Application of Functional Data Analysis:

Ontario Forest Fire

Jenny Huang

Brock University

**Abstract:**

2023 Canadian wildfires had the most area burned in Canada’s recorded history and not only posed a serious threat to human life, wildlife, the environment, and property, but affected air quality as well. In this study, we apply the functional linear regression model approach to learn about the number of forest fires each week using the precipitation curves in Ontario.

**Keywords:**

Functional Linear Regression Model, Functional Data Analysis

**Introduction:**

The motivation of the project comes from the 2023 Canadian wildfires, which caused serious air pollution problems and damage in wildlife. According to Canadian Interagency Forest Fire Centre (CIFFC), it had the most area burned in Canada’s recorded history. Moreover, the air pollutant was carried as far as New York in the USA and over the Atlantic. We are concerned about this topic and would like to further study the relationship between the number of forest fires and the past precipitation curves.

In this study, we apply the functional linear regression model approach to learn about the number of forest fires each week using the precipitation curves in Ontario. The Canadian National Fire Database provides information of fires of all sizes, which includes fire locations (point data), fire perimeters (polygon data) and data of fire occurrence for years 1959-2021. The Canadian National Fire point and polygon data are collected from Canadian fire management agencies including provinces, territories, and Parks Canada. The total number of forest fires for each week in Ontario can be calculated based on the records from the National Fire Database fire point data, which can be downloaded at https://cwfis.cfs.nrcan.gc.ca/datamart. We first filter our dataset and select the fire data obtained in Ontario and National Park located in Ontario, which includes Bruce Peninsula, Georgian Bay Is., Thousand Islands, Point Pelee, Pukaskwa, and St. Lawrence Islands in our dataset. We also create an excel to list the dates and their corresponding week number to count the total number of forest fires each week for our study.

The Adjusted and Homogenized Canadian Climate Data prepares the daily adjusted precipitation for over 460 locations across Canada for years 1950-2017. The data can be downloaded at https://open.canada.ca/data/en/dataset/d6813de6-b20a-46cc-8990-01862ae15c5f. We filter our dataset and select the precipitation data obtained in all 57 stations in Ontario and take the average of daily precipitation across all 57 stations located in Ontario, and then obtain for each week the average precipitation. The corresponding functional covariate is the weekly average precipitation curve for the previous 52 weeks. We collect the total number of forest fires in each week from 1980 to 2017. Figure 1 displays 10 randomly selected smoothed weekly average precipitation curves. The response is the logarithm of the total number of forest fires in a week and the functional predictor is the weekly average precipitation for the previous one year.

