
# Convolution Functional Linear Regression Model

by Maria G Asencio

---

# CONVOLUTION FUNCTIONAL LINEAR REGRESSION MODEL

A Thesis

Presented to the Faculty of the Graduate School

of Cornell University

In Partial Fulfillment of the Requirements for the Degree of

Master of Science

by

Maria G. Asencio

August 2010

# ABSTRACT

This thesis develops a new set of tools in Functional Data Analysis. A general historical functional linear model called "*convolution functional linear model"* is developed to handle a stationary response to the recent history of a covariate process. We argue that the dependent variable $y_i(t)$ is continuously related to values of covariate $\xi_{iP}(w)$ where $w \in [t-\delta_p, t]$ explained by $\beta_p(w)$. $\beta_p(w)$ is the coefficient function of the convolution of the independent variable and give a weight to the values of $\xi_{iP}(w)$ for the estimation of $y_i(t)$ as $w \to t$. The coefficient of the convolution $\beta_p(w)$ as well as $\delta_p$ are unknown and estimated by least squares and cross-validation methods. In addition, diagnostic methods are also developed with the purpose of estimating the variability of $\beta_p(w)$, once $\delta_p$ has been estimated, and the effects that each curve sample has over model estimation or solution. In particular, the methodology of the non-overlapping block bootstrapping (NBB) was expanded to provide a solution for the distortion caused by dividing the sample data into blocks of length "$l$" and resample them with replacement. This method handles the discontinuity caused at the interior points where the blocks ordinarily should be connected in the block bootstrapped sample. In other words, the block bootstrapped sample becomes less rough and maintains the smoothness of the original data. In addition, the block bootstrapped resembles the autocovariance structure of the original data allowing estimating accurate confidence interval for $\beta_p(w)$. The statistical methods of Cook's and Mahalanobis distance also were modified with the objective of accounting for the time structure and serial correlation of the data. This was specifically done by restricting the estimation of the covariance using the autocovariance estimates. These two methods allow identifying curve or functional

data that are causing a significant effect in the solution and estimates of the model. Finally, the model was used for the estimation of the continuous trajectory of particulate matter given driving behavior variables. Estimates of the parameters and diagnostics for this particular case are provided.

# BIOGRAPHICAL SKETCH

Maria Asencio started to have a fascination in mathematics since early years. This sentiment was reborn when she started a Computer Science major at the University of North Texas (UNT). The extensive requirement of mathematics classes stir up her excitement and passion with this subject. This let to strive for a double major in Mathematics and Computer Science. She had the opportunity to take a range of classes and work in projects that allow her to explore and acknowledge the usefulness of these two subjects in other areas such as environmental sciences and biology. As a McNair Scholar, she became a research assistant in Dr. Acevedo's lab and was involved in a research project for the development of a mathematical model for the vegetation regeneration. There, she learned different statistics methods under the guidance of Dr. Monticino while she contributed to the development and simulation of the statistical model. She wanted to proceed further in her purpose of working as a research assistant for projects with the objective of modeling different environmental dynamics. This led her to pursue a Master degree in Statistical Science at Cornell University. She acquired a strong theoretical and computational background in statistics. Specifically, she gained knowledge and had the opportunity to work in research projects with Dr. Giles in the area of Functional Data Analysis. This led to the development of the 'convolution functional linear model" for the prediction of responses that have been perturbed and their recordings do not represent the effects of the instantaneous changes of the each stimulus.

This thesis is dedicated to my dear parents, Mirna and Rafael, and two brothers, Alvaro and Raul

# ACKNOWLEDGMENTS

I would like to thank first of all to my family for their unconditional love and support. They are important key for my success and achievements. I am particularly grateful to Dr. Giles Hooker whom has been a great supporter in all these years at Cornell University. Under his guidance, I have the opportunity of working on research projects in the area of Functional Data Analysis. I want to thank Dr. Oliver Gao for letting me work in the project of modeling particulate matter and for his helpful recommendations. I want to thank the Sloan Foundation for their great support. A very important thanks to Dr. Acevedo and Dr. Monticino for showing me the path of research. Thank you to Ms. Judy and Ms. Diana from the McNair program that helped and guided me to pursue post-bachelor degree.

# TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# CHAPTER 1

# LITERATURE REVIEW

There is an interest on studying quantities measured continuously over time. This allows observing changes and variation on their behavior under specific conditions or dynamics. Function Data Analysis (FDA) takes advantage of the continuity and smoothness that this serial recording has and presents it as a function of time instead of discrete points [4]. Representing the data as a curve allows observing features and characteristics of the data that are important for the analysis and understanding the mechanism that is causing specific behavior. Functional data analysis has also propagated in the statistical area of linear regression model in which some or all variables are represented as a function of time [4]. We explain several functional linear regression models for specific type of relation between the dependent and independent variables.

Furthermore, we discuss about the development of emission models to predict the amount of emissions a vehicle can produced. These models are an important key for the planning and designing of air-quality programs that intend to control or minimize the emission levels in the air [3]. Emissions models give feedback about the appropriateness of the project prior to implementation. These models generally rely on chassis dynamometer measurements to inquire about how several factor such as driving behavior variables affect the production of the emissions in vehicles. Chassis dynamometer measurements allow conducting several experiments under almost the same conditions. However, it is also important to notice that the emissions' travel

time is affected while they are being transported toward the equipment that keeps track of them [3]. We discuss this problem and the different categories in which these models fall.

## 1.1 FUNCTIONAL DATA ANALYSIS

Most classical statistics methods analyze the data gathered in a sequential manner as a sample of i.i.d. points ignoring the order in which they were recorded. However, taking into account the time structure permits to observe whether the data have a periodic pattern or relevant peaks and it gives a better picture of the importance of the events in a continual manner. Several statistical and mathematical methods have been developed with the purpose of handling this time dependency. One of these methods is functional data analysis which is discussed by Ramsay and Silverman in their book "Functional Data Analysis (FDA)"[4]. FDA methods have found to be significantly useful in different research areas having the ability to represent complex patterns and curve structure by executing merely the least square method.

For example, consider the observations $y_j, j = 1, \dots, n$. The goal of FDA is to represent these discrete points as a function $x(t)$ [4]. This is accomplished by setting $y_j = x(t_j) + \varepsilon_j = \sum_{k=1}^{K} c_k \varphi_k (t_j) + \varepsilon_j$ and minimizing for $c_k$'s the expression $\sum_{j=1}^{n} (y_j - \sum_{k=1}^{K} c_k \varphi_k (t_j))^2$ [4]. The discrete points $y_j$ are linearly related to a set of $k$ basis functions $\varphi_k$. There exists several types of basis functions and their selection and definition depend on the curve structure of the data. It is common to use cubic B-splines for nonlinear curves since they have the ability to represent any pattern by having the same mechanism of polynomials. As a result, this method allows defining

a continuous function $x(t)$ with infinite elements by using finite number of basis function. We have the function definition of $x(t)$: $(t_1, t_n) \rightarrow (y_1, y_n)$.
Looking at the data as a function of time has also been adapted or expanded in the area of linear regression. The distinction between functional linear regression and classical linear regression is that one or more of the linear equation's components are represented as a function of time [4]. Several type of functional linear regression have been developed and in particular, we consider the following type of model:

$$y_i(t) = \alpha(t) + \int_a^b X_i(s)\beta(s,t)ds + \varepsilon_i(t), t \in [0,T].$$

We see that in this case all the components are considered as function of time and that the influence of $X_i(s)$ in $y_i(t)$ is accounted by the surface function of coefficients $\beta(s,t)$. By changing the values of $(a,b)$, a specific relation between the variables can be defined.

The simplest or less complex case is when $a = b$ which implies that $s = t$ and thus, current values of $y_i(t)$ are influenced only by the current values of the predictor [4]. This relation is expressed as the following:

$$y_i(t) = \alpha(t) + X_i(t)\beta(t) + \varepsilon_i(t), t \in [0,T].$$

The significance of the model is that the values of the coefficients are time variant which implies that the predictors do not have a fixed effect for the response [4].

Moreover, Ramsay and Silverman discussed the case when the current value $y_i(t)$ of the response is affected by the predictors' values $x_i(s)$ over the entire time interval. The values for $(a,b)$ are $(0,T)$ in this case.

We have the following model:

$$y_i(t) = \alpha(t) + \int_0^T X_i(s)\beta(s,t)ds + \varepsilon_i(t), t \in [0,T].$$

This expression allows $y_i(t)$ to be influenced even by future values of the predictors. This type of model holds for cyclical or periodic data since the dynamics have a consistent and repeated behavior.

However, responses with no periodic behavior are more common. Malfait and Ramsay [2] developed the model in which the dependent variable is influenced only by the current and past values of the independent variable. They considered the following model:

$$y_i(t) = \alpha(t) + \int_{t-\delta}^{t} x_i(t)\beta(s,t)ds + \varepsilon_i(t), t\epsilon[0,T].$$

The response $y_i(t)$ is only affected by the predictors $x_i(s)$ at $s \in [t-\delta, t]$. They called historical functional linear regression to this feed-forward type of model. In this type of modeling, it is also of interest to estimate the time lag $\delta$ from the data. The model developed in this research is the general case of the historical functional linear regression model. This model is explained in chapter 2.

## 1.2 EMISSION MODELS

As mentioned before, emission models are an important tool for the successful planning and application of projects for the control and reduction of the vehicles emission production. This is because measuring emissions at each traffic situation is not feasible. One way to measure how much emission a vehicle produces given a traffic situation and driving behavior is by using a chassis dynamometer. In this case,

prescribed driving cycles can be applied to different types of vehicles. This type of experiments not only get rids of the effects of non significant factors but also decrease the variation and uncertainty in the experiments. Chassis dynamometer measurements provide the opportunity to find out how instantaneous change in the engine affects the production of the emission in vehicles.

### 1.2.1 TYPE OF EMISSION MODELS

There exist several emissions models which provide an estimate of the emissions generated or produced by different type of vehicles [5]. In general, these emissions models can be classified in three main categories:

- Average model
- Map model
- Load-based model

Each type of model differs in the way that they explain the relation between the response (emission values) and independent variables (driving behavior variables).

**Average Model** is the least complex model. In this case, the average of the driving behavior variables, velocity and acceleration, are used to estimate the total emissions produced over a time interval. The total emissions are related to polynomial functions of the average values. The advantage of this model is that it provides an accurate estimate of the total emissions from a large area. However, any inference in smaller

scales such as a specific intersection or road is not feasible given that the instantaneous data of the driving behavior variables is averaged [5].

**Map Model**, in comparison with the average model, uses the instantaneous values of the variables for the estimation of the model. In this case, they classify the values of the instant velocities and acceleration in different numerical levels or categories and given this category an estimation of the rate of emission is given. This model relates emission rates with polynomial function of these velocity and acceleration levels. The advantage of this model is that it gives evidence of the effects that instantaneous values of acceleration and velocity have in the production of emissions. However, this model fails to account for driving cycles that were not included for the estimation of the model. This implies that the accurateness of the model depends on the driving cycle [5].

**Load Based Models** divide the dynamics of the emissions from their production, release and transportation into different sub-models [3]. The delay that happens in the dynamics of transportation is accounted by a sub-model [3]. These sub-models have their own function and parameters and some of them depend on the outcome of other sub-models. This implies a deeper understanding of the dynamics production and transportations of the emissions [5]. However, these models are complex and need the information of several variables making it inconvenient and difficult to estimate.

### 1.2.2 DISTORTION

It is necessary to take into consideration that the recorded trajectory of the emission is not the instantaneous response of the effects of the driving cycle. This is because the particles are affected by several factors such as the air flow, temperature, and interaction with other particles while being transported from the tailpipe to the analyzer [3]. There are some suggestions to solve this alteration while the emission model is being developed.

Some of these suggestions are:

- No change. It is assumed that the data does not suffer any distortion and thus the records represent the instantaneous response of the effects of the driving behavior variables. The data is not changed and used to estimate the model parameter as it is.
- Offset the data. The trajectory of the response is assumed to be offset $t$ seconds from the values of the driving behavior values. This offset time can be found by looking at the estimated sum square error (SSE) or correlation between the observation and the predictions [6]. That is, the data is offset a second each time until the minimum SSE or maximum correlation is found. This implies that the response values $y_i(t+s)$ are related with $x(t)$ values of the independent variables and the SSE (correlation) is found each time the offset time $s$ increments. The offset time $t$ is chosen by looking at the $s$ value with the smallest SSE or largest correlation. In this case, it is assumed that the emissions particles are affected in similar degrees and thus their travel are the

same. This is not the case since as it was mentioned earlier, there are many factors that can affect the particles and thus they might not travel uniformly.

- Model the distortion. Weilenmann suggested that the engine and the transport dynamics affect the travel of the particles and thus emissions that are recorded during the time of the experiment are not after-the-catalyst emission information [3]. This was fixed by modeling the pure time delay for the transport of the gas and a dynamic signal deformation phenomenon separately [3]. This adds complexity to development of emission model.

We show an alternative emission model that can be applied to this type of data. We first discuss the mechanics of the model at chapter 2 and then the results at chapter 4.

# CHAPTER 2

# MODEL

In this thesis, we develop a general case of the historical functional linear regression model discussed in previous chapter. This model identifies the relation between the instantaneous changes in the dynamics and its response even when these are not the result of the instantaneous effects but a distortion. This implies that the instantaneous reaction as a result of the stimulus is not the same as the recorded value. The response suffers a delay, and the measurement of the response at time t is the sum of portions of the instantaneous reactions happened at the interval $[t-\delta, t]$. The purpose of our model is to account for the continuity and smoothness of the data and find the continuous influence that the stimuli have over the response.

## 2.1 DESCRIPTION

Suppose we have the records $\{y_1(t), \dots, y_N(t)\}$ and with their own P stimuli given by $\{\xi_{11}(t), \dots, \xi_{1P}(t), \dots, \xi_{N1}(t), \dots, \xi_{NP}(t)\}$. This data was taken at time intervals given by $t \in [0, T_i]$ for $i = 1, \dots, N$. We assume that these records $y_i(t)$ can be explained by a linear combination of recent and current values of the stimuli $\{\xi_{i1}(t), \dots, \xi_{iP}(t)\}$ by

$$y_i(t) = \alpha + \sum_{p=1}^{P} \int_{t-\delta_p}^{t} \beta_p(w)\xi_{ip}(w)\, dw + \varepsilon_i(t).$$

Each stimulus $\xi_{iP}(t)$ influences $y_i(t)$ over a time lag of length $\delta_p$. We call the $\beta_p(w)$ coefficient of the convolution. The coefficient of the convolution tells us the effects that stimulus have over $y_i(t)$ as they change over time. We have that $\alpha$ is the fixed

intercept and $\varepsilon_i(t)$ is the random error. As in classical linear regression, we assume that $E(\varepsilon_i(t)) = 0$. However, given the nature of the data, the random error might have some auto-correlation and thus, we consider that $cov(\varepsilon_i(t), \varepsilon_i(s)) \neq 0$ for $\forall|t-s| \geq 0$ but $cov\left(\varepsilon_i(t), \varepsilon_{i'}(s)\right) = 0$ for $\forall|t-s| \geq 0$ when $i \neq i'$.

This model accounts for the smoothness and continuity of the data curves by looking at the components as function of time and the delay suffered by the response by integrating the convolution of length $\delta_p$ for each continuous stimulus. Our first task is to estimate the fixed intercept $\alpha$ and for each stimulus, the coefficient function $\beta_p(w)$ and the lag $\delta_p$ parameters by using the recordings $\{y_1(t), \dots, y_N(t)\}$ and stimuli $\{\xi_{11}(t), \dots, \xi_{1P}(t), \dots, \xi_{N1}(t), \dots, \xi_{NP}(t)\}$.

### 2.1.1 ESTIMATING $\alpha$ AND $\beta_p(w)$

We use the ordinary least square criterion which is the same method applied in classical linear regression and functional linear regression models for the estimation of $\alpha$ and $\{\beta_1(w), \dots, \beta_p(w)\}$. We are interested in representing the coefficients of each stimulus as function of time in which $\beta_p(w)$ is a functional object in an interval given by $[1, \delta_p]$ where $\delta_p$ is the time lag dependency and unknown. Assuming we know $\delta_p$ for each stimulus, we then represent the coefficients as function of time by using basis function expansion:

$$\beta_p(w) = \sum_{k=1}^{K_p} c_{pk}\,\phi_{pk}(w)\ , p = 1, \dots, P$$

where $P$ is the number of stimulus included in the model and $K_p$ the number of basis used to represent the convolution of a specific stimulus $p$. We have that $\phi_{pk}$ is the

basis functional building block which in this case is in b-spline and $c_{pk}$ is its respective coefficient which is unknown. By letting $\boldsymbol{c'}$ to be a vector of length $K_p$ with the coefficients $\{c_{p1}c_{p2} \dots c_{pK_p}\}$ and $\boldsymbol{\phi_p}(\boldsymbol{w})$ to be the functional vector of length $K_p$ with elements $\{\phi_{p1}(w)\ \phi_{p2}(w) \dots \phi_{pk_p}(w)\}$, we express the function of convolution in matrix notation as

$$\beta_p(w) = \boldsymbol{c'_p}\boldsymbol{\phi_p}(\boldsymbol{w}).$$

We can see now that the model is redefined as

$$y_i(t) = \alpha + \sum_{p=1}^{P} \int_{t-\delta_p}^{t} \boldsymbol{c'_p}\boldsymbol{\phi_p}(\boldsymbol{w})\xi_{ip}(w)dw + \varepsilon_i(t).$$

Looking at the last expression, we can see that that it is of interest in finding the values of the coefficients of the b-splines $c_{pk}$ in addition to the fixed intercept $\alpha$. Next, we know that our data is composed of records of the responses $\{y_1(t), \dots, y_N(t)\}$ and the stimulus $\{\xi_{11}(t), \dots, \xi_{1P}(t), \dots, \xi_{N1}(t), \dots, \xi_{NP}(t)\}$. However, the observations $y_i(t)$ and $\{\xi_{i1}(t), \dots, \xi_{iP}(t)\}$ do not have to have the same curve pattern and recording time length as the observations $y_j(t)$ and $\{\xi_{j1}(t), \dots, \xi_{jP}(t)\}$, where $i \neq j$ and $i, j \in \{1, \dots, N\}$. We use matrix methods to integrate the information of all the recordings and find the solution of the parameters.

