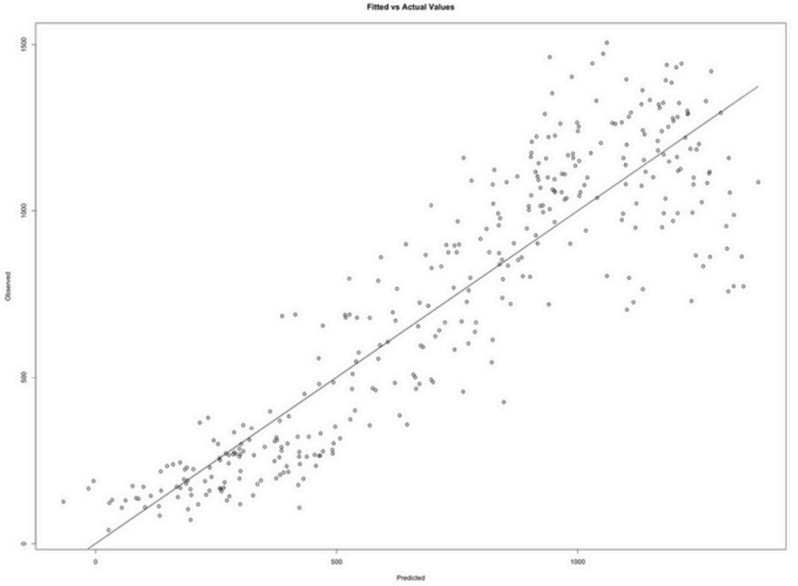
(c) Aligned With a Constant Coefficients



(d) Aligned with Constant Observed vs Predicted



(e) Unaligned With Constant



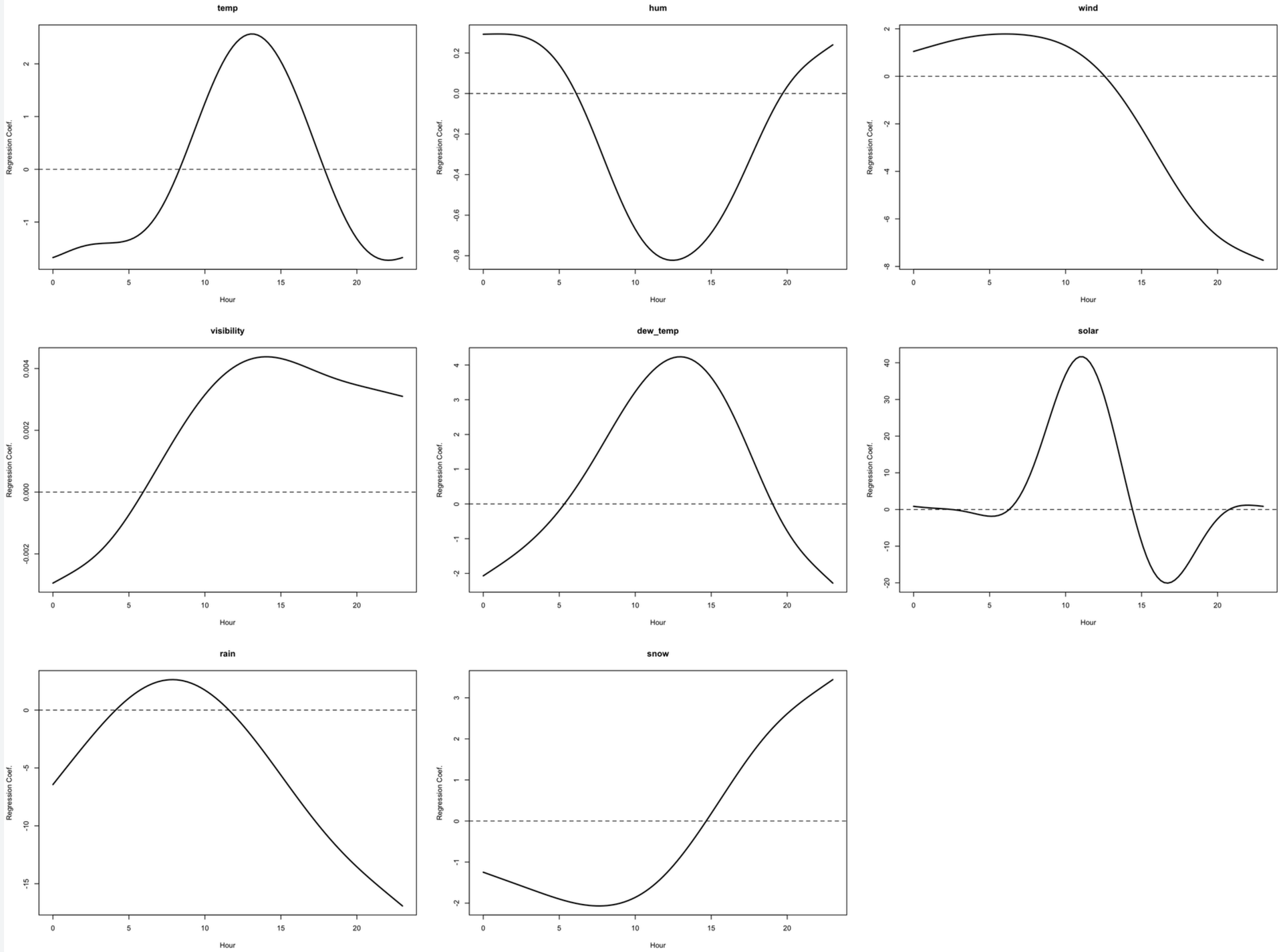
(f) Unaligned With Constant Observed vs Predicted

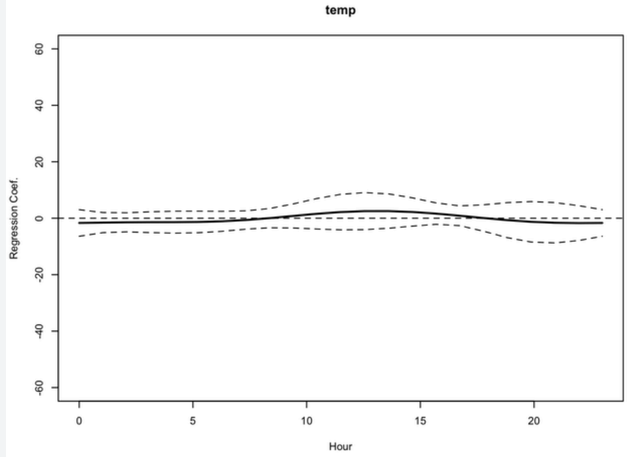
After we have seen which model performs better with our evaluation metric, we have conducted anothed scalar on function regression but this time we have used the FPCA.