**Figure 1 Figure 2**

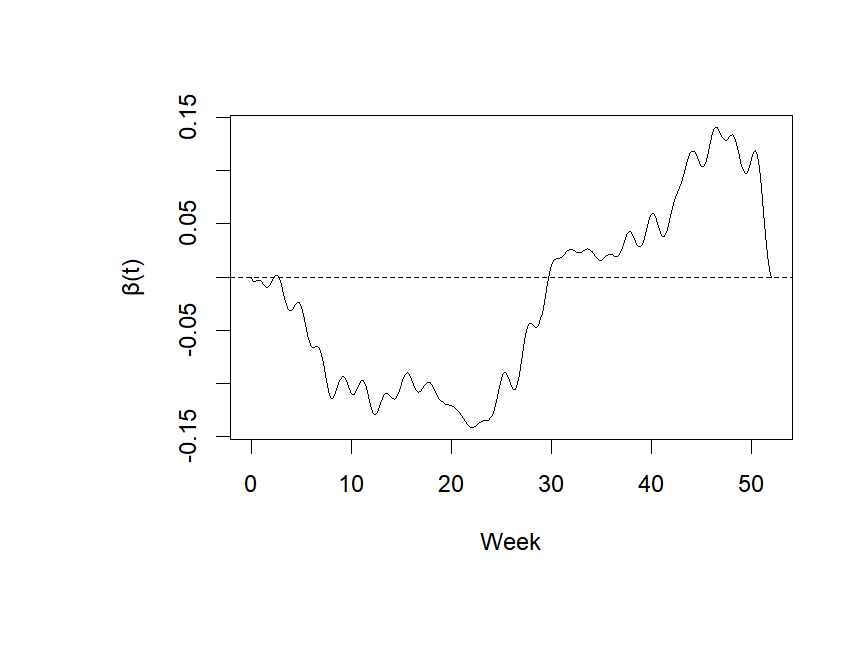
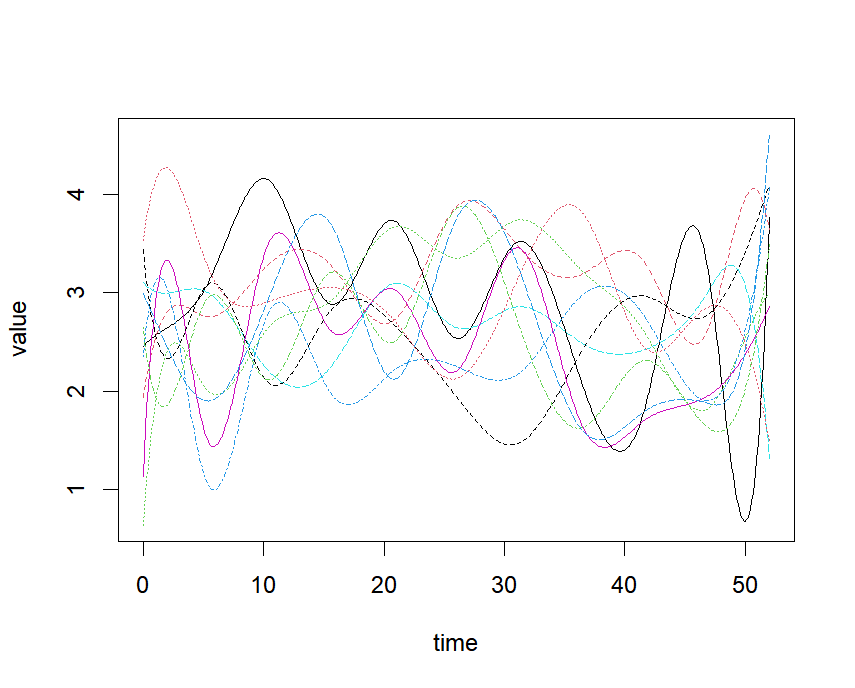


Figure 1: 10 randomly selected smoothed weekly average precipitation curves. Figure 2: The weight function that allows perfect prediction of log total number of forest fires in each week from observed annual pattern of precipitation. The arrows indicate the direction of time.

**Method:**

In our study, we use Functional Linear Regression Model (FLM) for our approach to predict the number of forest fires in each week using the precipitation curves in Ontario. Moreover, we use Linear Regression Model (LM) for comparison. LM takes the form of , which estimates the linear relationship between a scalar response and one or more explanatory variables. In our case, we take the precipitation curves into 52 explanatory variables for prediction, and LM becomes a multiple linear regression model. FLM takes the form of, which is an extension of LM with a functional covariate or response. In our case, it is a functional covariate, scalar response model. Before learning more about FLM, functional data needs to be introduced.

Functional data is multivariate data with an ordering on the dimensions. This type of data describes a process that changes smoothing, and continuously over time. Functional data analysis is a branch of statistics that analyses data that are functions, which involves repeated measures of the same process. We can use functional data analysis to study important sources of pattern and variation among the data. In our study, we use R package ‘fda’ to convert our precipitation dataset to Functional data object and get the precipitation curves as our figure 1 shown. Now our model takes the form of , which is a system of 1975 equations with 52 unknowns. The regression coefficient function is bound to be under-determined based on any finite sample . We want to fit a model of the form , where is the vector of values of log total weekly forest fire numbers predicted by the model and is a 52-vector of regression parameter estimates. In our study, we use the function ‘fRegress’ in R studio to help solve this problem.

Function ‘fRegress’ carries out a function regression analysis, where either the dependent variable or one or more independent variables are functional. All regression coefficient functions are functional as well. In our case of a scalar dependent variable, the regression coefficient for a scalar covariate is converted to a functional variable with a constant basis. All regression coefficient functions can be forced to be smooth through roughness penalties and are specified in the argument list as functional parameter objects.