First, we do this for a specific record $i$ and express the model as the following matrix notation:

$$\boldsymbol{Y}_i = \boldsymbol{Z}_i\boldsymbol{\theta} + \boldsymbol{\varepsilon}_i.$$

The elements of the matrix notation are defined as

$$\boldsymbol{Y_i} = \begin{bmatrix} y_{it_1} \\ y_{i(t_1+1)} \\ \vdots \\ y_{i(t_j)} \\ \vdots \\ y_{i(t_{T_i}-1)} \\ y_{i(t_{T_i})} \end{bmatrix}, \quad \boldsymbol{\theta} = \begin{bmatrix} \alpha \\ \boldsymbol{c_1} \\ \vdots \\ \boldsymbol{c_p} \\ \vdots \\ \boldsymbol{c_{P-1}} \\ \boldsymbol{c_p} \end{bmatrix}, \boldsymbol{\varepsilon_i} = \begin{bmatrix} \varepsilon_{it_1} \\ \varepsilon_{i(t_1+1)} \\ \vdots \\ \varepsilon_{i(t_j)} \\ \vdots \\ \varepsilon_{i(t_{T_i}-1)} \\ \varepsilon_{i(t_{T_i})} \end{bmatrix}, \text{ and}$$

$$\boldsymbol{Z_i} = \begin{bmatrix} 1 & \int_{t_1-\delta_1}^{t_1} \xi_{i1}(w)\boldsymbol{\phi'_1}(\boldsymbol{w})dw & \int_{t_1-\delta_2}^{t_1} \xi_{i2}(w)\boldsymbol{\phi'_2}(\boldsymbol{w})dw \ldots & \int_{t_1-\delta_p}^{t_1} \xi_{ip}(w)\boldsymbol{\phi'_p}(\boldsymbol{w})dw \ldots \\ 1 & \int_{t_1+1-\delta_1}^{t_1+1} \xi_{i1}(w)\boldsymbol{\phi'_1}(\boldsymbol{w})dw & \int_{t_1+1-\delta_2}^{t_1+1} \xi_{i2}(w)\boldsymbol{\phi'_2}(\boldsymbol{w})dw \ldots & \int_{t_1+1-\delta_p}^{t_1+1} \xi_{ip}(w)\boldsymbol{\phi'_p}(\boldsymbol{w})dw \ldots \\ \vdots & \vdots & \vdots & \vdots \\ 1 & \int_{t_j-\delta_1}^{t_j} \xi_{i1}(w)\boldsymbol{\phi'_1}(\boldsymbol{w})dw & \int_{t_j-\delta_2}^{t_j} \xi_{i2}(w)\boldsymbol{\phi'_1}(\boldsymbol{w})dw \ldots & \int_{t_j-\delta_p}^{t_j} \xi_{ip}(w)\boldsymbol{\phi'_p}(\boldsymbol{w})dw \ldots \\ \vdots & \vdots & \vdots & \vdots \\ 1 & \int_{t_{T_i}-1-\delta_1}^{t_{T_i}-1} \xi_{i1}(w)\boldsymbol{\phi'_1}(\boldsymbol{w})dw & \int_{t_{T_i}-1-\delta_2}^{t_{T_i}-1} \xi_{i2}(w)\boldsymbol{\phi'_1}(\boldsymbol{w})dw \ldots & \int_{t_{T_i}-1-\delta_p}^{t_{T_i}-1} \xi_{ip}(w)\boldsymbol{\phi'_p}(\boldsymbol{w})dw \ldots \\ 1 & \int_{t_{T_i}-\delta_1}^{t_{T_i}} \xi_{i1}(w)\boldsymbol{\phi'_1}(\boldsymbol{w})\xi_{i1}(w)dw & \int_{t_{T_i}-\delta_2}^{t_{T_i}} \xi_{i2}(w)\boldsymbol{\phi'_1}(\boldsymbol{w})dw \ldots & \int_{t_{T_i}-\delta_p}^{t_{T_i}} \xi_{ip}(w)\boldsymbol{\phi'_p}(\boldsymbol{w})dw \ldots \end{bmatrix} =$$

Putting everything together, the matrix notation of the model using records $\{1, \ldots, N\}$ is

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_N \end{bmatrix} = \begin{bmatrix} \boldsymbol{Z}_1\boldsymbol{\theta} + \boldsymbol{\varepsilon}_1{}' \\ \vdots \\ \boldsymbol{Z}_N\boldsymbol{\theta} + \boldsymbol{\varepsilon}_N \end{bmatrix} = \boldsymbol{Z\theta} + \boldsymbol{\varepsilon}.$$

The next step is to find the estimation of $\boldsymbol{\theta} = [\alpha, \boldsymbol{c}_1, \ldots, \boldsymbol{c}_k]$. This is done by first fixing $\boldsymbol{\delta} = \{\delta_1, \ldots, \delta_p\}$ and then using the least square methods or minimizing

$$SMSSE(Y|\theta) = (\boldsymbol{Y} - \boldsymbol{Z\theta})'(\boldsymbol{Y} - \boldsymbol{Z\theta}) = ||\boldsymbol{Y} - \boldsymbol{Z\theta}||^2.$$

The estimate of $\boldsymbol{\theta}$ is

$$\widehat{\boldsymbol{\theta}} = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{Y}.$$

This implies that the estimation of $\{\hat{\beta}_1(w), \dots, \hat{\beta}_p(w)\}$ is found by $\hat{\beta}_p(w) = \hat{\boldsymbol{c}}'_{\boldsymbol{p}}\boldsymbol{\phi}_{\boldsymbol{p}}(\boldsymbol{w})$ for $p = 1, \dots, P$.

### 2.1.2 ROUGNESS PENALTY $\boldsymbol{\lambda}$

In addition, we can introduce a roughness penalty to the least square expression so that it gives us smoother estimated curves and handles the problem of over fitting the data. This is included in the model expression as

$$PENSSE_\lambda = \sum_{i=1}^{N}\sum_{j=1}^{T_i}[y_{it_j} - \alpha - \sum_{p=1}^{P}\int_{t_j-\delta_p}^{t_j}\beta_p(w)\xi_{ip}(w)\,dw]^2$$

$$+\sum_{p=1}^{P}\lambda_p\int_{t_j-\delta_p}^{t_j}[D^m\beta_p(w)]^2dw.$$

We have that

- $\lambda_p$ is the smoothing parameter for the p covariate and should be greater than zero. The curve estimation is less penalized, varies more, and the bias is smaller when $\lambda_p \to 0$ and vice versa when $\lambda_p \to \infty$.
- The roughness penalty is expressed by $\int_{t_j-\delta_p}^{t_j}[D^m\beta_p(w)]^2dw$.

The roughness penalty is the $m$ derivative of function because they contain the information of how this is changing over time. This new expression accounts for the variability in the curve that the SMSSE is not able to fit.

In general, $\boldsymbol{\lambda} = \{\lambda_1, \dots, \lambda_p, \dots, \lambda_P\}$ need to be estimated and this will be discussed in the next section. Meanwhile, we show how to estimate $\widehat{\boldsymbol{\theta}} = [\hat{\alpha}, \hat{\boldsymbol{c}}_{\boldsymbol{1}}\hat{\boldsymbol{c}}_{\boldsymbol{2}}, \dots, \hat{\boldsymbol{c}}_{\boldsymbol{k}}]$ by fixing the values of $\boldsymbol{\lambda}$. We represent first $\beta_p(w)$ and $D^m\beta_p(w)$ as a basis expansion and redefined the sum square expression as

$$PENSSE_\lambda = \sum_{i=1}^{N}\sum_{j=1}^{T_i}[y_{it_j} - \alpha + \sum_{p=1}^{P}\int_{t_j-\delta_p}^{t_j} \boldsymbol{c_p}'\boldsymbol{\phi_p}(\boldsymbol{w})]^2\xi_{ip}(w)\,dw]^2$$

$$\sum_{p=1}^{P}\lambda_p\int_{t_j-\delta_p}^{t_j}[D^m\boldsymbol{c_p}'\boldsymbol{\phi_p}(\boldsymbol{w})]^2dw.$$

We can write the roughness penalty expression as a matrix notation by

$$\sum_{p=1}^{P}\lambda_p\int_{t_j-\delta_p}^{t_j}[D^m\boldsymbol{c_p}'\boldsymbol{\phi_p}(\boldsymbol{w})]^2dw = \boldsymbol{\theta}'\boldsymbol{R}(\boldsymbol{\lambda})\boldsymbol{\theta}$$

where $\boldsymbol{R}(\boldsymbol{\lambda}) =$

$$\begin{bmatrix} 0 & 0 & \cdots & 0 & 0 \\ 0 & \lambda_1\int[D^m\boldsymbol{\phi_1}(\boldsymbol{w})]^2dw & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \\ & & & \lambda_{P-1}\int[D^m\boldsymbol{\phi_{P-1}}(\boldsymbol{w})]^2dw & 0 \\ 0 & 0 & \cdots & & \\ 0 & 0 & & 0 & \lambda_P\int[D^m\boldsymbol{\phi_P}(\boldsymbol{w})]^2dw \end{bmatrix}$$

The matrix notation for the penalized sum square error is

$$PENSSE(Y|\theta) = (\boldsymbol{Y} - \boldsymbol{Z\theta})'(\boldsymbol{Y} - \boldsymbol{Z\theta}) + \boldsymbol{\theta' R(\lambda)\theta}.$$

We estimate $\boldsymbol{\theta}$ by minimizing the expression $PENSSE(Y|\theta)$ given by

$$\widehat{\boldsymbol{\theta}} = (\boldsymbol{Z'Z} + \boldsymbol{R(\lambda)})^{-1}\boldsymbol{Z'Y}.$$

In this case, we fix $\boldsymbol{\delta} = \{\delta_1, \dots, \delta_p\}$ and $\boldsymbol{\lambda} = \{\lambda_1, \dots, \lambda_p, \dots, \lambda_P\}$ and then estimate $\widehat{\boldsymbol{\theta}} = [\hat{\alpha}, \hat{\boldsymbol{c}}_{\boldsymbol{1}}\hat{\boldsymbol{c}}_{\boldsymbol{2}}, \dots, \hat{\boldsymbol{c}}_{\boldsymbol{k}}]$ which implies the estimation of $\{\hat{\beta}_1(w), \dots, \hat{\beta}_p(w)\}$ by

$$\hat{\beta}_p(w) = \hat{\boldsymbol{c}}'_{\boldsymbol{p}}\boldsymbol{\phi_p}(\boldsymbol{w})$$

for $p = 1, \dots, P$.

### 2.1.3 ESTIMATING $\boldsymbol{\delta}$ AND $\boldsymbol{\lambda}$

We discussed in the last section how to find $\widehat{\boldsymbol{\theta}}$ given some fixed values of $\boldsymbol{\delta} = \{\delta_1, \dots, \delta_p, \dots, \delta_P\}$ and $\boldsymbol{\lambda} = \{\lambda_1, \dots, \lambda_p, \dots, \lambda_P\}$. However, these values also need to be estimated and thus a direct method does not exist. Instead, they can be estimated by using the method of cross-validation. The basic idea of cross-validation is to divide the data into two subsets $\{A, B\}$. Subset $A$ is called the $training\ sample$ and used to fit the model and subset B which is called the $validation\ sample$ is used to validate the model by finding the difference between subset B and its estimation $\hat{B}$ [4]. The next procedure is an explanation of how this is done to find $\boldsymbol{\delta}$ and $\boldsymbol{\lambda}$ parameters.

Let assume that we have the observations of the recordings $\{y_1(t), \dots, y_i(t), \dots, y_N(t)\}$ and the stimulus $\{\xi_{11}(t), \dots, \xi_{1P}(t)\}, \dots \{\xi_{N1}(t), \dots, \xi_{NP}(t)\}$. First, we find the values of $\boldsymbol{\delta}$ by following the next steps:

1. Choose a starting value for $\boldsymbol{\delta}$.
2. Fix the values of $\boldsymbol{\delta}$. We have then that $\boldsymbol{\delta}_{(d)} = \{\delta_1, \dots, \delta_p, \dots, \delta_P\}$ where $\delta_p \geq 0, \forall p$ and $d$ represents the state or specific values of the $\delta_p's$. If we have $d' \neq d$ then this implies that $\boldsymbol{\delta}_{(d)} \neq \boldsymbol{\delta}_{(d')}$.
3. Remove recording $y_i(t)$ and its respective $\{\xi_{i1}(t), \dots, \xi_{iP}(t)\}$ from the data. The new data is defined as $Y^{(-i)} = \{y_1(t), \dots, y_{i-1}(t), \dots, y_N(t)\}$ for the recordings and $\xi^{(-i)} = \{\xi_{11}(t), \dots, \xi_{1P}(t), \dots, \xi_{N1}(t), \dots, \xi_{NP}(t)\}$ for the stimulus. We have that $(-i)$ represents the missing observation in the new data.
4. Estimate $\widehat{\boldsymbol{\theta}}^{(-i)d} = (\boldsymbol{Z}'^{(-i)d}\boldsymbol{Z}^{(-i)d})^{-1}\boldsymbol{Z}'^{(-i)d}\boldsymbol{Y}^{(-i)}$.
5. Find the estimates of $\boldsymbol{Y}^i$ by $\widehat{\boldsymbol{Y}}^{(i)d} = \widehat{\boldsymbol{\theta}}^{(-i)}\boldsymbol{Z}^{(i)}$.
6. Find the Sum Square Error by $SSE^{(-i)d} = (\boldsymbol{Y}^i - \widehat{\boldsymbol{Y}}^{(i)d})'(\boldsymbol{Y}^i - \widehat{\boldsymbol{Y}}^{(i)d})$.
7. Repeat steps 3-5 for i=1,..., N.
8. Find the Total Sum Square Error by $TSSE^d = \sum_{i=1}^{N} SSE^{(-i)d}$.
9. Repeat 2-7 for $d$ by changing the values of the vector $\boldsymbol{\delta}$.
10. Find the minimum of the $TSSE$ by $MTSSE = \min(TSSE^d)$.

Once we know the minimum, then we will know the best combination of $\delta_p's$. We set $\widehat{\boldsymbol{\delta}}_{\boldsymbol{BEST}} = \{\hat{\delta}_1, \dots, \hat{\delta}_p, \dots, \hat{\delta}_P\}$ which are the values that with the smaller TSSE. However, we want also to find the roughness penalty $\boldsymbol{\lambda} = \{\lambda_1, \dots, \lambda_p, \dots, \lambda_P\}$ for the coefficients of the stimulus. We follow similar steps as before and the difference is

that instead of changing the length or value of the $\delta_p{}'s$ we change the values of the $\lambda_p{}'s$. The penalized sum square error is used in this case instead. Next, we estimate $\boldsymbol{\lambda}$ by the following steps:

1. Get $\widehat{\boldsymbol{\delta}}_{\boldsymbol{BEST}}$ and thus we have $\widehat{\boldsymbol{\delta}}_{\boldsymbol{BEST}} = \{\hat{\delta}_1, \dots, \hat{\delta}_p, \dots, \hat{\delta}_P\}$ where $\hat{\delta}_p > 0 \forall p$.
2. Choose initial values for $\boldsymbol{\lambda}$.
3. Fix the values of $\boldsymbol{\lambda}$ and thus we have $\boldsymbol{\lambda}_{(d)} = \{\lambda_1, \dots, \lambda_p, \dots, \lambda_P\}$ where $\lambda_p \geq 0, \forall p$ and $(d)$ represents the state or specific values of the $\boldsymbol{\lambda}_p{}'s$ . If we have $d' \neq d$ then this implies that $\boldsymbol{\lambda}_{(d)} \neq \boldsymbol{\lambda}_{(d')}$.
4. Take out recording $y_i(t)$ and its respective $\{\xi_{i1}(t), \dots, \xi_{iP}(t)\}$ from the data and thus the new data is defined as $Y^{(-i)} = \{y_1(t), \dots, y_{i-1}(t), \dots, y_N(t)\}$ for the recordings and $\xi^{(-i)} = \{\xi_{11}(t), \dots, \xi_{1P}(t)\}, \dots \{\xi_{N1}(t), \dots, \xi_{NP}(t)\}$ for the stimulus. We have that $(-i)$ represents the missing observation in the new data.
5. Estimate by $\widehat{\boldsymbol{\theta}}^{(-i)d} = (\boldsymbol{Z}'^{(-i)d}\boldsymbol{Z}^{(-i)d} + \boldsymbol{R}(\boldsymbol{\lambda}))^{-1}\boldsymbol{Z}'^{(-i)d}\boldsymbol{Y}^{(-i)}$.
6. Find the estimates of $\boldsymbol{Y}^i$ by $\widehat{\boldsymbol{Y}}^{(i)d} = \widehat{\boldsymbol{\theta}}^{(-i)}\boldsymbol{Z}^{(i)}$.
7. Find the Sum Square Error by $SSE^{(-i)d} = (\boldsymbol{Y}^i - \widehat{\boldsymbol{Y}}^{(i)d})'(\boldsymbol{Y}^i - \widehat{\boldsymbol{Y}}^{(i)d})$.
8. Repeat steps 3-5 for i=1,..., N.
9. Find Total Sum Square Error by $TSSE^d = \sum_{i=1}^{N} SSE^{(-i)d}$.
10. Repeat 2-7 for $d$ by increasing the value of the $\lambda_p{}'s$.
11. Find the minimum of the $TSSE$ by $MTSSE = \min(TSSE^d)$.
12. Once we know the minimum, then we will know the best combination of $\delta_p{}'s$.

We set $\hat{\boldsymbol{\lambda}}_{\boldsymbol{BEST}} = \{\hat{\lambda}_1, \dots, \hat{\lambda}_p, \dots, \hat{\lambda}_P\}$ which are the values that with the smaller TSSE. Iterating these two cross-validation procedures will result in an algorithm that attempts

to minimize TSSE over both $\boldsymbol{\delta}$ and $\boldsymbol{\lambda}$ jointly. However, this is likely the result in a local minimum and will be computational expensive. We therefore run each procedure once, providing one-step approximation to the minimum TSSE estimate. At the end, we find also $\beta_p(t), p = 1, \ldots, P$ since $\widehat{\boldsymbol{\theta}} = [\hat{\alpha}, \hat{\boldsymbol{c}}_{\boldsymbol{1}}\hat{\boldsymbol{c}}_{\boldsymbol{2}}, \ldots, \hat{\boldsymbol{c}}_{\boldsymbol{k}}]$ by

$$\hat{\beta}_p(w) = \hat{\boldsymbol{c}}'_{\boldsymbol{p}}\boldsymbol{\phi}_{\boldsymbol{p}}(\boldsymbol{w})$$

for $p = 1, \ldots, P$.

## 2.2 MODEL DIAGNOSTICS

Once we estimate the parameters of our model the next step is to look at how well our model predicts the observations. In this section, we discuss the estimation for different components that at the end gives us a feedback about how our model is doing and if we can rely in its output.

### 2.2.1 ESTIMATING RESIDUALS

The analysis of residuals is an important step for the diagnostic of our model. This is because

- By analyzing the residuals, we can find out if the assumptions about our model are correct.
- The residuals give us feedback about how close we are to the observations.
- They allow seeing how each element of the data influences the estimation of the parameters.

In our case, we have that out model is the following:

$$y_i(t) = \alpha + \sum_{p=1}^{P} \int_{t-\delta_p}^{t} \beta_p(w)\xi_{ip}(w)\, dw + \varepsilon_i(t)$$

assuming the $E(\varepsilon_i(t)) = 0$ and $cov(\varepsilon_i(t), \varepsilon_i(s)) \neq 0$ for $\forall |t-s| \geq 0$ but $cov\left(\varepsilon_i(t), \varepsilon_{i'}(s)\right) = 0$ for $\forall |t-s| \geq 0$ when $i \neq i'$.

In matrix notation, we have that the model can be written as

$$\boldsymbol{Y} = \boldsymbol{Z\theta} + \boldsymbol{\varepsilon}\ E(\boldsymbol{\varepsilon}) = \mathbf{0},\ cov(\boldsymbol{\varepsilon}) = \boldsymbol{\Sigma}$$

where $\boldsymbol{Z}$ is the design matrix and its number of columns depends on the length of $\boldsymbol{\theta}$ which contains the unknown coefficients of the basis and the fixed intercept. The $\boldsymbol{\varepsilon}$ is the unobservable random variable and $\boldsymbol{\Sigma}$ the covariance matrix.