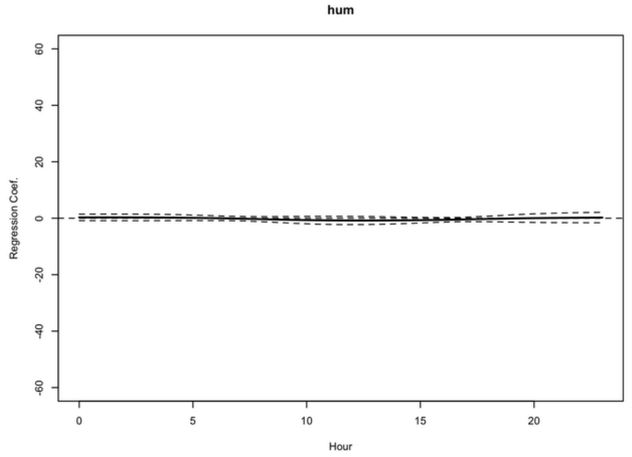
So we have relied on three of the principal scores in order to conduct the regression on our scalar response.

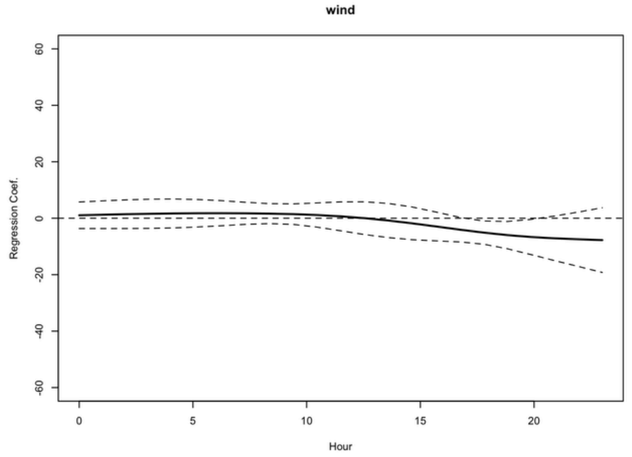
We have analysed the coefficients and the confidence interval for the results.

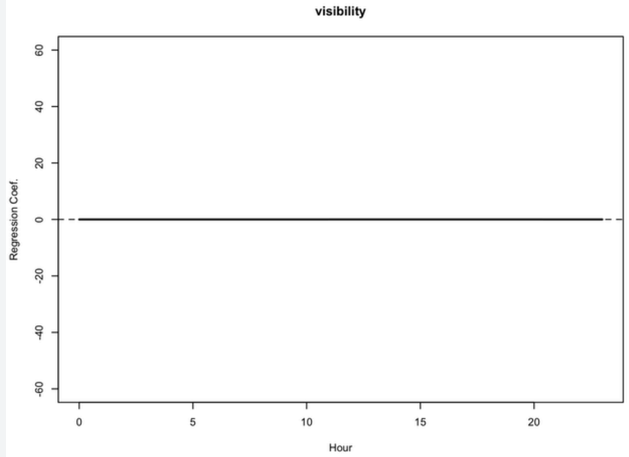
We saw that all of the bands of the confidence interval contain zero, except for specific hours for windspeed, rain and solar.