We choose to use 5 folds cross validation to find our parameter for smoothing. First, we separate our dataset into 5 folds, and each has 395 sets of data. We each choose 1 fold as our test data, while the other 4 folds consider as the training data. After repeating 5 times, we can then find the smoothing parameter for our estimation. Figure 3 shows clearly the process as well.

For comparison, we use LM and consider each week as a variable by using LASSO (Least Absolute Shrinkage and Selection Operator) regression method to fit the model

to avoid overfitting. LASSO regression is a regularization technique that applies to a penalty to prevent overfitting and enhance the accuracy of statistical models. It has automating feature selection, where the bias introduced by LASSO regression will artificially shrink the coefficients towards zero. Some variables will shrink exactly to zero, leaving the model with a subset of the most important variables to make predictions.

**Figure 3**



Figure 3: 5 folds Cross Validation process

**Figure 4**  **Figure 5**

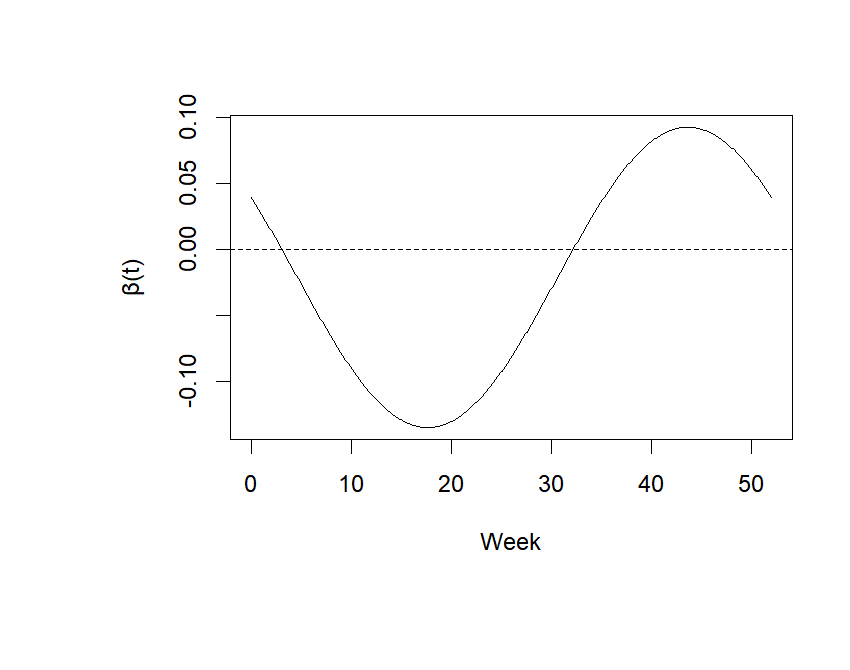
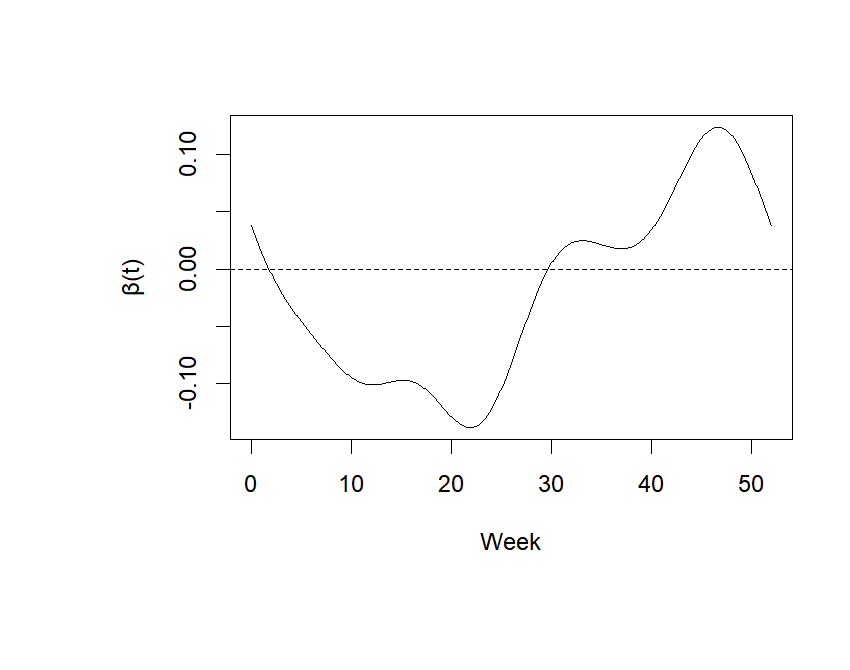


Figure 4: The graph choosing basis =13. Figure 5: The graph choosing smoothing =3. The arrows indicate the direction of time.