To estimate the residuals, we use the estimate of $\boldsymbol{\theta}$ discussed in previous sections. We have $\widehat{\boldsymbol{\theta}} = (\boldsymbol{Z'Z} - \boldsymbol{R}(\widehat{\boldsymbol{\lambda}}))^{-1}\boldsymbol{Z'Y}$ . Using the estimate $\widehat{\boldsymbol{\theta}}$, we find the fitted values of $\boldsymbol{Y}$ by $\widehat{\boldsymbol{Y}} = \boldsymbol{Z}\widehat{\boldsymbol{\theta}}$. The estimates of the residuals are just the difference between the observation and fitted values. That is, we have that the residuals are found by

$$\boldsymbol{e} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}}.$$

These residuals are then used for the estimation of the covariance of the model as it is explained in the following section.

### 2.2.2 ESTIMATING THE COVARIANCE

We assume that the covariance of the model is estimated as $\Sigma_{ts} = cov(y_i(t), y_i(s)) = cov\big(\varepsilon_i(t), \varepsilon_i(s)\big) = R(s,t)$. Under the assumption that $\{y_1(t), \dots, y_i(t), \dots, y_N(t)\}$ has stationary covariance, we can define $R(s,t) = R(|s-t|)$ for $s, t \in T_i$.

This means that we can write the $\Sigma_i$ as the following:

$$\Sigma_i = \begin{pmatrix} R(0) & R(1)\cdots & R(n-1) \\ R(1) & R(0)\cdots & R(n-2) \\ \vdots & \vdots\cdots & \vdots \\ R(n-1) & R(n-2)\dots & R(0) \end{pmatrix}$$

Given that the random errors are not observed, we use the residuals to estimate the covariance. For each $y_i(t)$, we estimate $\hat{\Sigma}_i$ by using the sample autocovariance of the residuals. This is done by

$$(\hat{\Sigma}_{\mathrm{i}})_l = \hat{R}(|s-t|) = \frac{1}{T_i}\sum_{s}^{T_i - l} e_{s+l} e_s.$$

We locate then these autocovariance values in the covariance matrix as

$$\hat{\Sigma}_i = \begin{pmatrix} \hat{R}(0) & \hat{R}(1)\cdots & \hat{R}(n-1) \\ \hat{R}(1) & \hat{R}(0)\cdots & \hat{R}(n-2) \\ \vdots & \vdots\cdots & \vdots \\ \hat{R}(n-1) & \hat{R}(n-2)\dots & \hat{R}(0) \end{pmatrix}$$

Then, we find the general estimation of $\boldsymbol{\Sigma}$ by the mean of the estimated covariance matrices

$$(\widehat{\boldsymbol{\Sigma}}_1 + \cdots + \widehat{\boldsymbol{\Sigma}}_i + \cdots + \widehat{\boldsymbol{\Sigma}}_N)\frac{1}{N} = \widehat{\boldsymbol{\Sigma}}.$$

This general estimation will be useful in the chapters of Block Bootstrapping and Influence and Outlier to make inference about the variability and effects of the data in the estimation of the model.

### 2.2.3 CONFIDENCE INTERVAL

Once we have an estimate of the coefficient function $\hat{\beta}_p(w)$ for each of the stimulus, we are interested in knowing about the accurateness and variability of our estimates. One simple way to observe the variability of the estimates of the model parameters is by using the delta method.

First, we show how the variance of $\boldsymbol{\theta}$

$$cov(\widehat{\boldsymbol{\theta}}) = cov((\boldsymbol{Z}'\boldsymbol{Z} - \boldsymbol{R}(\boldsymbol{\lambda}))^{-1}\boldsymbol{Z}'\boldsymbol{Y})$$

$$= (\boldsymbol{Z}'\boldsymbol{Z} - \boldsymbol{R}(\boldsymbol{\lambda}))^{-1}\boldsymbol{Z}'cov(\boldsymbol{Y})\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z} - \boldsymbol{R}(\boldsymbol{\lambda}))^{-1}$$

$$= (\boldsymbol{Z}'\boldsymbol{Z} - \boldsymbol{R}(\boldsymbol{\lambda}))^{-1}\boldsymbol{Z}'\widehat{\boldsymbol{\Sigma}}\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z} - \boldsymbol{R}(\boldsymbol{\lambda}))^{-1}.$$

The confidence interval of $\boldsymbol{\theta}$ then can be estimated by using the delta method as

$$\widehat{\boldsymbol{\theta}} \pm 2 * \sqrt{diag(cov(\widehat{\boldsymbol{\theta}}))}.$$

CHAPTER 3

# CASE STUDY

We found the solution for an alternative emission model using chassis dynamometer measurements gathered as a result of the application of several driving cycles to a group of medium heavy-duty trucks (MHDT). The objective of this chapter is to describe the data and inquire about the relation between particulate matter and the driving behavior variables. This provides evidence that the model developed in this thesis is good fit for the nature and behavior of the chassis dynamometer measurements. This is because the particulate matter (response) and driving behavior variables (predictors) had been recorded second-by-second over a time interval and the response was affected by various factors while being transported from the tailpipe to the emission analyzer equipment. This implies that these recordings do not represent the instantaneous effects of the independent variables in this case the effects of velocity and acceleration.

## 3.1 PARTICULATE MATTER

Particulate matter is a pollutant that can cause serious cardiovascular and respiratory illnesses and in some cases even the death of the individual [12, 13]. It has been shown it leads to approximately 100,000 early deaths per year in the United States [14]. Recent research has shown that certain characteristics of the particles such as size and chemical properties are linked to specific effects in the human health. In particular, fine particles are particulate matter smaller than 2.5 $\mu m$. This characteristic gives them the ability to stay suspended in the air and to be easily inhaled and get

attached to the human body affecting the functioning of respiratory or cardiovascular components. In addition, there is a significant connection between the exposure to these fine particles and daily deaths in six eastern U. S. cities [8]. These specific types of particles are mainly produced by mobile sources. Although the production of particulate matter by vehicle has decreased significantly in recent decades, it has been shown that they still are an important contributor [9]. In urban environments, almost 90% of traffic-generated particulate matter is from diesel exhaust [10]. This case study was done with the purpose to relate the particulate matter with driving behavior variables, velocity and acceleration. We wanted to come with an alternate emission model that gives the trajectory estimates of the particulate matter and helps to come up with a course of actions to regulate and control this substance in cities where its levels are significantly high and its main contributors are diesel trucks.

## 3.2 E-55/59 PROGRAM

We used data from the E-55/59 program to serve as an example for the application of this model. The purpose of this program was to quantify heavy and medium heavy-duty trucks emissions production in the South Coast Air Basin of California. Data was gathered from 76 trucks with similar characteristics as the trucks used in the roads of that area at the moment of the experiment. We are interested in the analysis and modeling of data gathered from the medium heavy-duty trucks.

### 3.2.1 CHASSIS DYNAMOMETER MEASUREMENTS

The data set of medium heavy-duty trucks consists in chassis dynamometer measurements of eleven trucks. Chassis dynamometers allow the consistent application of specific driving cycles to more than one truck and the almost continuous recording of variables linked with the instant physical change of the engine and emission production of the vehicles over a period of time. These measurements are frequently used for the development of emission models since they permit us to consider the effects that driving behavior variables have while maintaining an almost steady or unchanging environment.

However, the recording of the emissions does not represent the effect of the instantaneous change of velocity. This is because the tailpipe is not connected directly to the emission analyzer. Instead, the emissions are transported through a exhaust system to the emission analyzer. These particles can experience a delay caused by the interaction with other factors and particles. Weilenmann suggested that this delay is the consequence of many factors affecting the particles at the transportation dynamics. This problem was solved by modeling the pure time delay for the transport of the gas and a dynamic signal deformation phenomenon separately as it was discussed in chapter 1 [3, 7]. This model is complex and needs the information of several variables.

In contrast, our model is an alternative for the modeling of particulate matter that intends to take into consideration the delay experienced by some of the particles. We specifically observe the dependence between past and current values of the particulate

matter with the velocity and acceleration. In addition, we discuss the estimates of the model in chapter 4.

### 3.2.2 DRIVING CYCLES

Driving cycles consist on a velocity patterns and weight load applied to the vehicle in the chassis dynamometer. Figure 3.1 shows the four velocity patterns applied to the medium heavy-duty trucks for the observation and study of emission production. The purpose of this driving cycles is to represent the weight carried and driving behavior follow by these trucks in the roads of California. We can see that the velocity patterns differ in several ways such as the length of time, the wiggleness of the curve, and the limit of velocity.

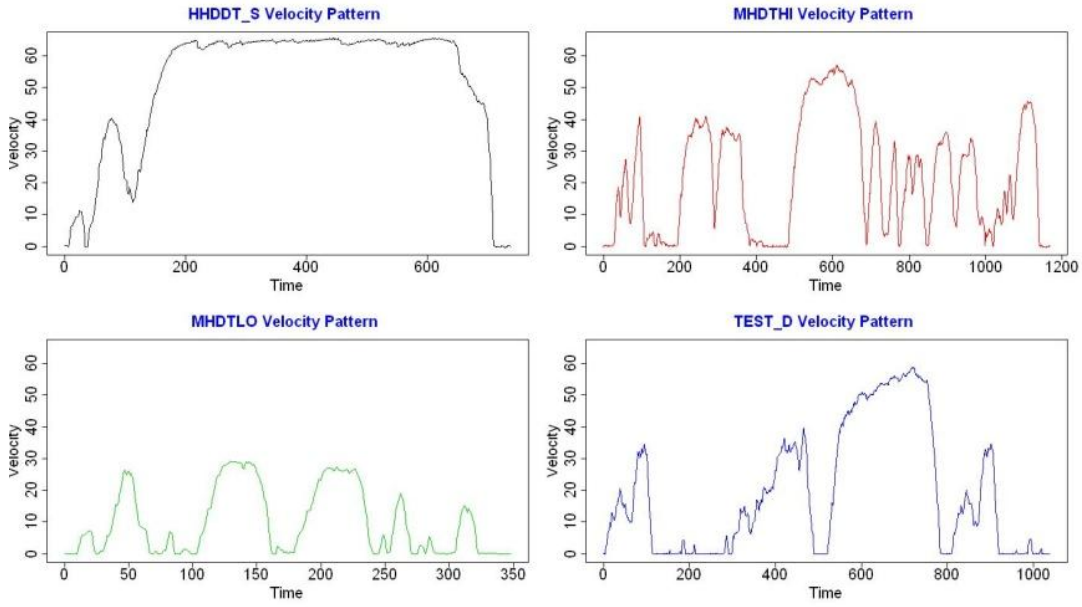



**Figure 3.1.** Velocity patterns applied to the medium heavy-duty trucks

These cycles can represent different traffic scenarios. For example, in the plot of HHDDT_S velocity pattern, the following can be observed:

- The velocity keeps incrementing until a little after 200 seconds
- The velocity remains constant at 60m/h.

This type of driving behavior is very common in highways were the limit of velocity is around 60 miles/hour and no stops are found.

In contrast, for the plot of MHDTLO velocity pattern, it can be seen that

- The trajectory fluctuates between 0 m/h to 30 m/h
- It has multiple stops
- The time between stops varies and can last various seconds

This velocity patterns represents the road or street traffic in which there are many stop signs or traffic lights and the maximum limit of velocity is 30 m/h.

Besides observing velocity patterns, it is also of interest to study the relation between acceleration and particulate matter. Figure 3.2 shows the acceleration pattern for each as a result of the velocity trajectories shown previously. Here, we see that

- The trajectories differ in pattern
- Their fluctuations are steadier than in the velocity patterns since they are between (-5, 3) in all the four trajectories.

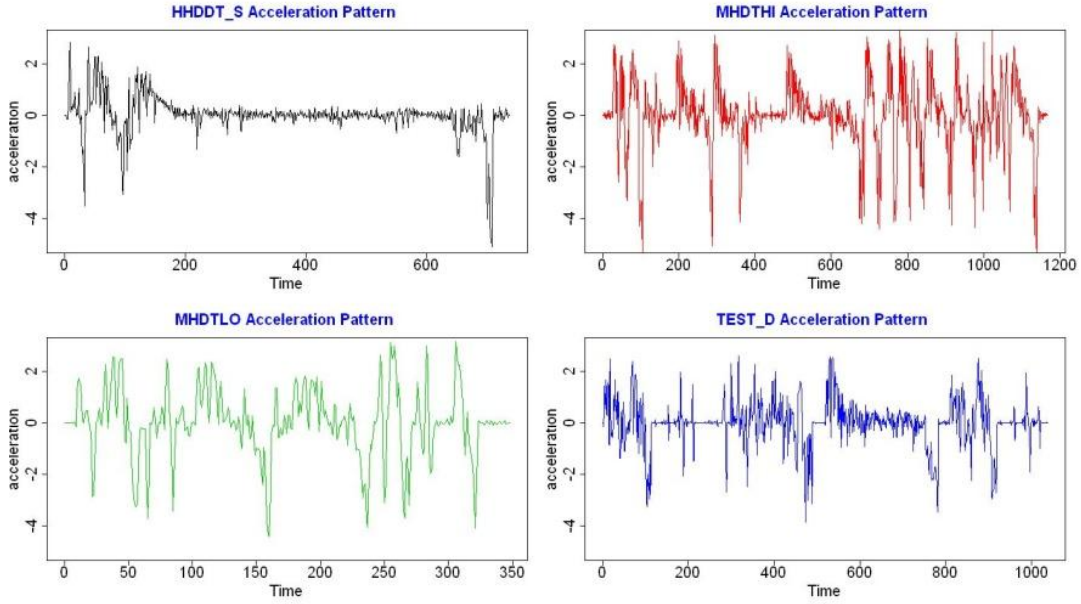



**Figure 3.2**. Acceleration trajectories as a result of a specific velocity pattern

In total, the data set consisted in 69 chassis dynamometer measurements gathered from different trucks. Figure 3.3 contains the curves of the particulate matter, the velocity and acceleration as a result of the application of MHDTHI velocity pattern to a group of trucks. We observe that most trajectories of the particulate matter follow a general pattern with exception of two curves. This implies that overall the trucks' emission production follows a particular trajectory that can be estimated by relating the PM with the velocity and acceleration. We also observe that the curves are smooth and continuous making it appropriate to apply methods from functional data analysis.

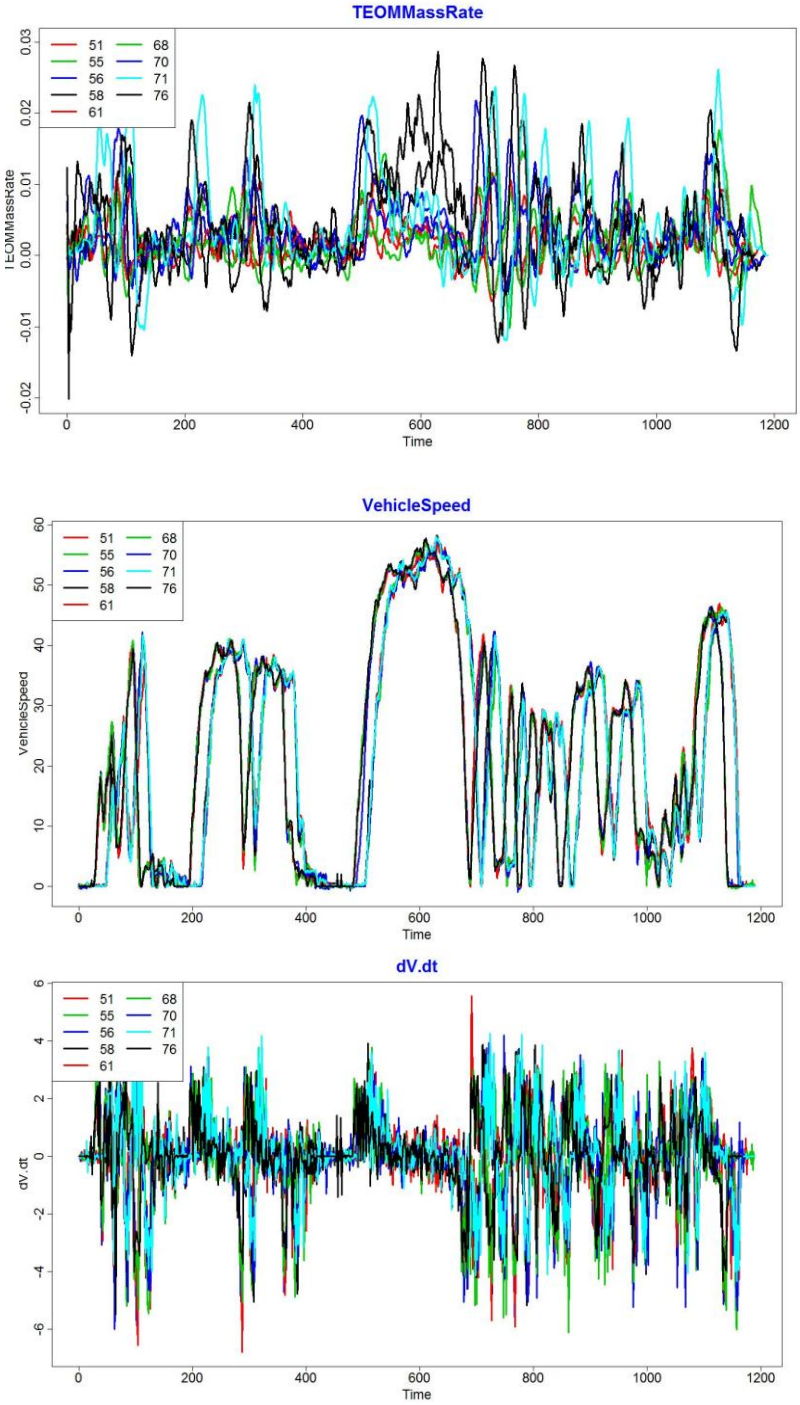



**Figure 3.3.** Comparison trajectories as a result of MHDTHI velocity pattern

## 3.3 DATA ANALYSIS

We present results of several analyses to infer about the relation between particulate matter with the velocity and acceleration. In fact, we show evidence that particulate matter does not only significantly relate to current values of the driving behavior variables but also to past values. This dependence is related to the distortion that the particles experience while they are transported from the tailpipe to the analyzer.

### 3.3.1 AVERAGE PARTICULATE MATTER AND AVERAGE DRIVING BEHAVIOR VARIABLES

Given that the total time interval varies among the samples, we look at the relation between the averages values of particulate matter and driving behavior variables. The relation among these variables is shown in Figure 3.4 and 3.5. In particular, we observe that the samples have similar average velocity if they belong to the same velocity pattern in Figure 3.4. For example, we can see that all average velocity values from the HHDDT_S velocity pattern are around 50 m/h. However, we cannot make any inference about the relation between the average particulate matter and the average velocity. This is because the average values of particulate matter do not show a significant correlation with the average values of the velocity. Furthermore, the values of the average acceleration do not follow the same behavior as the average velocity values as it is shown in Figure 3.5. That is, the values of the average acceleration do not cluster by driving cycle. In addition, we also conclude that there is not strong relation between average particulate matter and average acceleration values. These two figures demonstrate that by averaging the variables important information

is lost and as a result, the estimation of particulate matter using these values is not enough and possible.