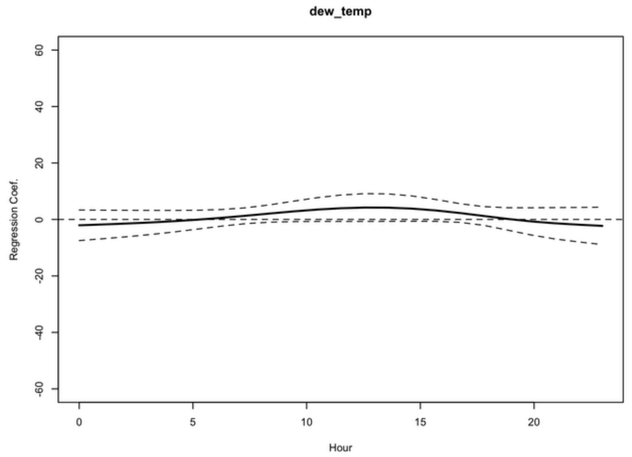

temp
hum
wind
visibility
dew_temp
solar
rain
snow
Regression Coef.
Hour

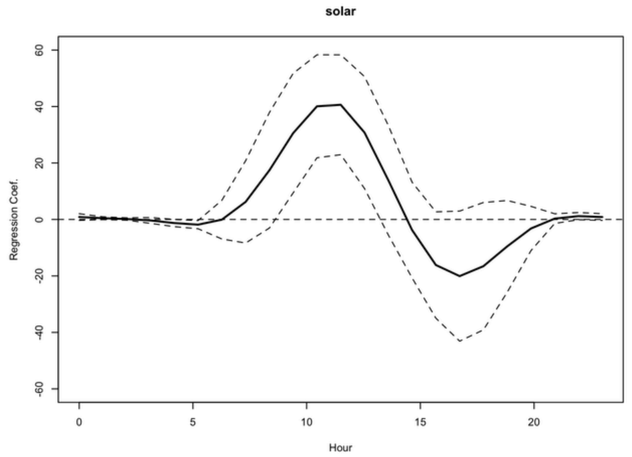

temp
Regression Coef.
60
40
20
0
-20
-40
-60
0
5
10
15
20
Hour

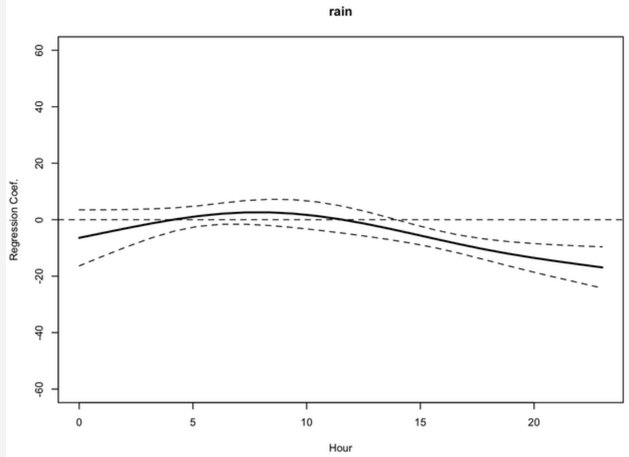

hum
Regression Coef.
60
40
20
0
-20
-40
-60
0
5
10
15
20
Hour

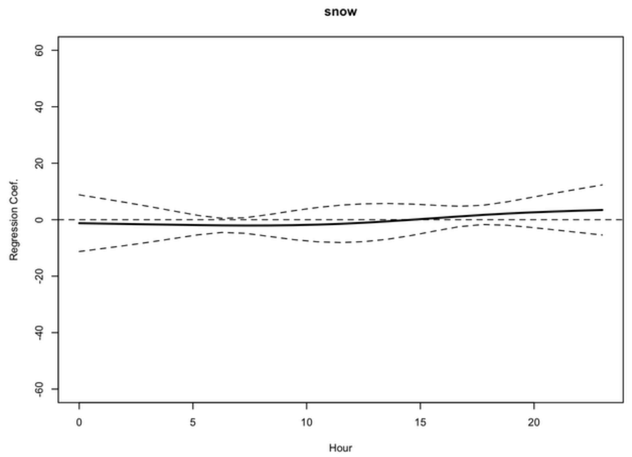

wind
Regression Coef.
60
40
20
0
-20
-40
-60
0
5
10
15
20
Hour


visibility
Regression Coef.
60
40
20
0
-20
-40
-60
0
5
10
15
20
Hour


dew_temp
Regression Coef.
60
40
20
0
-20
-40
-60
0
5
10
15
20
Hour


solar
Regression Coef.
60
40
20
0
-20
-40
-60
0
5
10
15
20
Hour


rain
Regression Coef.
60
40
20
0
-20
-40
-60
0
5
10
15
20
Hour

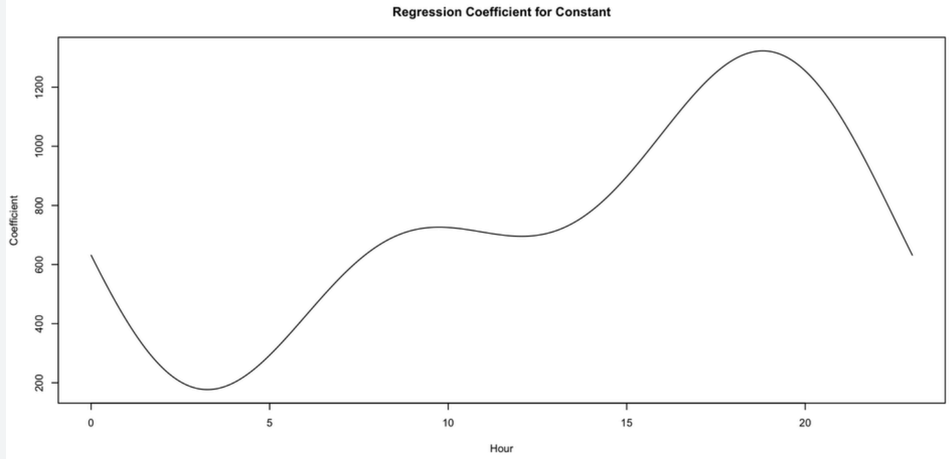

snow
Regression Coef.
60
40
20
0
-20
-40
-60
0
5
10
15
20
Hour

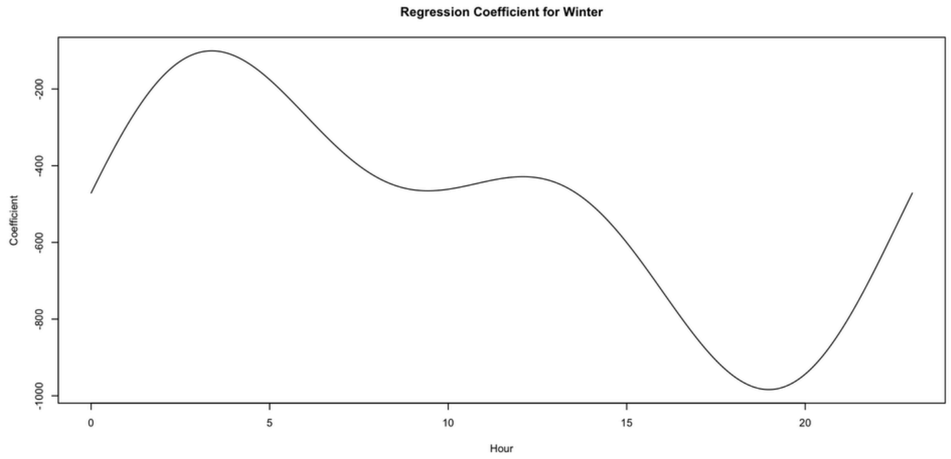
FANOVA - SEASONS

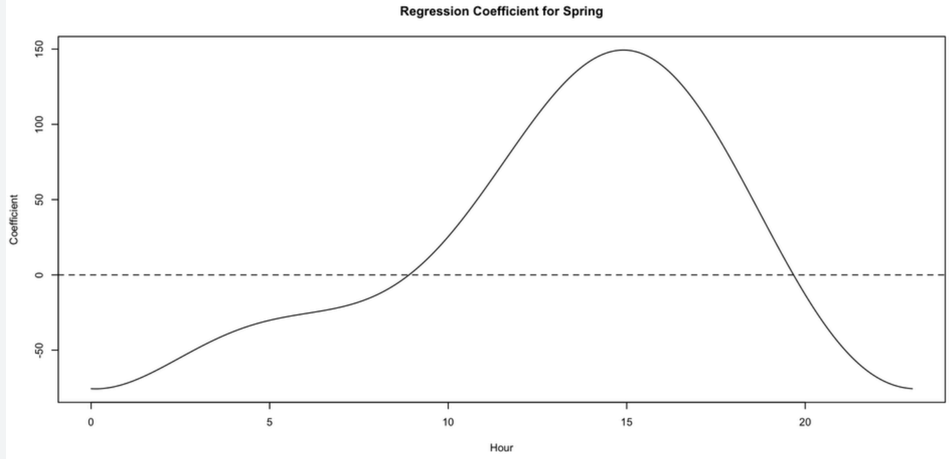
Our last analysis consists of the FANOVA on the different seasons, namely Winter, Spring, Autumn and Summer.