**Results:**

After using function ‘fRegress’ in R Studio to fit our FLM and estimate basis parameter , we get our basis parameter as 13 to get Figure 4. We then test parameter with the range of 3 to 13 and get our final smoothing parameter equlas to 3, which is shown in Figure 5. It is clear to learn that in the previous 4 months, the trend follows that when there is more precipitation, the number of forest fire will decrease. However, when the time is beyond the previous 5 months, there is little effect of the precipitation on the number of forest fire since the graph has a growing trend. When the time is beyond the previous 8 months, we could see have positive values, which means the datasets before that time are not significant.

For our estimation, we use the F test to check the overall effect of . Our F ratio for the basis parameter is 953.7926, which is much larger than the 95% quantile with the number of 2.609417. It indicates that has a significant effect on . We also test the smoothing parameter and get the F ratio of 635. 5393.The variance between groups becomes smaller after smoothing, which leads to a decrease in F ratio. However, since the second F ratio is still much larger than the 95% quantile, we could state that still has a significant effect on overall.

As we introduced in the previous section, we use LASSO regression method to fit the LM and compare the 52-vector coefficients with the FLM. Using R code to solve the coefficients, it is easy to calculate there is a 51.38% of goodness of fit. After plotting all coefficients, we get a graph in Figure 6.

**Figure 6 Figure 7**

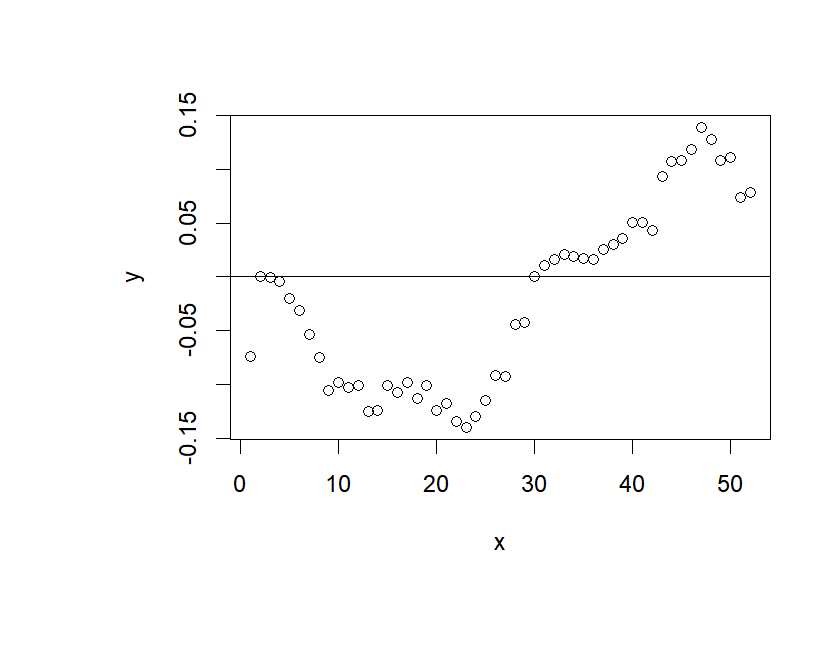
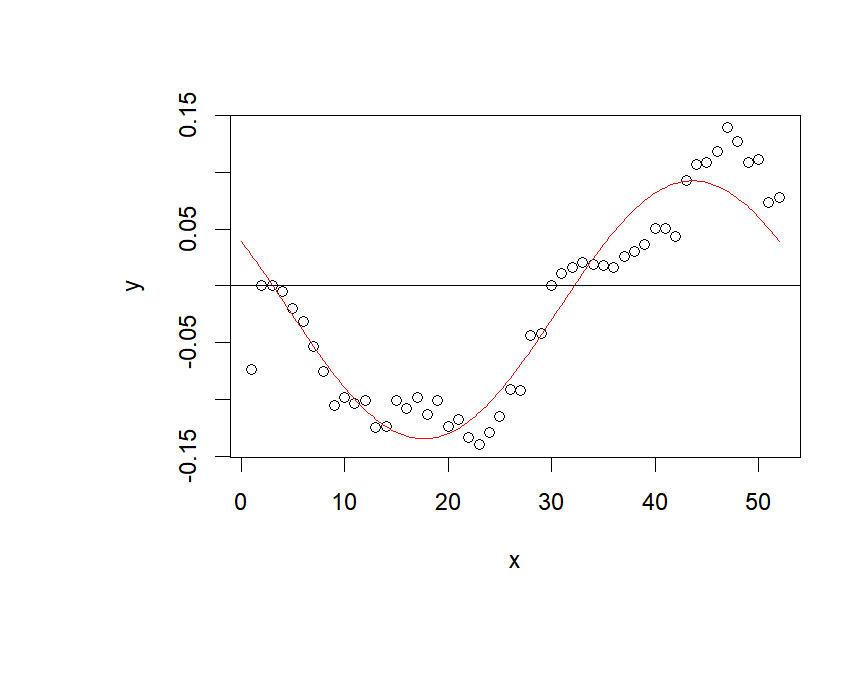
 