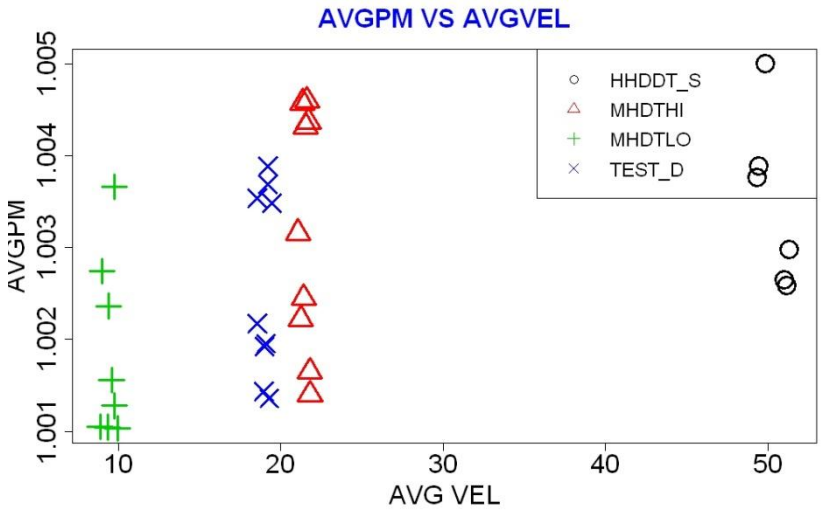



**Figure 3.4.** Relation between average pm and average velocity

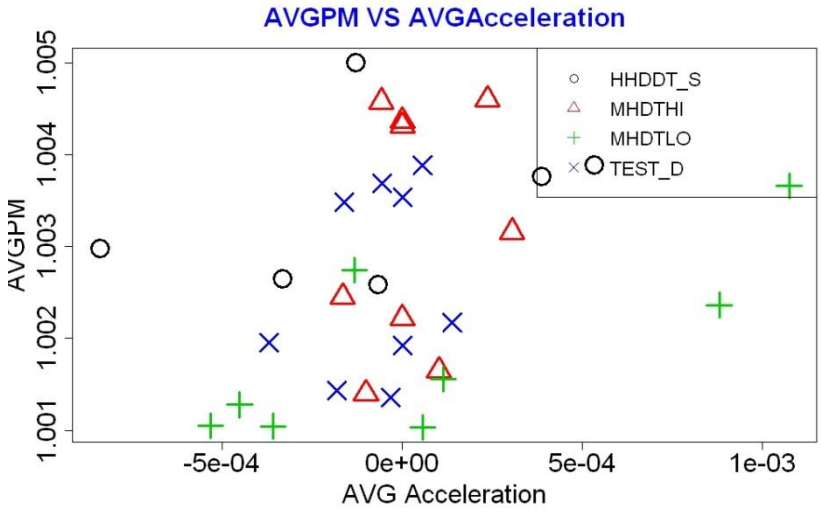
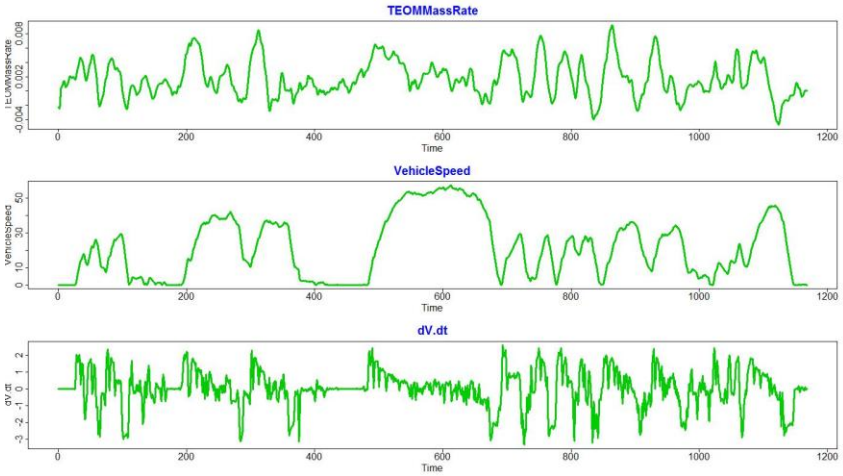


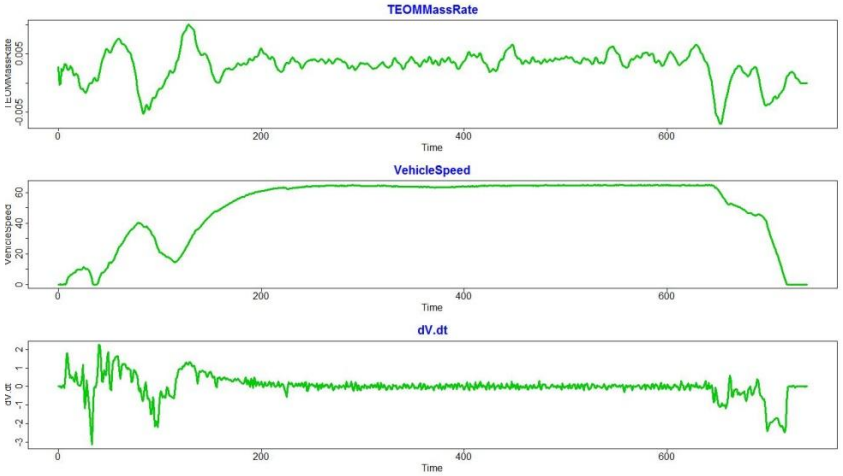

**Figure 3.5.** Relation between average pm and average acceleration

### 3.3.2 TRAJECTORIES

In this section, we show the trajectories as result of the application of two different velocity patterns in the same truck. We are interested to show that the instantaneous changes in velocity and acceleration are important for the estimation of the particulate matter. In Figure 3.6, we observe that the way the trajectories of the particulate matter behave depends in the instantaneous values of velocity and acceleration. That is, the wiggly (steady) behavior of the velocity trajectory in plot a (plot b) is reflected in the trajectory of the particulate matter. We can see that the particulate matter fluctuates more in plot (a) than in plot (b). Also, it seems that the peaks in the particulate matter trajectories are the result of major increments and decrements on the velocity and acceleration. However, we analyze further the relation between particular matter and the driving behavior variables by looking at the cross-correlation which is discussed in the next section.




a) TEST_D




b) HHDDT S

**Figure 3.6.** Trajectories given two different velocity patterns

### 3.3.3 CROSSCORRELATION

In this section, we study the relation of particulate matter with past values of the velocity and acceleration. One way to show the dependence over time between two different time series is by looking at the cross-correlation. We consider the following expressions

$$(\hat{R}(|s-t|) = \frac{1}{T_i}\sum_{s}^{T_i-l} VEL_{s+l}PM_s$$

$$(\hat{R}(|s-t|) = \frac{1}{T_i}\sum_{s}^{T_i-l} ACCEL_{s+l}PM_s$$

Figures 3.7 and 3.8 show that particulate matter does not only depend on present values of velocity and acceleration but also it is significantly related to past values of both. This is because the cross correlation curves do not go to zero at lag 1. That is, the correlation between velocity (acceleration) and particular matter is still significant even when the values of velocity (acceleration) go farther away from the current values of the particular matter. In addition, we observe that the cross correlation curves have similar patterns. In the case of cross correlation between particular matter and acceleration, we can see that the dependency is significant until lag 20 as it is shown in Figure 3.8. We can conclude that it is important to take in consideration past values of velocity and acceleration for the estimation of particulate matter.

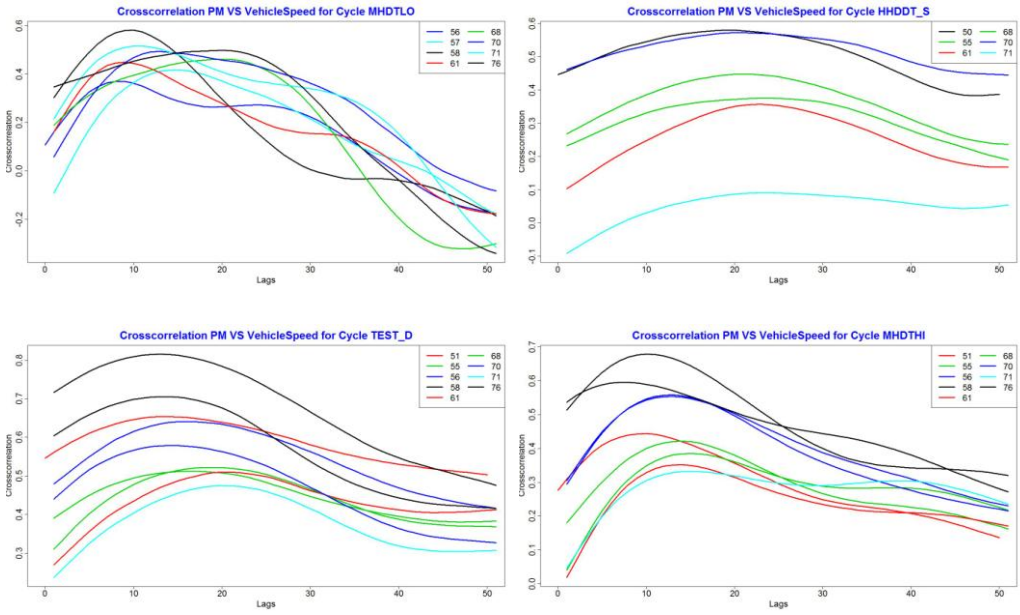



**Figure 3.7.** Cross-correlation between PM and velocity

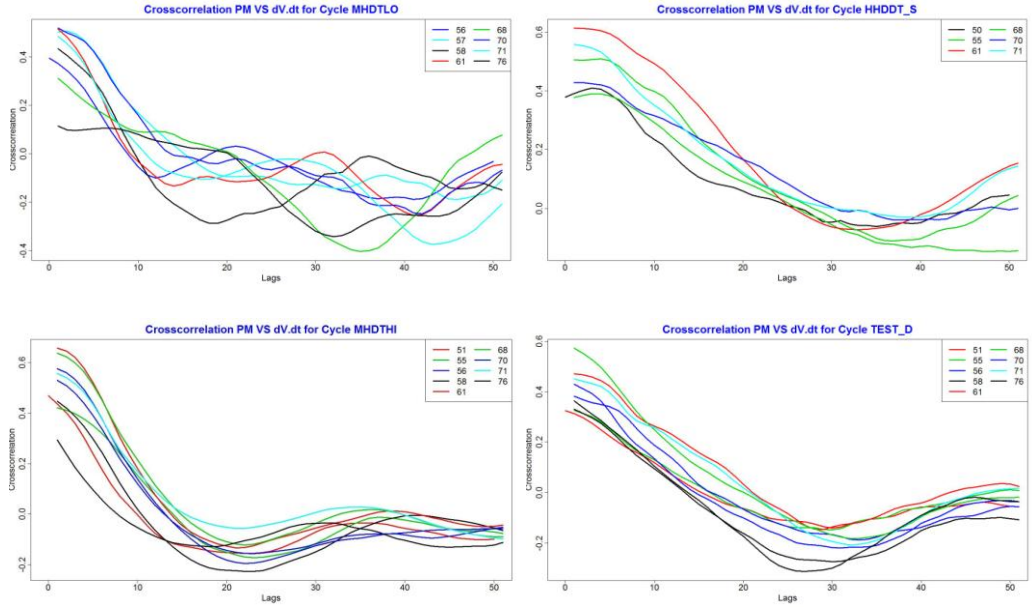



**Figure 3.8.** Cross-correlation between PM and acceleration

Based on the description and analysis of the data, we deduce that

- The trajectories of the variables are continuous and smooth.
- Values of particulate matter are significantly related to the instantaneous values of the driving behavior variables and not to the average of these.
- There is significant cross-correlation between particulate matter and driving behavior variables that is believed to be related to the distortion experienced by the response.

These are conclusive evidence that the model developed in this thesis is a good fit for the data. We discuss the estimate results in the next chapter.

CHAPTER 4

# PARTICULATE MATTER MODEL ESTIMATES

We examined general characteristics and behavior of the trajectories of particulate matter, velocity and acceleration in previous chapter. In this chapter, we discuss the estimation results of the statistical model for the prediction of the continuous trajectory of the particulate matter given a driving behavior. This model reports an alternative that is simpler and more accurate than the models discussed in chapter 1. In particular, we want to consider the smoothness and continuity of the data and the dependence over time between particular matter and the driving behavior variables.

## 4.1 MODEL

We consider particulate matter as the dependent variable and velocity and acceleration as the stimuli. The model developed for this case study identifies the relation between the instantaneous change of velocity and acceleration and the values of particulate matter even when the amount recorded by the emission analyzer is not the same amount of particulate matter released at tailpipe. The model is expressed as

$$PM_{ij}(t_i) = \alpha + \int_{t_{ij}-\delta_{vel}}^{t_{ij}} \beta_{vel}(w)\xi_{ijvel}(w)dw + \int_{t_{ij}-\delta_{accel}}^{t_{ij}} \beta_{accel}(w)\xi_{ijaccel}(w)dw + \varepsilon_{ij}(t)$$

where $i = 1, \ldots, 4$ (number of the driving cycles) and $j = 1, \ldots, 11$ (number of the trucks) and

$PM_{ij}(t_i)$ is the response or dependent variable

$\xi_{ijvel}(w)$ and $\xi_{ijaccel}(w)$ are the stimuli

$\beta_{vel}(w)$ and $\beta_{accel}(w)$ are the coefficient function

$\delta_{vel}$ and $\delta_{accel}$ are the length of the convolution

$\varepsilon_{ij}(t)$ is the random error

From the expression of the model, we can see that

- The components of the equation are represented as function of time. This is because we believe that the chassis dynamometer data are continuous and smooth.
- The model relates the current value of particulate matter with the convolution of the velocity and acceleration. This type of relation pertains to account for the distortion that the particles experienced when they were being recorded.
- The coefficient function weights the values of the velocity and acceleration which depends of how far away they are from the current value of the particulate matter.
- The model gives the trajectory of particulate matter.

The next section shows the values of the estimates for $\delta_{vel}$, $\delta_{accel}$, $\lambda_{vel}$, $\lambda_{accel}$, $\beta_{vel}(w)$ and $\beta_{accel}(w)$.

## 4.2 ESTIMATING PARAMETERS

We apply the same procedures discussed in chapter 2 to estimate the parameters $\delta_{vel}$, $\delta_{accel}$, $\beta_{vel}(w)$ and $\beta_{accel}(w)$.

### 4.2.1 VALUES FOR $\hat{\delta}_{vel}$ AND $\hat{\delta}_{accel}$

To estimate $\delta_{vel}$ and $\delta_{accel}$, we apply the cross-validation method discussed in previous chapter. In this case, the starting values were $\boldsymbol{\delta} = (\delta_{vel}, \delta_{accel}) = (1,1)$. The algorithm was stopped at $\boldsymbol{\delta} = (25,25)$. Figure 4.1 contains the results of this procedure. We observe that the values of $\delta_{vel}$ and $\delta_{accel}$ that minimize the Total Sum Square Error are $(\delta_{vel}, \delta_{accel}) = (21,21)$.

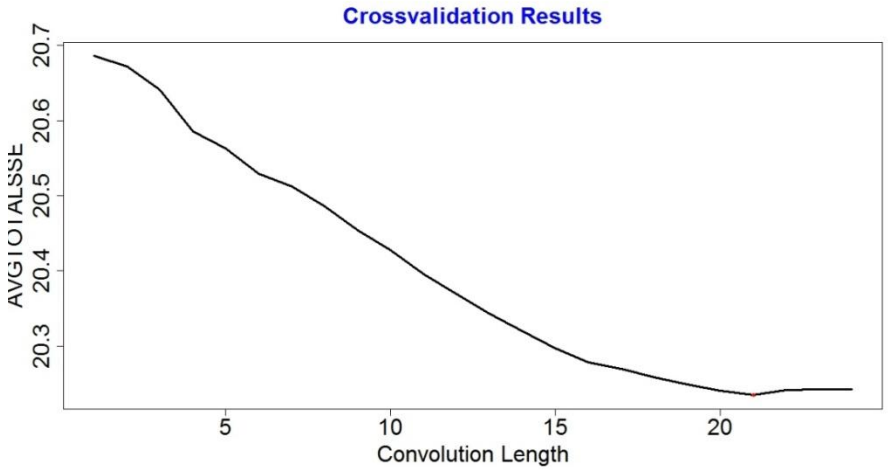



**Figure 4.1.** Total sum square error results to estimate $\boldsymbol{\delta}$

Figure 4.2 shows the sum square error results by vehicle. Here, most curves follow the same pattern and in general, they are minimizes by values greater than 15. However, there is one curve that shows strange behavior. This behavior is explained

in the chapter 6 where it is shown that one of the data samples from truck 76 is an outlier and the one causing this strange behavior.

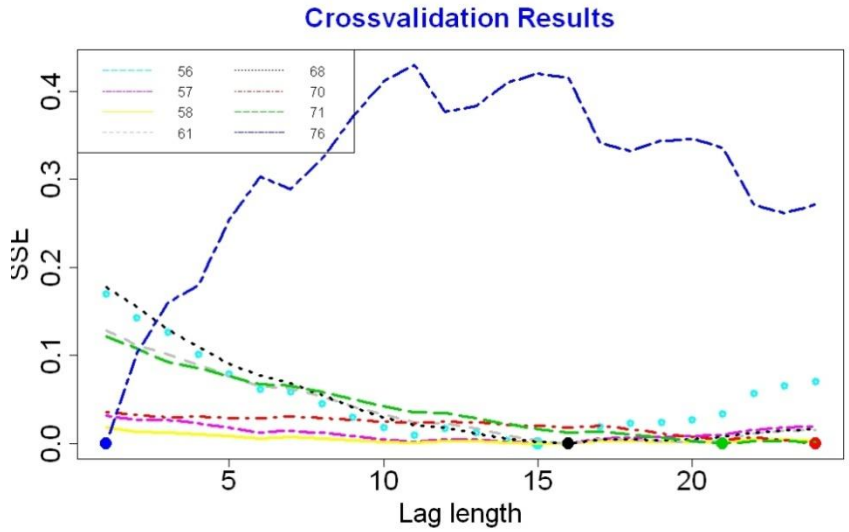



**Figure 4.2**. Sum square error results by vehicle

### 4.2.2 VALUES FOR $\lambda_{vel}$ AND $\lambda_{accel}$

To estimate $\boldsymbol{\lambda} = (\lambda_{vel}, \lambda_{accel})$, we also use the method of cross-validation. Unlike the plots shown in the last section, specific values of $\boldsymbol{\lambda}$ that minimize the Total Sum Square Error could not be found. That is, the TSSE did not reach a minimum value since it keeps decreasing as the values of $\boldsymbol{\lambda}$ were increasing as it is shown in Figure 4.3. This might be as result of the large amount of data to be considered. For that reason, we let $\hat{\boldsymbol{\lambda}} = \left(\hat{\lambda}_{vel}, \hat{\lambda}_{accel}\right) = (100,100)$.

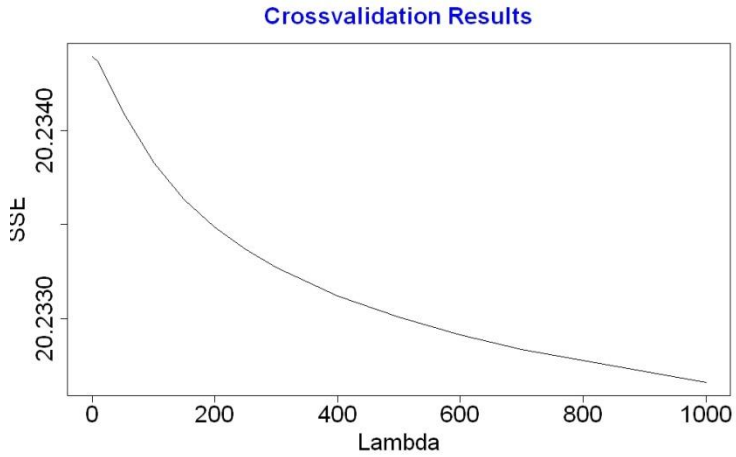



**Figure 4.3**. Total sum square error to estimate $\boldsymbol{\lambda}$

### 4.2.3 ESTIMATES FOR $\boldsymbol{\beta_{vel}(w)}$ AND $\boldsymbol{\beta_{accel}(w)}$

By previous section, we found that the length of convolution for the acceleration and velocity is length 21. Since they have the same length, we use the same linear basis expansion for each of them. That is, we represent the coefficients function as a basis expansion of 7 b-splines which are shown in Figure 4.4.