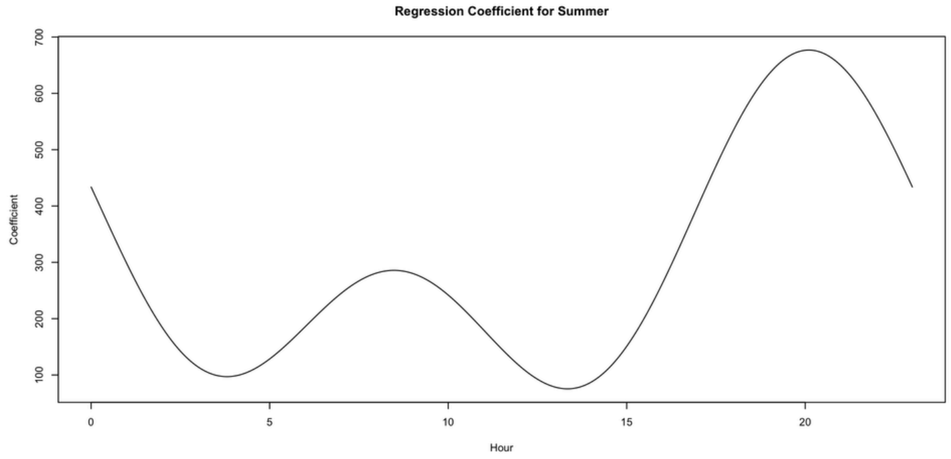
We computed the estimate, predictions and the confidence interval for each season.

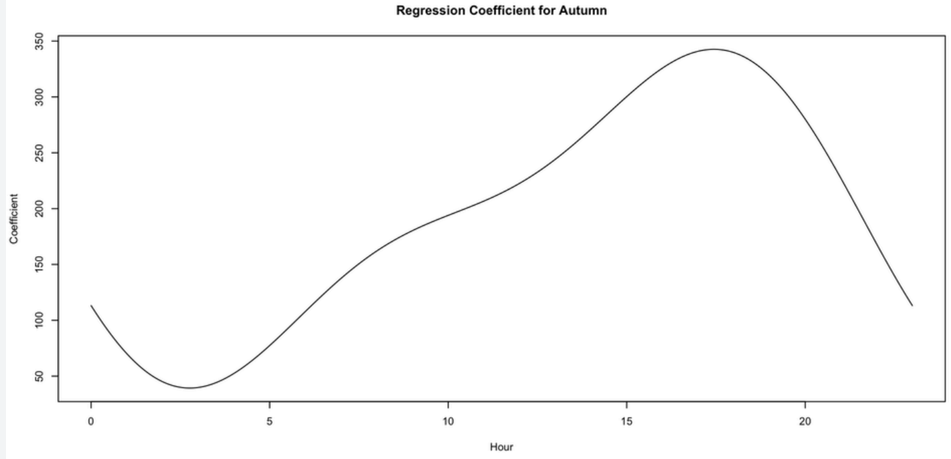
Then we finished off with a permutation F-test and t-test for Winter and Summer.

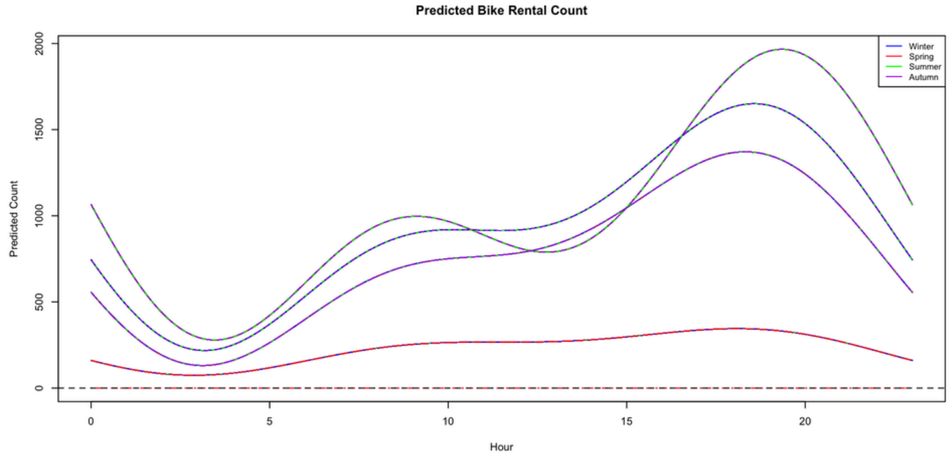

Regression Coefficient for Constant
Coefficient
200
400
600
800
1000
1200
0
5
10
15
20
Hour