Figure 6: The LM coefficients graph. Figure 7: The combination of LM coefficients and smoothing FLM graph. The arrows indicate the direction of time.

According to the result, there are 2 vectors that have 0 as their solutions, which are V2 and V30 (representing the previous 2 week and previous 30 week). We decide to ignore the V2 result since the edge could get an error, especially V1 and V3 are both negative values. Luckily, we still have the V30 result as a turning point, where the coefficients after V30 become positive values. It also follows that when the time is beyond the previous 8 months (30 weeks), there is little effect of the precipitation on the number of forest fire. Figure 7 is the combination of LM coefficients and smoothing FLM graph, where we could find out that they are suitable for each other for the overall trend.

**Conclusions:**

According to the result, we can conclude that it trends to follow by a negative linear relationship, which is when the precipitation increases in the previous months, the number of forest fires decreases. Moreover, there is little effect of the precipitation on the number of forest fires when the time is beyond the previous 8 months. This is still a good estimation with the usage of Functional Linear Regression Model for analyzing in general.

However, there are still some limitations during our study. To begin with, the data about precipitation could be more accurate, since this is a dataset released in 2017, which could be different from these years’ situation. Moreover, our topic is related to the dataset in Ontario, which can filter the data into different regions since Ontario is a large province that contains different environments in each area. Last but not least, we can discuss more about the method since the LASSO regression in LM only has 51.38% goodness of fit, and the basic and smoothing parameters could be discussed more during the smoothing process in FLM.

In general, this study shows a brief relationship between the precipitation curves and the number of forest fires, which helps us predict the future number of forest fires with the previous precipitation curves. Moreover, this is an interesting application to the Functional Data Analysis that could build a model with functional data but not only limit it to several points, which could also promote more in the future study.

**References:**

Functional Data Analysis. (n.d.). . Scholars Portal.

Functional Data Analysis with R and MATLAB. (n.d.). . Scholars Portal.

CIFFC. (n.d.). https://ciffc.net/statistics

Adjusted and homogenized Canadian climate data – daily temperature and precipitation (AHCCD – daily T&P). Open Government Portal. (n.d.). https://open.canada.ca/data/en/dataset/d6813de6-b20a-46cc-8990-01862ae15c5f

Canada, N. R. (n.d.). Canadian National Fire Database (CNFDB). Canadian Wildland Fire Information System. https://cwfis.cfs.nrcan.gc.ca/ha/nfdb

Caojiguo. (n.d.). Caojiguo/HFL: Estimating historical functional linear models with a nested group bridge approach. GitHub. https://github.com/caojiguo/HFL

Caojiguo. (n.d.-a). Caojiguo/FDAworkshop: Materials for Functional Data Analysis Workshop. GitHub. https://github.com/caojiguo/FDAworkshop