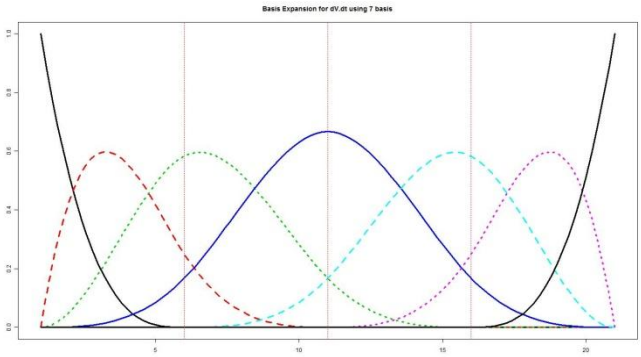



**Figure 4.4.** Basis functions

In addition, the order of these b-splines was chosen to be 4 and the range is from 1 to 21 which is the length of the convolution.

Let

$$\hat{\beta}_{vel}(w) = \sum_{k=1}^{7} \hat{c}_{velk}\, \phi_{velk}(w) \text{ and } \hat{\beta}_{accel}(w) = \sum_{k=1}^{7} \hat{c}_{accelk}\, \phi_{accelk}(w)$$

be the estimates of the coefficients by applying the procedure discussed in chapter 2.

Also, we estimate their confidence of interval (CI) by

$$\hat{\beta}_{vel}(w) \pm 2 * sd(\hat{\beta}_{vel}(w)) \text{ and } \hat{\beta}_{accel}(w) \pm 2 * sd(\hat{\beta}_{accel}(w)).$$

Figures 4.5 and 4.6 show the estimates of the coefficient function of both velocity and acceleration with their respective confidence intervals. We can see that both have different patterns and thus they affect the estimate of particulate matter differently.

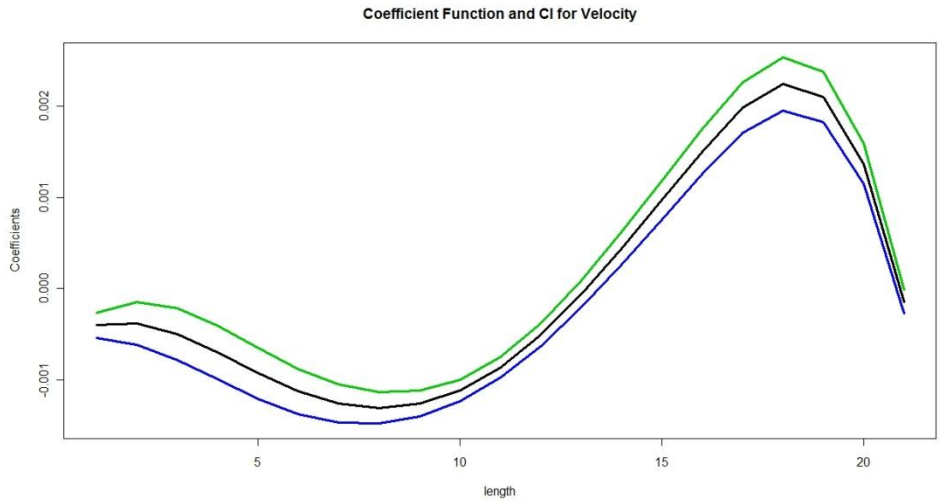



**Figure 4.5.** Coefficient function and CI for velocity

We see that the coefficient function give more weight to the past values of the velocity than to the present values as it is shown in Figure 4.5. This means that values of the velocity that are father away from the current value of particulate matter are more

important for the estimation of the particulate matter than the ones that are closer to it. The confidence intervals are close to the estimates of the coefficient estimates of the velocity.

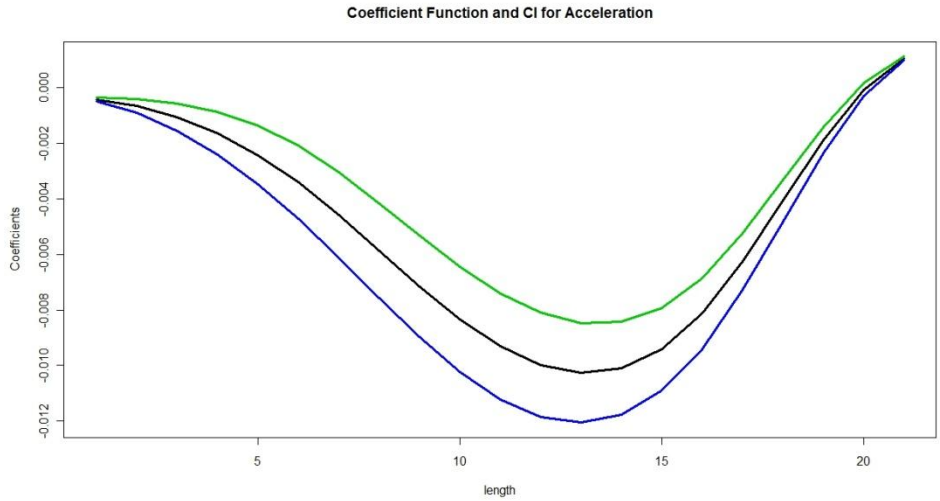



**Figure 4.6.** Coefficient function and CI for acceleration

The coefficient function for the acceleration instead weights more negatively the values of acceleration that are between 10 and 15 seconds away from the current value of particulate matter. We observe that the confidence intervals are wider than those found for the coefficient function of the velocity. There is more variability in the estimates of the acceleration than in the estimates of the velocity.

## 4.3 PREDICTIONS

Once the coefficient functions are estimated, we can predict the observations. This is done by

$$\widehat{PM}_{ij}(t_i) = \alpha + \int_{t_{ij}-\hat{\delta}_{vel}}^{t_{ij}} \hat{\beta}_{vel}(w)\xi_{ijvel}(w)dw + \int_{t_{ij}-\hat{\delta}_{accel}}^{t_{ij}} \hat{\beta}_{accel}(w)\xi_{ijaccel}(w)dw$$

where $i = 1, \dots, 4$ (number of the driving cycles) and $j = 1, \dots, 11$ (number of the trucks) and

$\widehat{PM_{ij}}(t_i)$ is prediction for the $i^{th}$ cycle from $j^{th}$ vehicle

$\xi_{ijvel}(w)$ and $\xi_{ijaccel}(w)$ are the stimuli

$\hat{\beta}_{vel}(w)$, $\hat{\beta}_{accel}(w)$ , $\hat{\delta}_{vel}$ and $\hat{\delta}_{accel}$ were estimated in previous sections

Figures 4.7 and 4.8 show the predictions of two observations or curves as a result of two different velocity patterns. We observe that although the patterns of the velocity are different, the model give good estimates of the trajectory of the particulate matter.

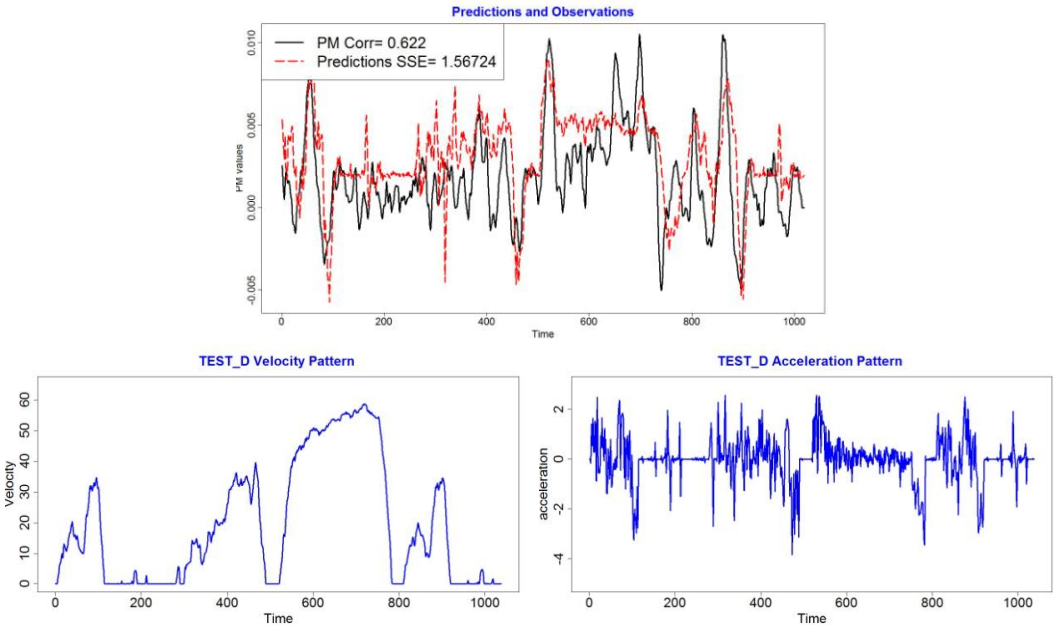



**Figure 4.7.** Comparing prediction and observation given TEST_D velocity pattern

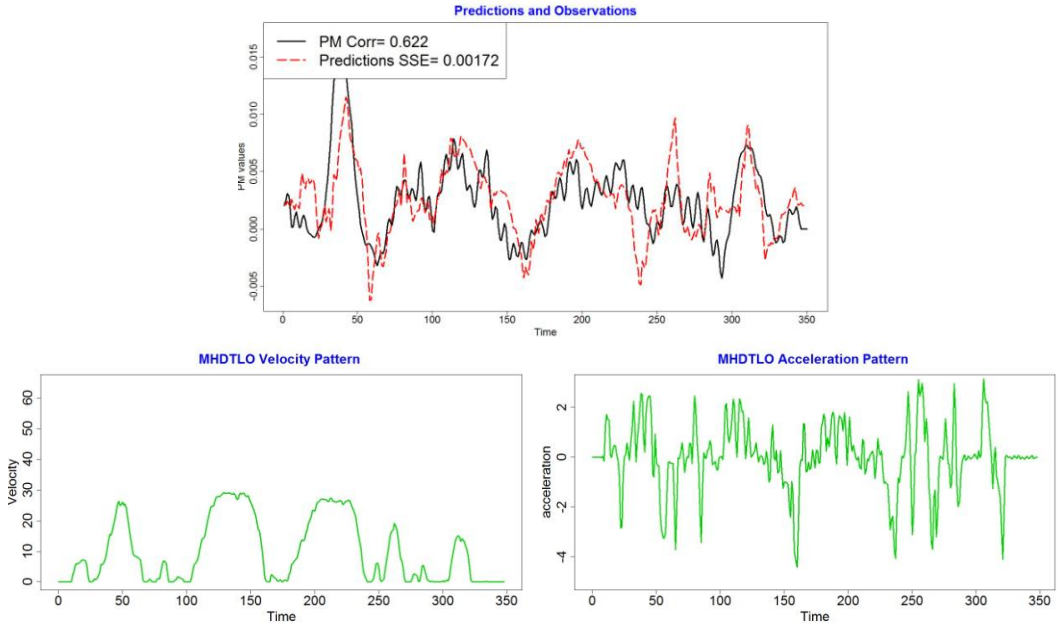



**Figure 4.8.** Comparing predictions and observation given MHDTLO velocity pattern

Looking at the previous figures, we can see that this model is capable of:

- Estimating the second-by-second trajectory of the particulate matter.
- Giving estimates for different trajectories of the particulate matter which depend on the applied velocity pattern.

This model accounts for the time structure, smoothness, and distortion of the data by integrating the components as function of time and convolution function for each of the driving behavior variables. In conclusion, we have a simpler model that can infer about how much particulate matter is produced during a time interval by a medium heavy-duty truck.

# CHAPTER 5

# BOOTSTRAPPING

In this chapter, the variability of the estimated coefficient function is explored by a novel bootstrapping method. This computational method allows inference about the certainty of the estimation of parameters without knowing about the distribution of the population of the sample [15]. Instead, the observations are used as the principal source of information of the entire population. This method was developed by Efron who showed its usefulness in several areas of statistics [16, 17]. In particular, we discuss the ”general” bootstrapping and the residual bootstrapping used for the classical linear regression. In these two techniques, the data bootstrapped is assumed to be identically distributed and independent (i.i.d.). However, there are cases in which the data does not fulfill the condition of independence. Several modifications of bootstrapping have been suggested with the purpose of taking into consideration the characteristic and dynamics of the model and specific parameter but also for data that do not follow the general assumptions accounted by Efron. One of these modifications is the method of block bootstrapping which permits to infer about the parameters when the data is fitted in a time series model and assumed to be serially correlated or weakly dependent [15]. We look upon to the general process of this statistical technique to examine its usefulness for the statistical model developed in this thesis. We specifically exploit the bootstrapping process used in the linear regression and block bootstrapping to extend this computational method to find the confidence interval of the estimated coefficients for the model developed in this

research project. This new method and its results are discussed at the end of the chapter.

## 5.1 GENERAL BOOTSTRAPPING

In the simplest case, the observations are assumed to be independent and identically distributed random variables. The main idea of bootstrapping is to resample with replacement the sample observations allowing an element of the sample observation to be repeated. The bootstrapped data have to be the same size as the sample observation. This "bootstrapped" data is treated as a new observation and used it to make inference about the parameter of interest. The key of this method is to realize and take advantage of the randomness characteristic of the observations and parameter estimates so that each resample or bootstrap is a new case or observation of the population [15]. By repeating the same process a number of times, we can state the accurateness and stability of the parameter estimation without needing to specify the data distribution but only relying in the observations. We show two different cases of how this method allows us to find the confidence interval and infer about how close the estimate is from the true.

**CASE 1**. Confidence Interval for mean $\theta$

Consider the i. i. d. $\{X_1, \dots, X_i, \dots, X_n\}$ from the unknown population distribution F. We are interested in estimated the mean which is estimated by $\hat{\theta} = E[X_i]$. Let assume that $\{x_1, \dots, x_i, \dots, x_n\}$ is the available sample observation drawn from the population.

The sample mean is estimated by

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

To find the confidence interval, the following process is done:

1. Resample with replacement sample observation. Let the new sample to be $\{x_{(1)}^b, \dots, x_{(i)}^b, \dots, x_{(n)}^b\}$.
2. Estimate the sample mean by
$$\hat{\theta}^b = \frac{1}{n}\sum_{i=1}^{n} x_i^b.$$
3. Repeat 1-2 B times.
4. Find the mean of $\{\hat{\theta}^1, \dots, \hat{\theta}^b, \dots, \hat{\theta}^B\}$ by
$$\bar{\hat{\theta}} = \frac{1}{B}\sum_{b=1}^{B} \hat{\theta}^b.$$
5. Find the standard deviation by
$$sd_{boot}(\hat{\theta}) = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(\hat{\theta}^b - \bar{\hat{\theta}})^2}.$$
6. Estimate the confidence interval by
$$\hat{\theta} \mp 2 * sd_{boot}(\hat{\theta}).$$

**CASE 2**. Confidence interval for the coefficient in the linear regression scenario

Consider the classical linear model

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$$

where

$\boldsymbol{Y}$ is a $nx1$ vector of responses or observations

$\boldsymbol{X}$ is a $nxp$ full rank matrix of known constant or predictors

$\boldsymbol{\beta}$ is a px1vector of unknown coefficients of the known constants

$\boldsymbol{\varepsilon}$ is a nx1 vector of unknown i. i. d. random errors

We are interested in finding the confidence interval of the coefficient $\boldsymbol{\beta}$. In this case, we resample the residuals since these are the components that are assumed to be random. Let $\widehat{\boldsymbol{\beta}}$ be the estimate of the coefficients. We follow the next steps to estimate the CI:

1. Estimate the predictions and residuals by $\widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{e}} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}}$.
2. Resample the residuals and let $\boldsymbol{e}^b = \{e^b_{(1)}, \dots, e^b_{(2)}, .., e^b_{(3)}\}$ be the new residuals.
3. Estimate the new responses by $\widehat{\boldsymbol{Y}}^{\boldsymbol{b}} = \widehat{\boldsymbol{Y}} + \boldsymbol{e}^b$.
4. Estimate the coefficient using the new responses by $\widehat{\boldsymbol{\beta}}^b = (\boldsymbol{X'X})^{-1}\boldsymbol{X'}\widehat{\boldsymbol{Y}}^b$.
5. Repeat steps 2-4 B times.
6. Estimate the bootstrap mean by
$$\bar{\widehat{\boldsymbol{\beta}}} = \frac{1}{B}\sum_{b=1}^{B}\widehat{\boldsymbol{\beta}}^b.$$
7. Estimate the standard deviation by
$$sd_{boot}\left(\widehat{\boldsymbol{\beta}}\right) = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(\widehat{\boldsymbol{\beta}}^b - \bar{\widehat{\boldsymbol{\beta}}})^2}.$$
8. Find CI by $\widehat{\boldsymbol{\beta}} \mp 2 * sd_{boot}\left(\widehat{\boldsymbol{\beta}}\right)$.

In these two cases, the data bootstrapped are discrete independent and identically distributed random variables. In the next section, we discuss the case in which the i. i. d. condition does not apply.

## 5.2 BLOCK BOOTSTRAPPING

The situation is more complicated when the observations are dependent such as in time series data. This is because not only is the population distribution unknown but the dependence model of the sample is also unknown [18]. This data is said to be weakly dependent. It is important to take in consideration this dependence to have an accurate estimate of the variance particular estimates. For this purpose, the block bootstrapping method has been proposed. In general, it follows the same steps as the previous cases discussed in the last section but with a small alteration which helps to take into account the dependency of the data. The idea is to divide the available sample of observations $(X_1, \dots, X_n)$ into subseries of length "$l$" which is the length of dependence of the data. These subseries are then considered to be i. i. d. and resampled with replacement. This block bootstrapped samples are then assumed to be new observations and used to infer about the parameter of interest. Although there exist several types of block bootstrap method, we discuss the one developed by Carlstein called non-overlapping block bootstrapping.

Let assume that $\{x_1, \dots, x_i, \dots, x_n\}$ is the available sample observation drawn from the population and thus, the sample mean is estimated by

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

To find the confidence interval, the following process is done:

1. Assume that observation have a length of dependence "$l$"

2. Divide the observations on non-overlapping blocks as $\left\{B_1, \dots, B_j, \dots, B_{\frac{n}{l}}\right\} = \{(x_1, \dots, x_l), \dots . (x_{n-l}, \dots, x_n)\}$ .
3. Resample with replacement the blocks $\left\{B_1, \dots, B_j, \dots, B_{\frac{n}{l}}\right\}$. Let the new sample to be $\{B^b_{(1)}, \dots, B^b_{(j)}, \dots, B^b_{(n/l)}\}$.
4. Estimate the sample mean $\hat{\theta}^b$ using the new sample.
5. Repeat 3-4 B times.
6. Find the mean of $\{\hat{\theta}^1, \dots, \hat{\theta}^b, \dots, \hat{\theta}^B\}$ by
$$\bar{\hat{\theta}} = \frac{1}{B}\sum_{b=1}^{B} \hat{\theta}^b.$$
7. Find the standard deviation by
$$sd_{boot}\left(\hat{\theta}\right) = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(\hat{\theta}^b - \bar{\hat{\theta}})^2}.$$
8. Estimate the confidence interval by
$$\hat{\theta} \mp 2 * sd_{boot}\left(\hat{\theta}\right).$$

The data has to be divided into blocks before it is resample and once this is done, the next steps are similar as before. The following section shows the results of the modification of block bootstrapping technique for the functional data and that helps to retain the smoothness properties of this type of data.