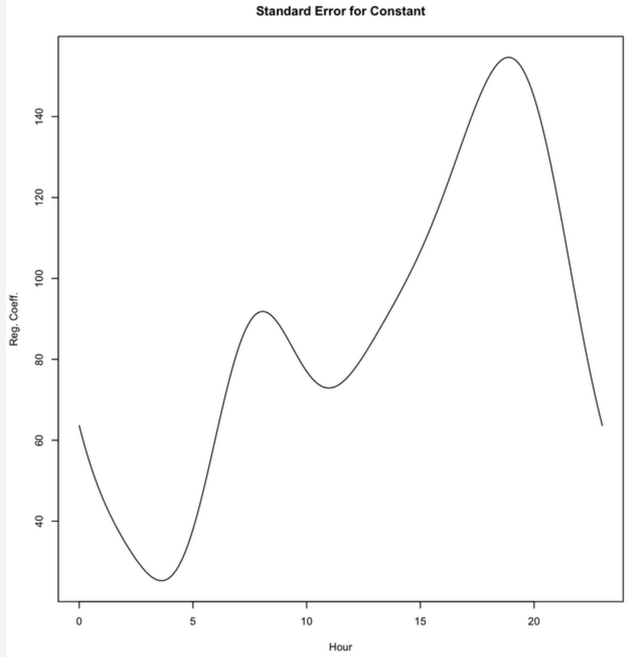

Regression Coefficient for Winter
Coefficient
-200
-400
-600
-800
-1000
0
5
10
15
20
Hour

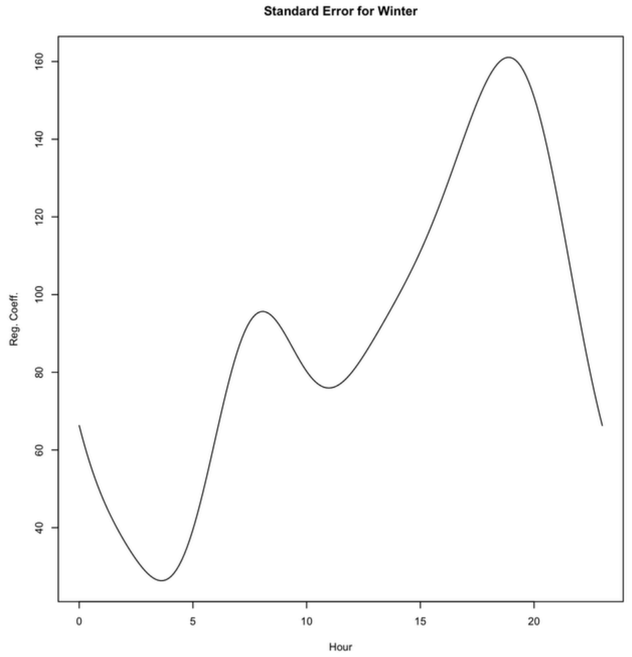

Regression Coefficient for Spring
Coefficient
-50
0
50
100
150
0
5
10
15
20
Hour

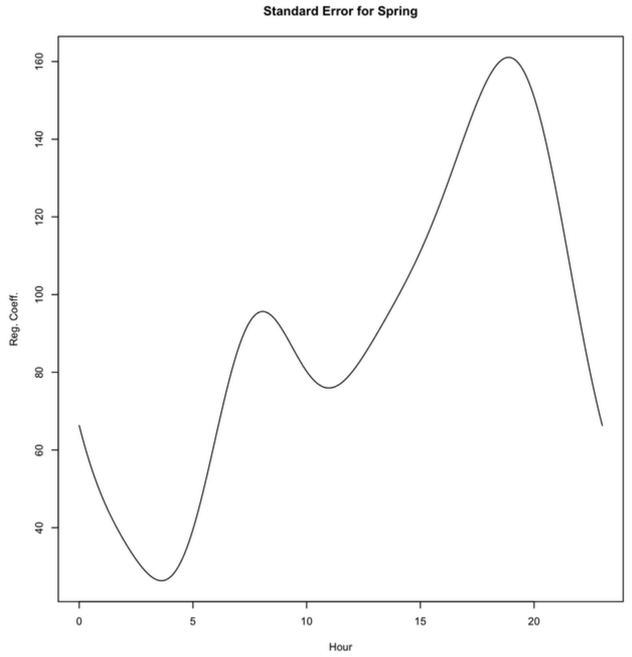

Regression Coefficient for Summer
Coefficient
100
200
300
400
500
600
700
0
5
10
15
20
Hour

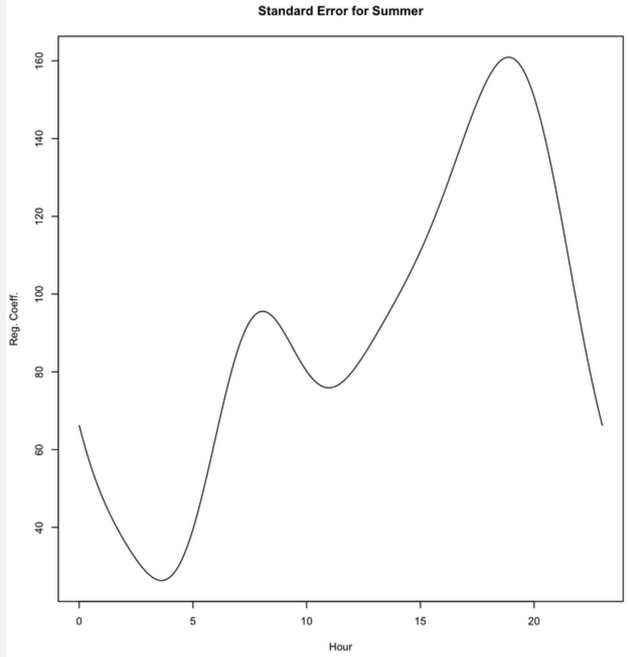

Regression Coefficient for Autumn
Coefficient
50
100
150
200
250
300
350
0
5
10
15
20
Hour

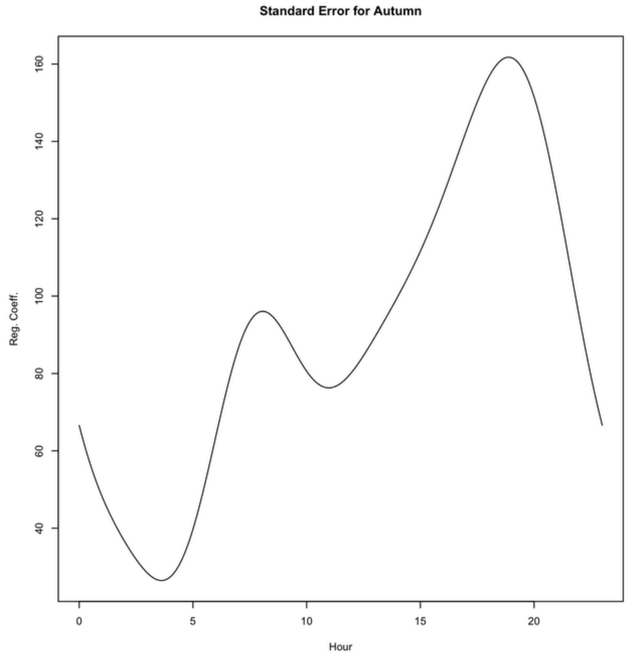

Predicted Bike Rental Count
Predicted Count
0
500
1000
1500
2000
0
5
10
15
20
Hour
Winter
Spring
Summer
Autumn