## 5.3 MODIFICATION OF BLOCK BOOTSTRAPPING

We are interested in inferring about the variability of the coefficients of the model developed in this research. As in the classical linear regression method, the residuals play an important role for the inference about the accuracy of the coefficient estimates. This is because they carry information concerning the appropriateness of the assumptions of the model. The model considered in this research is the following:

$$y_i(t) = \alpha + \sum_{p=1}^{P} \int_{t-\delta_p}^{t} \beta_p(w)\xi_{ip}(w)\, dw + \varepsilon_i(t)$$

where

$y_i(t)$ is the response or dependent variable

$\xi_{ip}(w)$ is the $p$ stimulus

$\beta_p(w)$ is the coefficient function of the $p$ stimulus

$\delta_p$ is the length of the convolution of the $p$ stimulus

$\varepsilon_i(t)$ the random error

$E(\varepsilon_i(t)) = 0$ and by letting $\varepsilon_i(t) = \{\varepsilon_{i1} \dots, \varepsilon_{iT}\}$ the covariance of the random error can be found by $cov(\varepsilon_{is}, \varepsilon_{is}) = \Sigma_{ss}$ $and$ $cov(\varepsilon_{is}, \varepsilon_{it}) = \Sigma_{st}$ where both are not equal to zero. That is, the random errors in this case have some serial dependency unlike of the common assumption for most linear regression models where the random errors are i. i. d. This is a restriction of the residual variance, requiring it to have a strict autocorrelation structure. The random error dependency can be found by looking at the autocovariance of the residuals. We take into consideration this behavior of the residuals to estimate the "general covariance". This "general covariance" is a key

component in the modification of the non-overlapped block bootstrapping. Moreover, the main purpose of this modification is to take advantage of the usefulness of the block bootstrapped method while also maintaining the smoothness of the data. We discuss first how this "general covariance" is estimated using the estimated residuals.

### 5.3.1 ESTIMATING THE GENERAL COVARIANCE

We assume that the covariance of the model is estimated as $\Sigma_{ts} = cov(y_i(t), y_i(s)) = cov\big(\varepsilon_i(t), \varepsilon_i(s)\big) = R(s,t)$. Under the assumption that $\{y_1(t), \dots, y_i(t), \dots, y_N(t)\}$ has stationary covariance, we can defined $R(s,t) = R(|s-t|)$ for $s, t \in T_i$.

This means that we can write the $\Sigma_i$ as the following:

$$\Sigma_i = \begin{pmatrix} R(0) & R(1) \cdots & R(n-1) \\ R(1) & R(0) \cdots & R(n-2) \\ \vdots & \vdots \cdots & \vdots \\ R(n-1) & R(n-2) \dots & R(0) \end{pmatrix}$$

Given that the random errors are not observed, we use the residuals to estimate the covariance. For each $y_i(t)$, we estimate the residuals by $\hat{\varepsilon}_i(t) = y_i(t) - \hat{y}_i(t)$ and the covariance matrix $\hat{\Sigma}_i$ by using the sample autocovariance. This is expressed as

$$(\hat{\Sigma}_i)_l = \hat{R}(|s-t|) = \frac{1}{T_i} \sum_{s}^{T_i - l} e_{s+l} e_s.$$

This is done for $\forall i$ where $i = 1, \dots, N$. To find the length of dependency, we look at the autocovariances estimates of all residuals. The value of *"$l$"* is decided to be the maximum lag value in which autocovariance values are significantly different from

zero. We let the estimate of the covariance estimate for each residual $i$ to be defined by the following expression:

$$\widehat{\mathbf{\Sigma}}_{\mathrm{i}} = \begin{pmatrix} \hat{R}_i(0) & \hat{R}_i(1) \cdots \hat{R}_i(l) & \dots 0 \\ \hat{R}_i(1) & \hat{R}_i(0) \cdots \hat{R}_i(l-1) & \dots 0 \\ \vdots & \vdots \cdots & \vdots \\ 0 & \dots & \hat{R}_i(0) \end{pmatrix}$$

Then, we find the general estimation of $\mathbf{\Sigma}$ by the mean of the estimated covariance matrices

$$(\widehat{\mathbf{\Sigma}}_1 + \cdots + \widehat{\mathbf{\Sigma}}_i + \cdots + \widehat{\mathbf{\Sigma}}_N)\frac{1}{N} = \widehat{\mathbf{\Sigma}}.$$

### 5.3.2 MODIFICATION

We developed a modification of the non-overlapping block bootstrapping method with the purpose of fixing the perturbation that the block bootstrapping methods caused to the functional residuals. We have that this method causes

- Discontinuity to the trajectory
- Alteration in the autocovariance

We consider the following relationship between the random error and residuals:

$$\varepsilon_i\,(t) \sim N(\mathbf{0}, \mathbf{\Sigma}^*) \Rightarrow \hat{\varepsilon}_i(t) \sim N(\mathbf{0}, \widehat{\mathbf{\Sigma}})$$

where $\widehat{\mathbf{\Sigma}}$ is the ”general covariance” which was estimated previously. We see that the covariance of the residuals has a defined structure $\widehat{\mathbf{\Sigma}}$. However, the smoothness

structure of the residuals and its covariance is perturbed or modified when block bootstrapping is applied. Instead, we have the following:

$$\hat{\varepsilon}_i(t)^b \sim N(\mathbf{0}, A\widehat{\mathbf{\Sigma}}\mathrm{A})$$

This implies

$$\hat{\varepsilon}_i(t)^b \sim N(\mathbf{0}, A\widehat{\mathbf{\Sigma}}\mathrm{A}) \neq \hat{\varepsilon}_i(t) \sim N(\mathbf{0}, \widehat{\mathbf{\Sigma}})$$

To recover the continuity and covariance structure of the data, the flowing procedure is applied:

1. Divide the curve residual $\hat{\varepsilon}_i(t)$ in blocks of length 10 as $\omega_i = \{\omega_{i1}, \dots, \omega_{1q_i}\} = \{(\hat{\varepsilon}_1, \dots, \hat{\varepsilon}_{10}), \dots (\hat{\varepsilon}_{N_i-10}, \dots, \hat{\varepsilon}_{N_i})\}$ where $q_i$ is the total number of blocks for the $i^{th}$ sample and defined by
$$q_i = \frac{T_i}{l}.$$
2. Resample with replacement the block residuals of case $i^{th}$ and let $\omega_i^b = \{\omega_{i(1)}^b, \dots, \omega_{i(2)}^b, \dots, \omega_{i(q_i)}^b\}$ be the bootstrapped residuals.
3. Apply step 1-2 for $\forall i$ where $i = 1, \dots, N$.
4. Assume that $\{\omega_1^b, \dots, \omega_i^b, \dots, \omega_N^b\} \sim N\left(0, A\widehat{\mathbf{\Sigma}}A'\right) = N(0, \widehat{\mathbf{\Sigma}}^b)$ .
5. Estimate the covariance of the block bootstrapped residuals $\widehat{\mathbf{\Sigma}}^b$ as discussed in previous section using $l$ and the autocovariance of $\omega_1^b, \dots, \omega_i^b, \dots, \omega_N^b\}$.
6. Estimate $A$ by
$$\widehat{\mathbf{\Sigma}}^{1/2}A' = \widehat{\mathbf{\Sigma}}^{b\,1/2}, A = \widehat{\mathbf{\Sigma}}^{-1/2}\widehat{\mathbf{\Sigma}}^{b\,1/2}.$$
7. Find the new smooth residuals by modifying the bootstrapped residuals $\{\omega_1^b, \dots, \omega_i^b, \dots, \omega_N^b\}$ by
$$\omega_1^b(t) = A\omega_1^b, \dots, \omega_N^b(t) = A\omega_N^b.$$

8. Estimate the new responses by $\hat{y}_i^b(t) = \hat{y}_i(t) + \omega_i^b(t)$.
9. Apply step 8 $\forall i$ where $i = 1, \dots, N$.
10. Estimate the coefficient using the new responses by

$$\widehat{\boldsymbol{\theta}}^b = (\boldsymbol{Z'Z} - \boldsymbol{R}(\boldsymbol{\lambda}))^{-1}\boldsymbol{Z'}\widehat{\boldsymbol{Y}}^b.$$

Refer to chapter 2 to look at the description of $\boldsymbol{\theta}, \boldsymbol{Z}, \boldsymbol{R}(\boldsymbol{\lambda})$ and $\boldsymbol{Y}$.

11. Repeat steps 2-10 B times.
12. Estimate the bootstrap mean of coefficients by

$$\bar{\widehat{\boldsymbol{\theta}}} = \frac{1}{B}\sum_{b=1}^{B}\widehat{\boldsymbol{\theta}}^b.$$

13. Estimate the standard deviation by

$$sd_{boot}\left(\widehat{\boldsymbol{\theta}}\right) = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(\widehat{\boldsymbol{\theta}}^b - \bar{\widehat{\boldsymbol{\theta}}})^2}.$$

### 5.3.3 RESULTS

In this section, we discuss two case scenarios for the application of the modification of block bootstrapping. These are done with the purpose of looking at the difference between parametric bootstrapping and non-parametric bootstrapping for the estimation of the variability.

#### *5.3.3.1 PARAMETRIC BOOTSTRAP*

Parametric bootstrap method uses data simulated from a distribution instead of the real data. This is because it is assumed that the data follows a distribution. Once the parameters of the distribution of the data or observations are estimated, then random variables are simulated from the distribution using the estimated parameters. These

random variables are then block bootstrapped. This is done with the purpose of showing that our modification works in other statistical settings.

We alter our block bootstrapped modification to handle parametric bootstrapping and follow the next steps:

1. Simulate 200 hundred random variables $\hat{x}_i(t)$of length 100 from the distribution $N(0, \widehat{\boldsymbol{\Sigma}})$ where $\widehat{\boldsymbol{\Sigma}}$ was estimated by using the residuals.
2. Divide the curve residual $\hat{x}_i(t)$ in blocks of length 10 as $\omega_i = \{\omega_{i1}, \dots, \omega_{10}\} = \{(\hat{x}_1, \dots, \hat{x}_{10}), \dots (\hat{x}_{100-10}, \dots, \hat{x}_{100})\}$ .
3. Resample with replacement the block of case $i^{th}$and let $\omega_i^b = \{\omega_{i(1)}^b, \dots, \omega_{i(2)}^b, \dots, \omega_{i10}^b\}$ be the bootstrapped samples.
4. Apply step 2-3 for $\forall i$ where $i = 1, \dots, 200$.
5. Assume that $\{\omega_1^b, \dots, \omega_i^b, \dots, \omega_{200}^b\} \sim N(0, A\widehat{\boldsymbol{\Sigma}}A') = N(0, \widehat{\boldsymbol{\Sigma}}^b)$ .
6. Estimate the covariance of the block bootstrapped residuals $\widehat{\boldsymbol{\Sigma}}^b$ as discussed in previous section using $l = 20$ and the autocovariance of $\{\omega_1^b, \dots, \omega_i^b, \dots, \omega_N^b\}$.
7. Estimate $A$ by
$$\widehat{\boldsymbol{\Sigma}}^{1/2}A' = \widehat{\boldsymbol{\Sigma}}^{b\,1/2}$$
$$A = \widehat{\boldsymbol{\Sigma}}^{-1/2}\widehat{\boldsymbol{\Sigma}}^{b\,1/2}.$$
8. Find the new smooth residuals by modifying the bootstrapped residuals $\{\omega_1^b, \dots, \omega_i^b, \dots, \omega_N^b\}$ by
$$\omega_1^b(t) = A\omega_1^b, \dots, \omega_N^b(t) = A\omega_N^b.$$
9. Estimate the autocovariance of $\omega_1^b(t) = A\omega_1^b, \dots, \omega_N^b(t) = A\omega_N^b$.

Figure 5.1 shows the autocovariance of the 200 simulated random variables, block bootstrapped random variables, and the modified block bootstrapped random variables. We observe that our modification reduced the distortion of autocovariance due to the block bootstrap method.

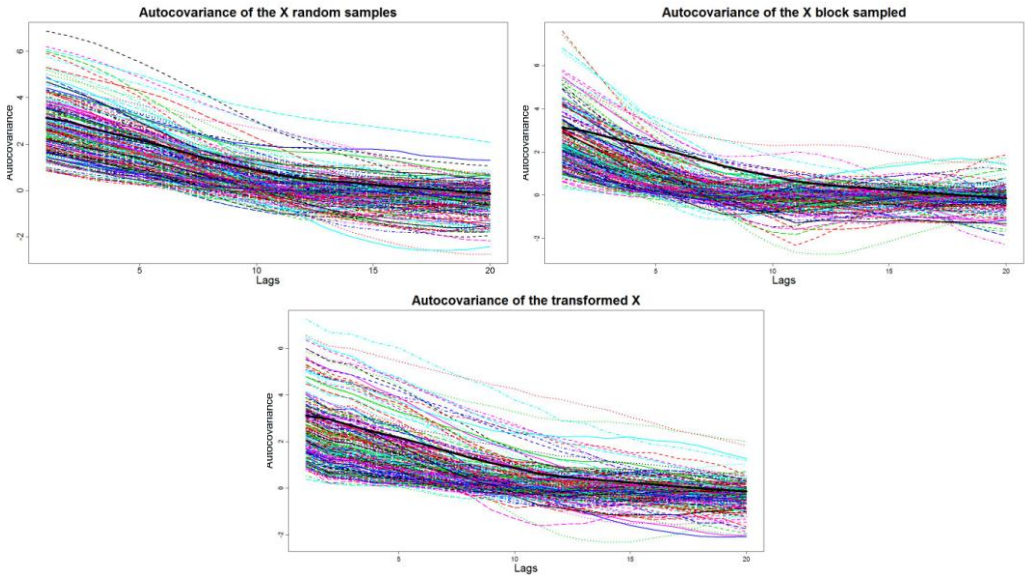



**Figure 5.1.** Autocovariance of the rv, block bootstrapped rv, and modified rv

Figure 5.2 shows what happens with the block bootstrapped random variables when the modification is applied. The black curve is the random variable simulated from the normal distribution. The red curve is the result of the random variable being block bootstrapped. As it can be seen, the curve of the block bootstrapped random variable is not continuous anymore. However, applying the transformation gives a smother curve (green line) by minimizing the discontinuity of the block bootstrapped random variable.

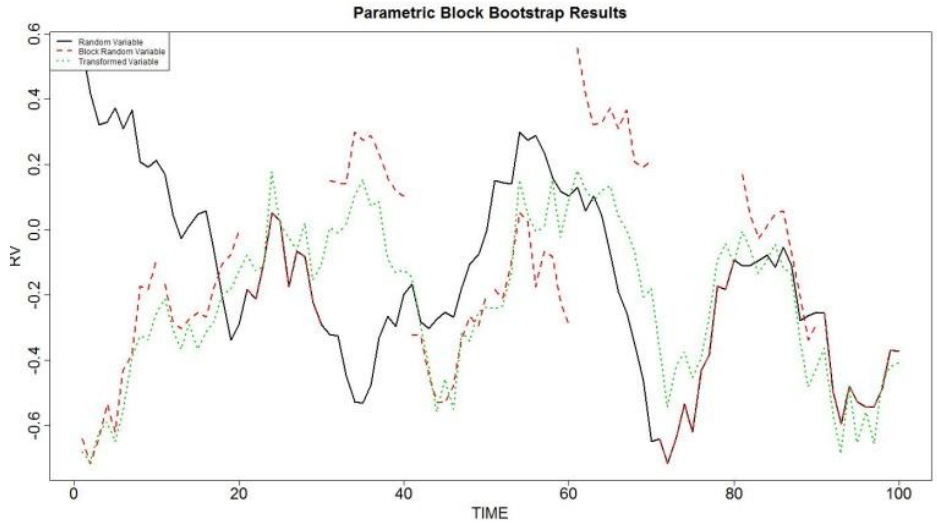



**Figure 5.2.** Curve of the rv, block bootstrapped rv and the modified rv

### *5.3.3.2 BOOTSTRAP USING REAL DATA*

In this section, we show the application of the modification of the block bootstrapping method using the functional residuals from the case study. We experienced some difficulty in the process of applying this method using the full data. We believe that this was as result of the high variability of the data. To show the application, we decided to use the samples of one particular truck which is referred in here as case $j$. First, we discuss how the "general covariance" is estimated using $n_i$ samples gathered from this truck $j$.

Let consider

$$\widehat{\boldsymbol{\theta}} = (\boldsymbol{Z_j}'\boldsymbol{Z_j} - \boldsymbol{R}(\boldsymbol{\lambda}))^{-1}\boldsymbol{Z_j}'\boldsymbol{Y_j} \text{ and } \widehat{PM}_i(t) = \widehat{\boldsymbol{\theta}}\boldsymbol{Z_{ij}} \text{ for } i = 1, \dots, n_j$$

where $n_j$ is the number of samples from truck $j$. The elements $\boldsymbol{Z_j}$ and $\boldsymbol{Y_j}$ contain only the information of truck $j$.

We estimate the residuals by $\hat{\varepsilon}_i(t) = PM_i(t) - \widehat{PM}_i(t)$ for $\forall i$ where $i = 1, \dots, n_j$. The autocovariance values for this residual is found by

$$\hat{R}_i(h) = \frac{1}{T_i} \sum_{t=1}^{T_i - h} (\hat{\varepsilon}_{it+h})(\hat{\varepsilon}_{it}), h = 0, \dots, T_i - 1$$

To find the length of dependency or $l$, we look at the plots of the sample autocovariances for this case. We see that most of them decrease and get close to zero around lag 20 as it is shown in 5.3.

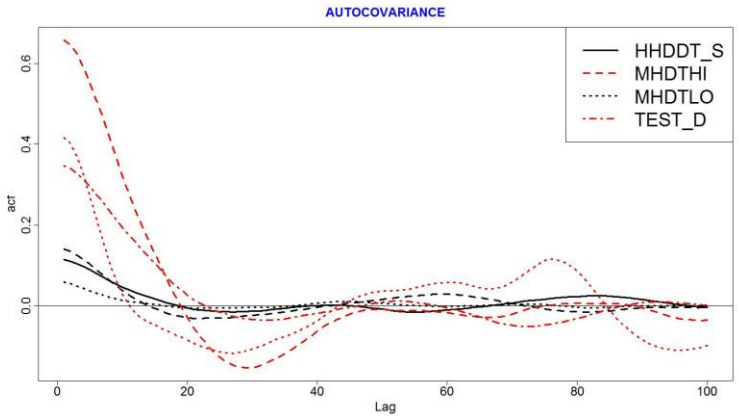



**Figure 5.3.** Autocovariance for the residuals of one vehicle

Based on this information, we considered the following:

$$R_i(h) = \begin{cases} \neq 0 \; if \; h \leq 20 \\ 0 \; if \; h > 20 \end{cases}, \quad \forall i$$

This implies that $l = 20$.