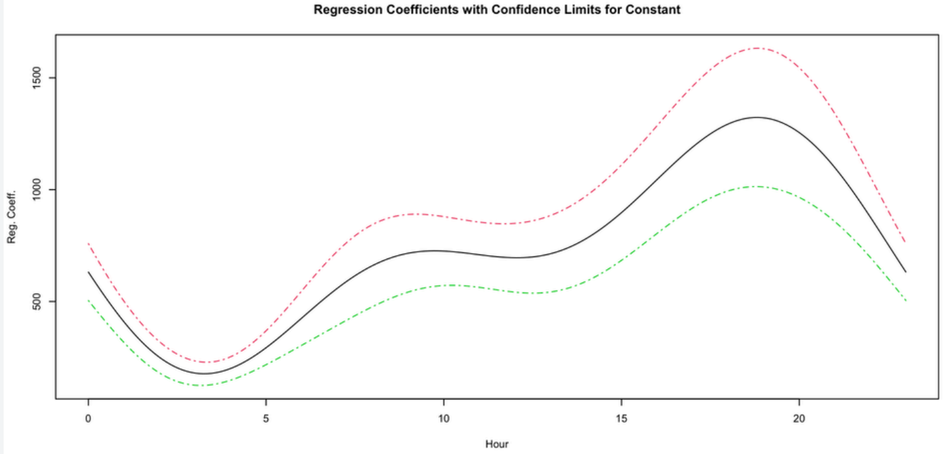

Standard Error for Constant
Reg. Coeff.
40
60
80
100
120
140
0
5
10
15
20
Hour

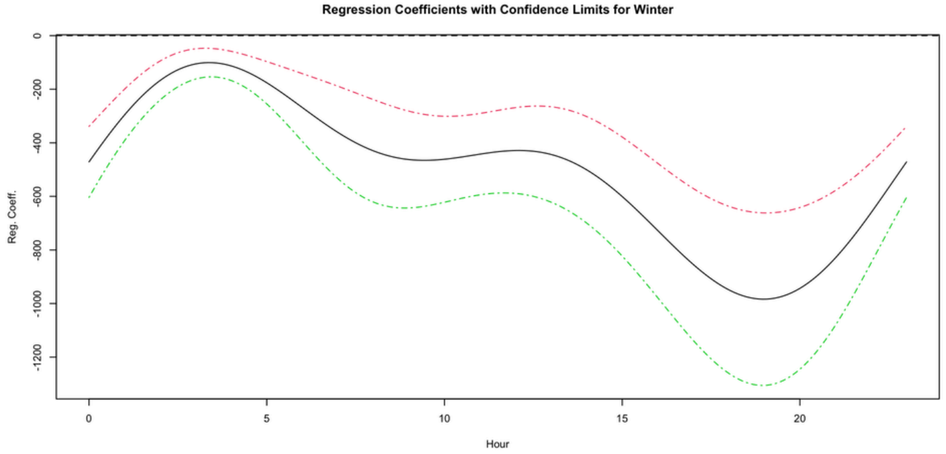

Standard Error for Winter
Reg. Coeff.
40
60
80
100
120
140
160
0
5
10
15
20
Hour

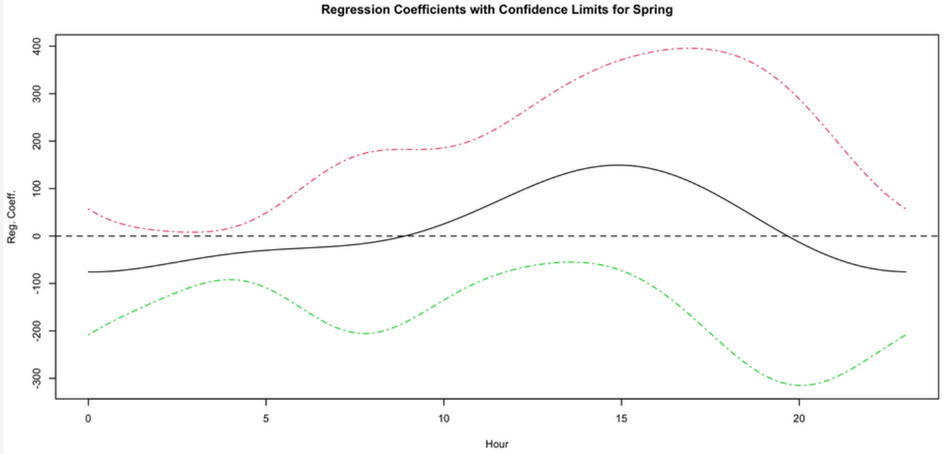

Standard Error for Spring
Reg. Coeff.
40
60
80
100
120
140
160
0
5
10
15
20
Hour