We let the estimate of the covariance estimate for each sample or functional residual $i^{th}$ to be defined as

$$\widehat{\boldsymbol{\Sigma}}_{\mathrm{i}} = \begin{pmatrix} \hat{R}_i(0) & \hat{R}_i(1) \cdots \hat{R}_i(20) & \dots 0 \\ \hat{R}_i(1) & \hat{R}_i(0) \cdots \hat{R}_i(19) & \dots 0 \\ \vdots & \vdots \cdots & \vdots \\ 0 & \dots & \hat{R}_i(0) \end{pmatrix}$$

Then, we find the general estimation of $\boldsymbol{\Sigma}$ by the mean of the estimated covariance matrices

$$(\widehat{\boldsymbol{\Sigma}}_1 + \cdots + \widehat{\boldsymbol{\Sigma}}_i + \cdots + \widehat{\boldsymbol{\Sigma}}_{n_i}) \frac{1}{n_j} = \widehat{\boldsymbol{\Sigma}}.$$

Then, the modified block bootstrapping is applied for this case as follow:

1. Let $\widehat{\boldsymbol{\theta}} = (\boldsymbol{Z_j}'\boldsymbol{Z_j} - \boldsymbol{R}(\boldsymbol{\lambda}))^{-1}\boldsymbol{Z_j}'\boldsymbol{Y_j}$
2. Divide the curve residual $\hat{\varepsilon}_i(t)$ in blocks of length 10 as $\omega_i = \{\omega_{i1}, \dots, \omega_{1q_i}\} = \{(\hat{\varepsilon}_1, \dots, \hat{\varepsilon}_{10}), \dots (\hat{\varepsilon}_{T_i-10}, \dots, \hat{\varepsilon}_{T_i})\}$ where $q_i$ is the total number of blocks for the $i^{th}$ sample for the $j$ vehicle and defined by
$$q_i = \frac{T_i}{10}.$$
3. Resample with replacement the block residuals of case $i^{th}$ and let $\omega_i^b = \{\omega_{i(1)}^b, \dots, \omega_{i(2)}^b, \dots, \omega_{i(q_i)}^b\}$ be the bootstrapped residuals.
4. Apply step 1-2 for $\forall i$ where $i = 1, \dots, n_j$.
5. Assume that $\{\omega_1^b, \dots, \omega_i^b, \dots, \omega_{n_j}^b\} \sim N\left(0, A\widehat{\boldsymbol{\Sigma}}A'\right) = N(0, \widehat{\boldsymbol{\Sigma}}^b)$ .
6. Estimate the covariance of the block bootstrapped residuals $\widehat{\boldsymbol{\Sigma}}^b$ as discussed in previous section using $l = 20$ and the autocovariance of $\{\omega_1^b, \dots, \omega_i^b, \dots, \omega_{n_j}^b\}$.

7. Estimate $A$ by

$$\widehat{\boldsymbol{\Sigma}}^{1/2}A' = \widehat{\boldsymbol{\Sigma}}^{b\,1/2}, A = \widehat{\boldsymbol{\Sigma}}^{-1/2}\widehat{\boldsymbol{\Sigma}}^{b\,1/2}.$$

8. Find the new smooth residuals by modifying the bootstrapped residuals $\{\omega_1^b, \dots, \omega_i^b, \dots, \omega_{n_j}^b\}$ by

$$\omega_1^b(t) = A\omega_1^b, \dots, \omega_{n_j}^b(t) = A\omega_{n_j}^b.$$

9. Estimate the new responses by $\widehat{PM}_i^b(t) = \widehat{PM}_{\boldsymbol{i}}(t) + \omega_i^b(t)$.
10. Apply step 9 $\forall i$ where $i = 1, \dots, n_j$
11. Estimate the coefficient using the new responses by

$$\widehat{\boldsymbol{\theta}}^b = (\boldsymbol{Z_j}'\boldsymbol{Z_j} - \boldsymbol{R}(\boldsymbol{\lambda}))^{-1}\boldsymbol{Z_j}'\boldsymbol{Y}_j^b.$$

Refer to chapter 2 to look at the description of $\boldsymbol{\theta}, \boldsymbol{Z}, \boldsymbol{R}(\boldsymbol{\lambda})$ and $\boldsymbol{Y}$.

12. Repeat steps 2-11 B times.
13. Estimate the bootstrap mean of coefficients by

$$\bar{\widehat{\boldsymbol{\theta}}} = \frac{1}{B}\sum_{b=1}^{B}\widehat{\boldsymbol{\theta}}^b.$$

14. Estimate the standard deviation by

$$sd_{boot}\left(\widehat{\boldsymbol{\theta}}\right) = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(\widehat{\boldsymbol{\theta}}^b - \bar{\widehat{\boldsymbol{\theta}}})^2}$$

Figure 5.4 shows the result of applying the block bootstrapped technique to the residuals. The green curve shows the autocovariance of the block bootstrapped residuals. We can see that these autocovariances are smaller than the "general autocovariance". Using the information from these block bootstrapped residuals might give us the wrong estimate for the variability and thus the wrong confidence interval estimates.

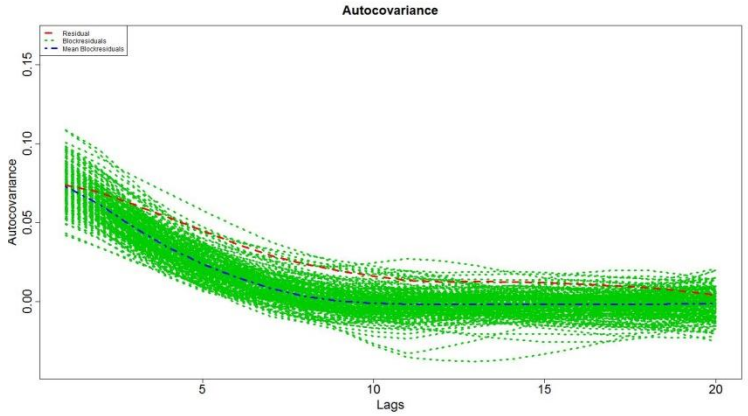



**Figure 5.4.** Autocovariance results after the block bootstrapping application

Figure 5.5 shows that the autocovariance of the restored block bootstrapped residuals. We can see that the autocovariance of these are scattered around the "general covariance". In particular, their mean (black curve) is close to the "general covariance" (red curve).

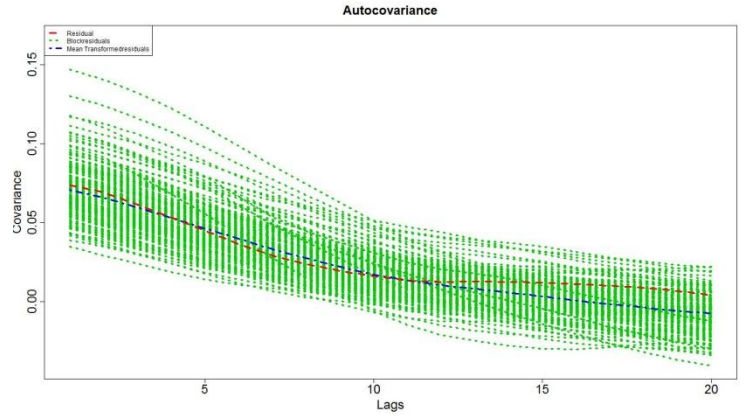



**Figure 5.5.** Autocovariance result after applying the modification

Figure 5.6 shows the result of applying the block bootstrapped and modification to one residual. We observe that the block bootstrapped residual (red curve) is not

continuous but after the application of the modification this discontinuity decreases (green curve).

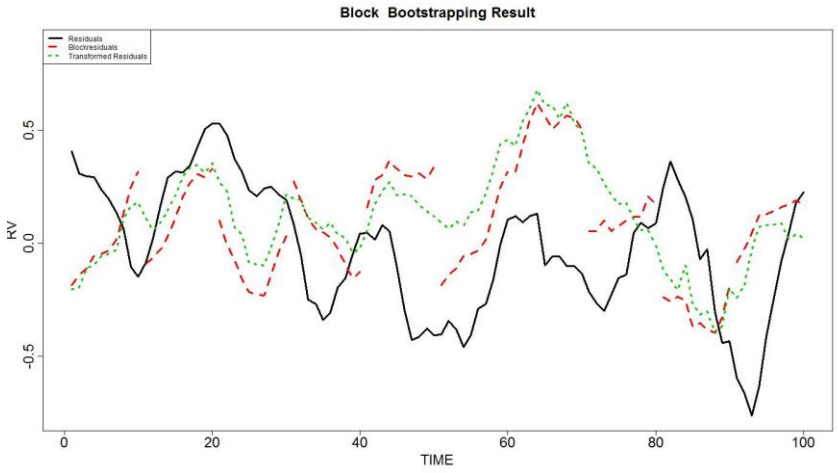



**Figure 5.6.** Example of one block bootstrapped residual

In addition, we show the coefficient of the convolution and confidence intervals for the velocity and acceleration in Figures 5.7 and 5.8 respectively.

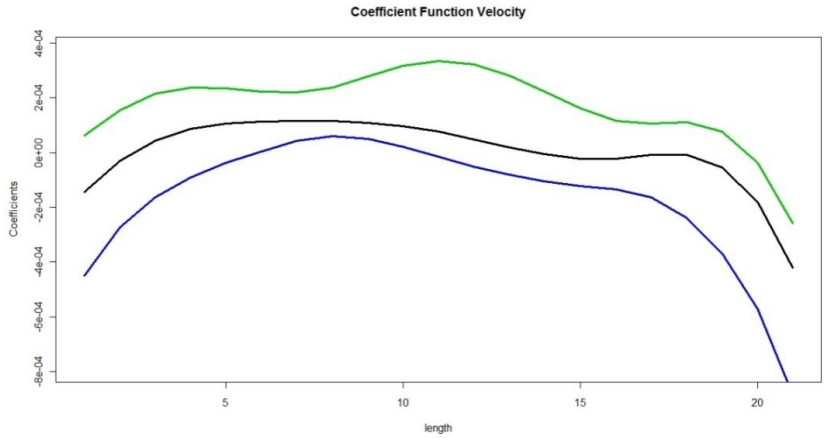



**Figure 5.7.** Confidence Interval for the convolution function of the velocity

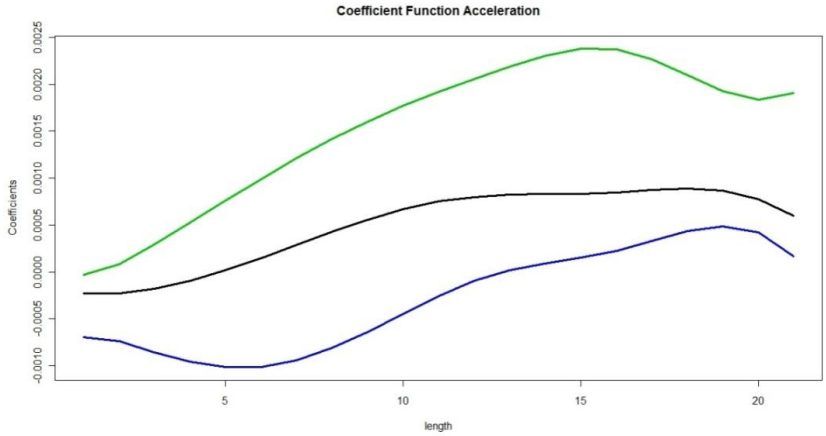



**Figure 5.8.** Confidence interval for the convolution of the acceleration

The confidence intervals were found by using the modification of block bootstrapping. We observe that

- The patterns of the coefficient of the convolutions differ from the patterns found in chapter 4.
- The confidence intervals are wider than the confidence interval found by the delta method.

We plan to study further this phenomenon. The discussion is expanded in chapter 7.

# CHAPTER 6

# INFLUENCE AND OUTLIERS

We discussed the methodology for estimating and finding the variability of the parameter of our statistical model in previous chapters. It is also important to know how these estimates depend on the information given by the data. That is, we want to know how robust and stable the estimates of the coefficients are and if we can rely on these estimates whenever the data is perturbed or changed. One method commonly use to consider this situation is removing one or a group of points at a time from the full data and quantifying the effects in the estimates of the linear regression using statistical methods developed for this purpose [19]. In the classic linear regression, clear examples have been shown in which the deletion of points in the data causes a significant effect in the coefficient estimates. These cases are commonly called influence and/or outlier points. After additional examination of the nature of the points, it might be decided that their contribution is not important and be deleted completely from the full data [19, 20]. Cook's and Mahalanobis distance are two common methods use for finding influence and outliers cases in the data respectively. Our motive of this chapter is to expand the applicability of these two statistics for the identification of curve data which are serially correlated or weakly dependent. In the following sections, we discuss their standard application in the linear regression, then we elaborate on the case of using curve data instead of discrete points, and lastly, we show the results for a specific case scenario.

## 6.1 INFLUENCE AND OUTLIER IN CLASSICAL LINEAR REGRESSION

There is interest on knowing how the estimates of the coefficients are affected when the data is perturbed or modified [19]. This perturbation allows deducing about the robustness and stability of the coefficients and their reliability on estimating the responses. The method of case deletion of one element or a group of elements from the full model has usually been used for coefficient inference of the classic linear regression model [19].

Consider the usual linear model

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}, var(\boldsymbol{\varepsilon}) = \sigma^2\boldsymbol{I}$$

where

$\boldsymbol{Y}$ is a $nx1$ vector of responses or observations

$\boldsymbol{X}$ is a $nxp$ full rank matrix of known constant or predictors

$\boldsymbol{\beta}$ is a px1vector of unknown coefficients of the known constants

$\boldsymbol{\varepsilon}$ is a nx1 vector of unknown i. i. d. random errors

$\boldsymbol{I}$ is the nxn identity matrix

$\sigma^2$is the common scalar variance

The model states that the response $\boldsymbol{Y}$ is linearly related to the predictors $\boldsymbol{X}$ and this relation is quantified by the coefficients $\boldsymbol{\beta}$ which need to be estimated. In addition, it is also assumed that the random error $\boldsymbol{\varepsilon}$ are uncorrelated and each has common variance $\sigma^2$. The estimate of $\boldsymbol{\beta}$ can be found by $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$. By using the full

data for the estimation of the coefficients, it is explicitly suggested that this specific model is appropriate for all the elements of the data. The model is said to be robust and this depends on the well behavior and the appropriate assumptions about the data.

However, a point or a group of points of the data might behave differently from the rest of the other points in the data. These types of disagreement on the behavior of these points can sometimes be identified visually however this is not always the case, especially for functional data. It is important to determine if there exist points in the data that do not agree with the assumptions made. Once this has been done, further investigation is advised to justify this behavior which can be a remote or extreme result of the experiment or error measurement. As consequence, a decision can be made on whether to remove it permanently from the full data or classifying it as an important point for the model estimation.

A common way to infer about the effects of each point has in the model estimation is the method of case deletion. The basic idea is summarize in the following steps for a particular $i$ point of the data:

1. Estimate $y_i$ by $\hat{y}_i = \boldsymbol{X}_i\widehat{\boldsymbol{\beta}}$ where $\boldsymbol{X}_i$ is the $1 \times p$ vector of known constants.
2. Take out the element $y_i$ and respective covariates $\{x_1, \dots, x_p\}$ from the full data.
3. Define the components of the linear model as $\boldsymbol{Y}^{(-i)}, \boldsymbol{X}^{(-i)}, \boldsymbol{\beta}^{(-i)}$ to express that the $i^{th}$ case has been taken out from the data. These are the components of the linear regression using the reduce data.
4. Estimate $\widehat{\boldsymbol{\beta}}^{(-\boldsymbol{i})} = (\boldsymbol{X}'^{(-i)}\boldsymbol{X}^{(-i)})^{-1}\boldsymbol{X}'^{(-i)}\boldsymbol{Y}^{(-i)}$.

As it can be seem, the estimates of the coefficients using the full and reduced data are found. We are interested to see if deleting the $i^{th}$ data made a difference in the estimation of the coefficients $\boldsymbol{\beta}$ or in the predictions. A simple way to observe this is by subtracting the estimates from both cases. For example, we can find the difference between $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\beta}}^{(-i)}$ by

$$EIC_i = (\widehat{\boldsymbol{\beta}}^{(-i)} - \widehat{\boldsymbol{\beta}})$$

This difference is called the empirical influence curve $(EIC)$ [19]. The value or values of $EIC_i$ clearly show how the coefficients are modified by the perturbation of the data. We can apply the same process for all points and find the empirical influence curve values for $\forall i$ where $i = 1, \dots, n$. Clearly, the values $\{EIC_1, \dots, EIC_i, \dots, EIC_n\}$ give us feedback about the effect that each point has in the estimation of the coefficients. Nonetheless, seeing which point has the largest effect is not a direct task unless $p = 1$ or $p = 2$. This is because they can be represented and analyzed geometrically [19]. The following two methods provide a better way to compare the differences between the case using full and reduced data whenever $p > 1$. Cook's and Mahalanobis Distance allow us to expose the data points that are influence and/or outlier cases respectively.

### 6.1.1 COOK'S DISTANCE

Cook developed a modification of empirical influence curve for the purpose of finding points that have a significant influence in the estimation of components of the linear regression model when the perturbation imposed to the data is deletion of cases [19]. This statistic is called Cook's Distance and allows inferring about the stability and variation in the estimation results. This statistic is useful to compare how influential each point is in p-dimensional space by given a scalar value for the quantification.

#### *6.1.1.1 LINEAR REGRESSION CASE*

Consider the same linear model explained in section 6.1 and the perturbation of deleting the $\boldsymbol{i^{th}}$ element from the full data. The Cook's distance is then defined as

$$D_i(\boldsymbol{X}'\boldsymbol{X}, ps^2) = \frac{(\widehat{\boldsymbol{\beta}}^{(-i)} - \widehat{\boldsymbol{\beta}})'(\boldsymbol{X}'\boldsymbol{X})(\widehat{\boldsymbol{\beta}}^{(-i)} - \widehat{\boldsymbol{\beta}})}{ps^2}$$

$$= \frac{EIC_i{}'(\boldsymbol{X}'\boldsymbol{X})EIC_i}{ps^2}$$

where

$\widehat{\boldsymbol{\beta}}^{(-i)}$ is the estimation of the coefficient using the reduced data

$\widehat{\boldsymbol{\beta}}$ is the estimation of the coefficient using the full data

$\boldsymbol{X}$ the design matrix for the full data

$p$ is degree of freedom

$EIC_i$ is the empirical influence curve discuss previously

$s^2$ is the sample variance estimated as

$$s^2 = \frac{\left(\widehat{\boldsymbol{Y}} - \boldsymbol{Y}\right)^{'}(\widehat{\boldsymbol{Y}} - \boldsymbol{Y})}{n-p} = \frac{\boldsymbol{e'e}}{n-p}$$

The $D_i(\boldsymbol{X}^{'}\boldsymbol{X}, ps^2)$ statistic is the measure or distance between $\widehat{\boldsymbol{\beta}}^{(-i)}$ and $\widehat{\boldsymbol{\beta}}$ standardized and scaled by $\boldsymbol{X'X}$ and $ps^2$ respectively [19]. This statistic does not change whenever the scale $ps^2$ is changed and rows are removed from $\boldsymbol{X}$ [21]. However, the most powerful property of this statistic is that, once the values $\{D_1, \dots, D_i, \dots, D_n\}$ are found, they can be categorized as influential by looking if $D_i \geq 1$ for $\forall i$ [19]. This cut-off is suggested when it can be assumed that $D_i's$ follow the $F(\,p, n-p)$ distribution which implies that $(1-\alpha)x100\%$ confidence regions can be found for $\boldsymbol{\beta}$ by

$$\{\boldsymbol{\beta'}|\,\frac{(\boldsymbol{\beta'} - \widehat{\boldsymbol{\beta}})^{'}(\boldsymbol{X}^{'}\boldsymbol{X})(\boldsymbol{\beta'} - \widehat{\boldsymbol{\beta}})}{ps^2} < F(1-\alpha, p, n-p\}$$

and since most random variables with distribution $F(.5, p, n-p)$ are close to 1, we delete the $i^{th}$ case if $D_i \geq 1$. This is because deleting $i^{th}$ move the estimate of $\boldsymbol{\beta}$ to 50% confidence region which is significant [19, 20, 21]. It is also important to notice that this not always true and thus we have to be careful with this cut-offs. It is better to examine further the results to find out if this is an influential point.