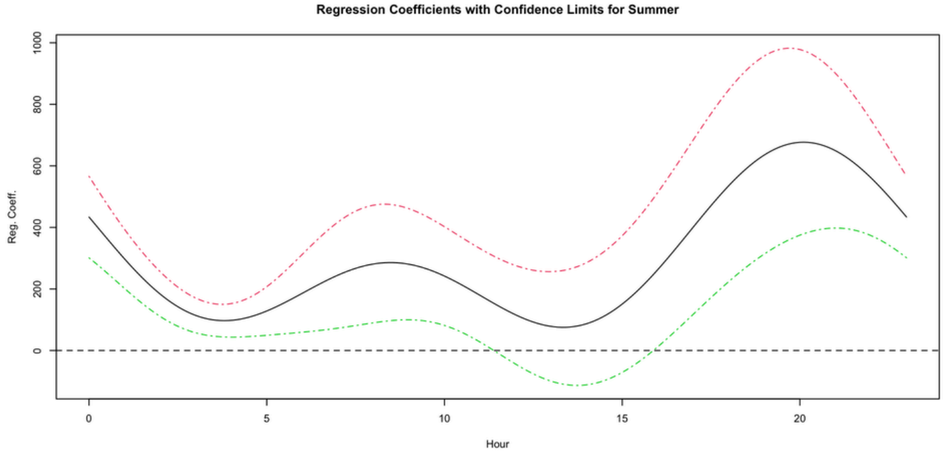

Standard Error for Summer
Reg. Coeff.
40
60
80
100
120
140
160
0
5
10
15
20
Hour

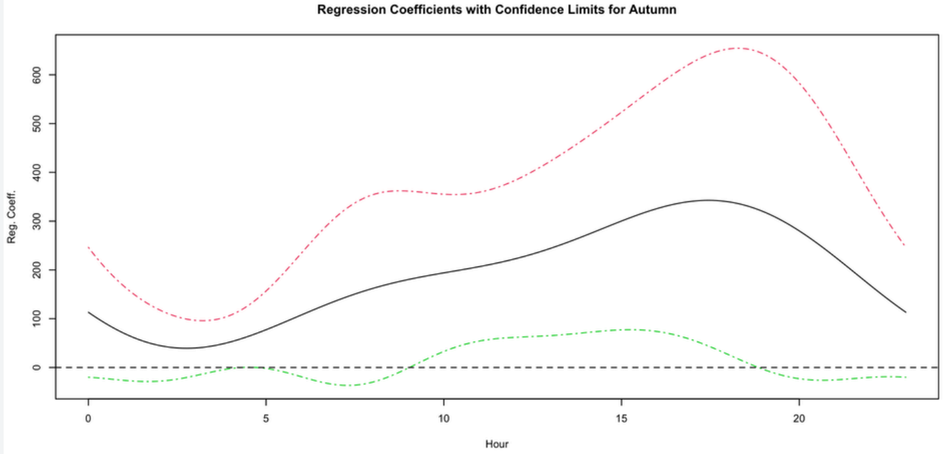

Standard Error for Autumn
Reg. Coeff.
40
60
80
100
120
140
160
0
5
10
15
20
Hour

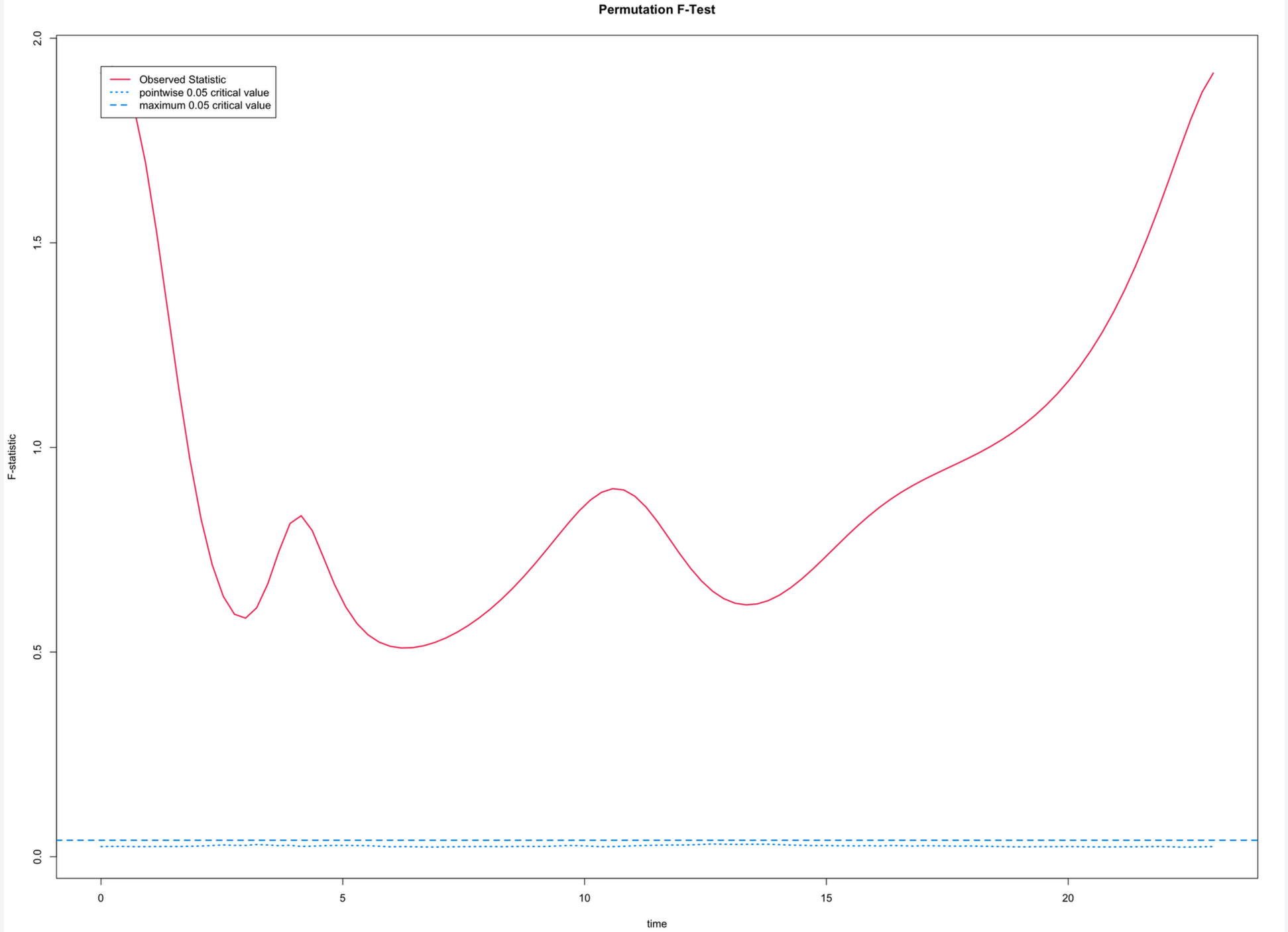

Regression Coefficients with Confidence Limits for Constant
Reg. Coeff.
500
1000
1500
0
5
10
15
20
Hour

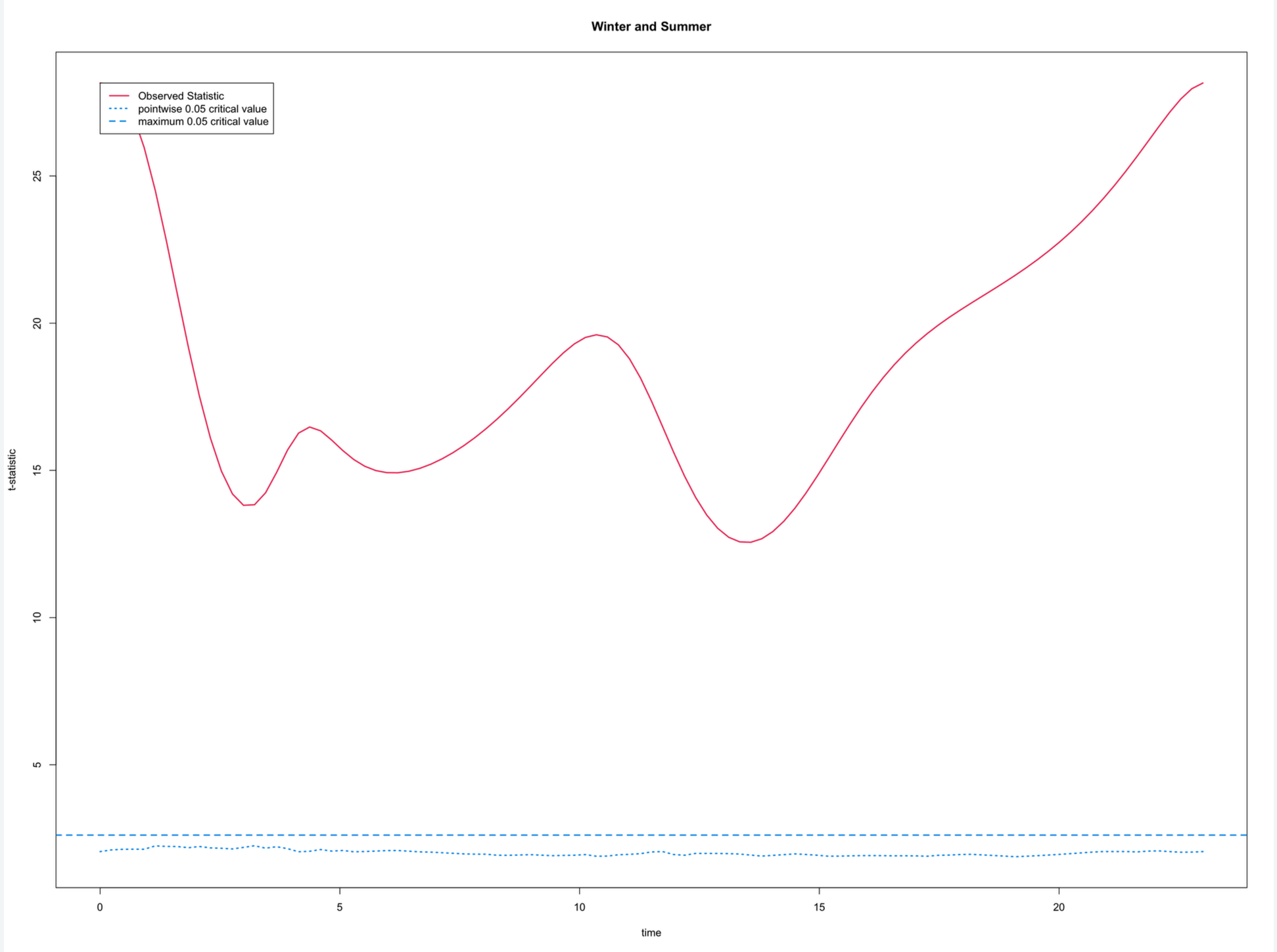

Regression Coefficients with Confidence Limits for Winter
Reg. Coeff.
0
-200
-400
-600
-800
-1000
-1200
0
5
10
15
20
Hour


Regression Coefficients with Confidence Limits for Spring
Reg. Coeff.
400
300
200
100
0
-100
-200
-300
0
5
10
15
20
Hour


Regression Coefficients with Confidence Limits for Summer
Reg. Coeff.
1000
800
600
400
200
0
0
5
10
15
20
Hour


Regression Coefficients with Confidence Limits for Autumn
Reg. Coeff.
600
500
400
300
200
100
0
0
5
10
15
20
Hour


Permutation F-Test
Observed Statistic
pointwise 0.05 critical value
maximum 0.05 critical value
F-statistic
0.0
0.5
1.0
1.5
2.0
0
5
10
15
20
time


Winter and Summer
Observed Statistic
pointwise 0.05 critical value
maximum 0.05 critical value
t-statistic
5
10
15
20
25
0
5
10
15
20
time

CONCLUSION

In this project we analysed the bike rental counts in the city of Seoul and regressed it to the weather conditions that were observed in the city.

We have soothed with a moderate and an over saturated bais with roughness penalization using the GCV method to detect the optimal lambda.

We have also aligned the data based on the phase variation using the time warping method.

We also assess the FPCA truncating the representation of our functional curves by capturing the most variation. Compaed the results with the VARIMAX to find that they are more distributed.

We applied scalar on functional regression on both the aligned and unaligned curves to find that the unaligned curves without constant perform the best. FPCA indicated that rian, windspeed were statistically sifinicant for certain hours of the day.

FANOVA results indicate that the confidence interval for the constant does not include zero in the band suggesting that the there is the overall mean of the different season is statistically signidicant.
F-test and t-test between Winter and Summer also show that there is a pointwise significant throught the day.

It would of been interesting to include interaction terms and also use a concurrent model to compare the results with the different approaches we have used.

THANK YOU!