$D_i(\boldsymbol{X}^{'}\boldsymbol{X}, ps^2)$ can be rewritten as the following:

$$D_i(\boldsymbol{X}^{'}\boldsymbol{X}, ps^2) = \frac{(\widehat{\boldsymbol{Y}}^{(-i)} - \widehat{\boldsymbol{Y}}\,)^{'}(\widehat{\boldsymbol{Y}}^{(-i)} - \widehat{\boldsymbol{Y}}\,)}{ps^2}$$

In this case, we can interpret this statistic as how important the data point $i^{th}$ is for the estimation of all responses.

### 6.1.2 MAHALANOBIS DISTANCE

Mahalanobis distance allows finding how far away or different a specific point is from the other points [20, 21]. By using this statistic, we can observe if there is a data point that is out of the ordinary or do not follow the same behavior as the others. One example or case scenario in which this statistic has been largely used is to find how far away the random variables $\{x_1, \dots, x_n\}$ are from their expected value $u$ [20].

In this case, we have that the Mahalanobis Distance is defined as

$$D_i^2 = (x_i - u)'\Sigma^{-1}(\mathrm{x_i} - \mathrm{u})$$

where $E(x_i) = u$ and $cov(x_i) = \Sigma$ for $\forall i$ where $i = 1, \dots, n$[20].

The distance is estimated by using

$$\widehat{D}_i^2 = (x_i - \bar{x})' S^{-1}(x_i - \bar{x})$$

where $\bar{x}$ and $S$ are the sample mean and covariance respectively [20]. We can see that larger the distance is the less likely the sample mean can characterize the distribution of the $i^{th}$ random variable. We look at how this statistic can be also applied to the linear regression case in the following section.

#### *6.1.2.1 LINEAR REGRESSION CASE*

Consider the same linear model discussed in section 6.1 and the perturbation of deleting the $\boldsymbol{i^{th}}$ element from the full data. The Mahalanobis Distance is defined as

$$D_i^2 = (y_i - \hat{y}_i^{(-i)})' \Sigma^{-1} (y_i - \hat{y}_i^{(-i)})$$

where

$y_i$ is the $i^{th}$observation

$\hat{y}_i^{(-i)}$ is the estimation of the $i^{th}$ observation using the reduce data

$\Sigma$ is the sample covariance

In this case, we can interpret this statistic as how important the data point $i^{th}$ is for the estimation its own estimation.

## 6.2 DISTANCES MODIFICATION

We are interested in inferring about how influential or out of ordinary each individual sample is accounting that they are curves or function of time. We also consider that $\varepsilon_i(t) \sim N(0, \Sigma)$ for $\forall i$, using the estimated covariance $\widehat{\boldsymbol{\Sigma}}_i$ for each residual $i$ obtained from the autocovariance estimate described in section 5.3. We explain how the dependency of the data is integrated in the Cook's and Mahalanobis Distance method.

### 6.2.1 COOK'S DISTANCE

To take into consideration that the residuals are serially correlated, we use the "general covariance" discussed in chapter 5 for the estimation of Cook's distance. We follow the next procedure:

1. Estimate $\widehat{\boldsymbol{\theta}} = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{Y}$.
2. Estimate
$$cov(\boldsymbol{Y}) = (\widehat{\boldsymbol{\Sigma}}_1 + \cdots + \widehat{\boldsymbol{\Sigma}}_i + \cdots + \widehat{\boldsymbol{\Sigma}}_N)\frac{1}{N} = \widehat{\boldsymbol{\Sigma}}.$$
3. Remove recording $y_i(t)$ and its respective $\{\xi_{i1}(t), \dots, \xi_{iP}(t)\}$ from the data and thus the new data is defined as $Y^{(-i)} = \{y_1(t), \dots, y_{i-1}(t), \dots, y_N(t)\}$ for the recordings and $\xi^{(-i)} = \{\xi_{11}(t), \dots, \xi_{1P}(t)\}, \dots \{\xi_{N1}(t), \dots, \xi_{NP}(t)\}$ for the stimulus. We have that $(-i)$ represents the missing observation in the new data.
4. Estimate $\widehat{\boldsymbol{\theta}}^{(-i)} = (\boldsymbol{Z}'^{(-i)}\boldsymbol{Z}^{(-i)})^{-1}\boldsymbol{Z}'^{(-i)}\boldsymbol{Y}^{(-i)}$.
5. Find the estimates of $\boldsymbol{Y}^i$ by
    a. $\widehat{\boldsymbol{Y}}^i = \widehat{\boldsymbol{\theta}}\boldsymbol{Z}$
    b. $\widehat{\boldsymbol{Y}}^{(i)} = \widehat{\boldsymbol{\theta}}^{(-i)}\boldsymbol{Z}^{(i)}$
6. Find the Cook's Distance by
$$D_i = \frac{(\widehat{\boldsymbol{Y}}^i - \widehat{\boldsymbol{Y}}^{(i)})(' (\boldsymbol{Y}^i - \widehat{\boldsymbol{Y}}^{(i)})}{trace\,((\boldsymbol{Z}'\boldsymbol{Z} - \boldsymbol{R}(\boldsymbol{\lambda}))^{-1}\boldsymbol{Z}')(\frac{sum\,(dia\,g(\Sigma))}{T_i})}.$$
7. Repeat steps 3-6 for i=1,..., N.

### 6.2.2 MAHALANOBIS DISTANCE

In the same manner as in Cook's Distance, we modify Mahalanobis Distance by substituting the estimated autocovariance for Σ.

1. Estimate $\widehat{\boldsymbol{\theta}} = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{Y}$.
2. Estimate
$$cov(\boldsymbol{Y}) = (\widehat{\boldsymbol{\Sigma}}_1 + \cdots + \widehat{\boldsymbol{\Sigma}}_i + \cdots + \widehat{\boldsymbol{\Sigma}}_N)\frac{1}{N} = \widehat{\boldsymbol{\Sigma}}.$$
3. Remove recording $y_i(t)$ and its respective $\{\xi_{i1}(t), \dots, \xi_{iP}(t)\}$ from the data and thus the new data is defined as $Y^{(-i)} = \{y_1(t), \dots, y_{i-1}(t), \dots, y_N(t)\}$ for the recordings and $\xi^{(-i)} = \{\xi_{11}(t), \dots, \xi_{1P}(t)\}, \dots \{\xi_{N1}(t), \dots, \xi_{NP}(t)\}$ for the stimulus. We have that $(-i)$ represents the missing observation in the new data.
4. Estimate $\widehat{\boldsymbol{\theta}}^{(-i)} = (\boldsymbol{Z}'^{(-i)}\boldsymbol{Z}^{(-i)})^{-1}\boldsymbol{Z}'^{(-i)}\boldsymbol{Y}^{(-i)}$.
5. Find the estimates of $y_i$ by $\hat{y}_i = \widehat{\boldsymbol{\theta}}^{(-i)}\boldsymbol{Z}_i$.
6. Find the Mahalanobis Distance by $D_i = (y_i - \hat{y}_i)'(\widehat{\boldsymbol{\Sigma}})^{-1}(y_i{}^i - \hat{y}_i)$.
7. Repeat steps 3-6 for i=1,..., N.

### 6.2.3 RESULTS

In this section, we discuss the application and results of Cook's and Mahalanobis distance for the case study. Figures 6.1 and 6.2 have the values of these distances. We observe that most samples do not seem to be significantly influential with exception to the samples from trucks 68 and 70. The distances values in both methods are larger than distances values from the other trucks. This implies that these samples are farther away from the other samples and they might be categorized as outliers.

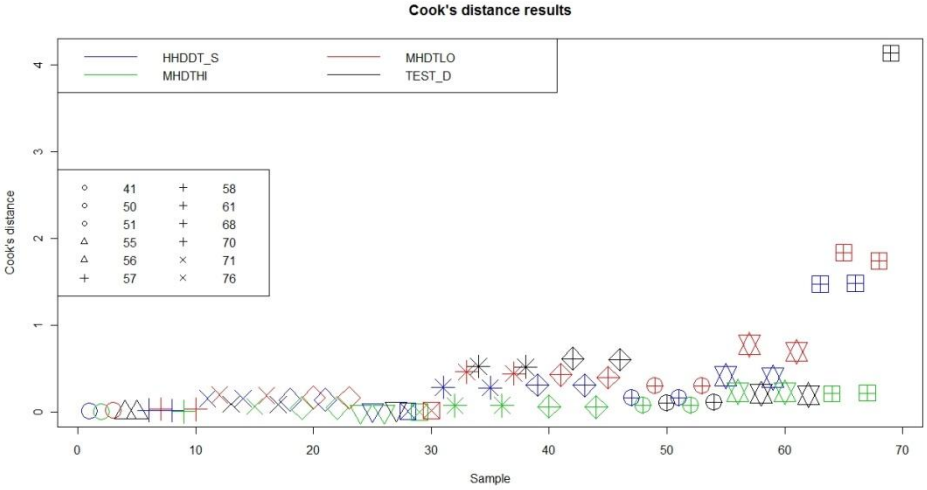



**Figure 6.1.** Results for Cook's Distance grouped by vehicle and velocity pattern

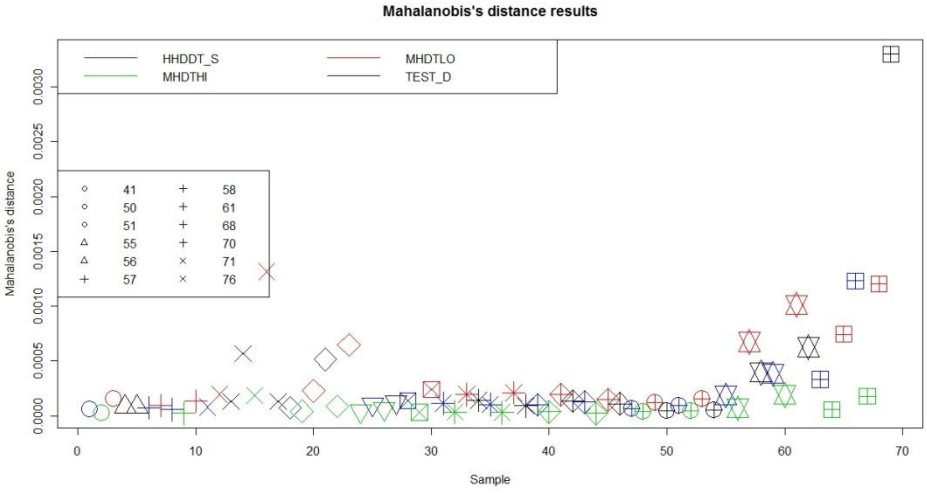
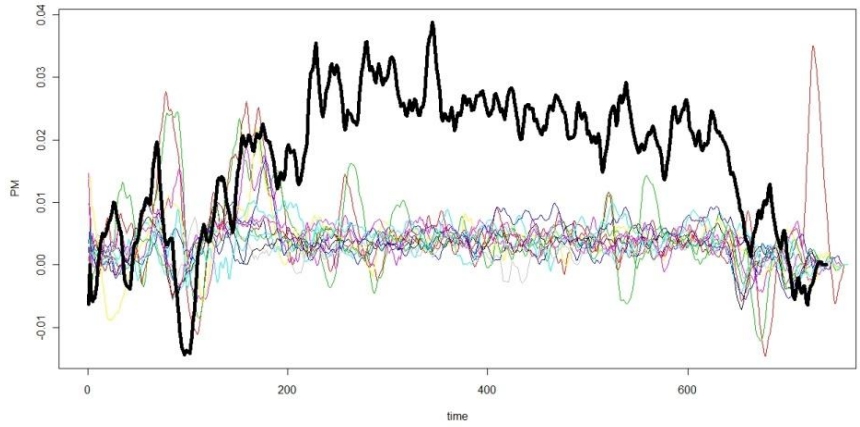
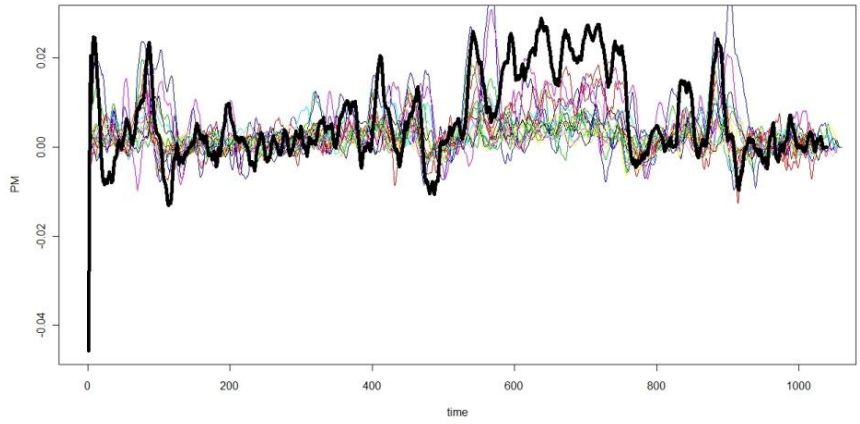



**Figure 6.2.** Results for Mahalanobis Distance grouped by vehicle and velocity pattern

Figure 6.3 has two of the curves with significant high Mahalanobis and Cook's distance values. They are also compared with the other curves that resulted from the same type of experiment as them.




a)




b)

**Figure 6.3.** Curves with the highest Mahalanobis and Cook's distance

Plot (a) of Figure 6.3 shows the curves resulted from the HHDDT_S velocity pattern. We can observe that the curve with highest Cook's and Mahalanobis distance (highlighted in black) is out of the ordinary and does not follow the same behavior as the other curves. Given these results, we infer that this case does not belong to this set of samples and thus need to be deleted. In the other hand, we can see in plot (b) of Figure 6.3 the curves resulted from the MHDTHI velocity pattern. The curve

highlighted is one with larger Cook's and Mahalanobis distance value than other samples but with smaller value than the curve discussed previously. In fact, this curve was mentioned in 3.2.2 since it seems to not follow closely the general pattern of the other curves. However, we cannot be certain that this is an outlier.

# CHAPTER 7

# CONCLUSION

We developed a functional linear regression model for fitting almost continuously data in which the responses have been distorted during the time recording by other factors in the environment or/and measure process. This implies that the recordings do not represent the instantaneous effects of the independent variables but combination of responses happening at different times. We argue that the distortion or alteration of the responses is well modeled by the convolution of the stimuli $\xi_{iP}(w)$ with length $\delta_p$ for$p = 1, \dots, P$. The relation between the recording $y_i(t)$ and stimulus $\xi_{iP}(w)$ is continuous and defined as $\beta_p(w)$ which is the coefficient of the convolution. This coefficient function gives a weight to the values of $\xi_{iP}(w)$ for the estimation of $y_i(t)$ as$w \to t$. In addition, diagnostic methods were also developed with the purpose of estimating the variability of $\beta_p(w)$ and the effects that the individual data record have in the model estimation. Specifically, we developed a block bootstrapped method that handles the discontinuity caused by the division in blocks and their resampling. The statistical methods of Cook's and Mahalanobis distance were also modified with the objective of accounting for the time structure and serial correlation of the data and to identify curve or functional data that causes a significant effect in the solution and estimates of the model.

We presented the results for the statistical model for the prediction of the continuous trajectory of the particulate matter given driving behavior. The model uses the data collected from a program called E-55/59 which had as an objective to improve the

emissions inventory in California. Given the data from medium heavy-duty trucks, we were able the following:

- The length of dependence between the particulate matter with the velocity and acceleration
- The coefficient functions for both covariates
- The confidence interval by the delta method for both covariates
- The trajectories of the particular matter gathered from different vehicles and given different velocity patterns
- Distance values that allows us to find curves that influenced the estimate of the model

However, we experienced difficulty when we applied the modification of the block bootstrapping method to the residuals to find the variability of this emission model. As a result, we were not able to find confidence interval for the coefficient function of velocity and acceleration by using the full data. Instead, we chose one vehicle to demonstrate the applicability of the modification of block bootstrapping. In this case, we were able to estimate confidence interval for the coefficient function of the velocity and acceleration for this specific vehicle. The plots of these coefficient functions are in Figure 5.6 and 5.7 in Chapter 5. Given this, we observe that the pattern of the coefficient functions differ to the ones found by using the full data. These plots are in Figure 4.5 and 4.6.

We believe that this problem is caused by the variability of the data and other factors that have not been accounted by the model.

As mentioned before, the data used for the development of this emission model comes from eleven medium heavy-duty trucks. We notice that there is some variety among the trucks even though they pertain to the medium heavy-duty group. Table 7.1 has the general physical characteristics of the trucks used for this particular study. We observe that the range of engine year is from1988 to 2001 and vehicle weight is from 25000 to 33000. We are interested in adding this variability to the model as a random effect.

**Table 7.1.** General characteristics of the medium heavy-duty trucks

| ID | VEH YEAR | VEH_MAN | VEH GVW | ENG MAN | ENG YEAR |
|---|---|---|---|---|---|
| 50 | 2001 | INTERNATIONAL | 26000 | INTERNATIONAL | 2001 |
| 51 | 1994 | INTERNATIONAL | 29000 | INTERNATIONAL | 1994 |
| 55 | 1992 | FORD | 31000 | FORD | 1991 |
| 56 | 1988 | FORD | 33000 | CATERPILLAR | 1988 |
| 57 | 2000 | FREIGHTLINER | 26000 | CATERPILLAR | 1999 |
| 58 | 1982 | FORD | 25000 | DETROIT | 1999 |
| 61 | 2000 | GMC | 25950 | CATERPILLAR | 1999 |
| 68 | 1995 | INTERNATIONAL | 33000 | INTERNATIONAL | 1995 |
| 70 | 1998 | FREIGHTLINER | 26000 | CUMMINS | 1997 |
| 71 | 1995 | FORD | 33000 | CUMMINS | 1994 |
| 76 | 1993 | FORD | 33000 | FORD | 1993 |

Furthermore, Table 7.2 has the general results of the chassis dynamometer experiments. This table has the identification of the truck, the velocity pattern and the load weight (test weight) applied to a specific truck, and the duration of the experiment.

**Table 7.2.** General results of the driving cycles

| ID | TEST CYCLE | TEST WEIGHT | TOTAL TIME |
|---|---|---|---|
| 50 | HHDDT_S | 14000 | 739.9 |
| 51 | TEST_D | 14500 | 1060 |
| 51 | MHDTHI | 14500 | 1190 |
| 55 | TEST_D | 15500 | 1060 |
| 55 | MHDTHI | 15500 | 1190 |
| 55 | HHDDT_S | 15500 | 760 |
| 70 | TEST_D | 13000 | 1039.9 |
| 70 | MHDTLO | 13000 | 348 |
| 70 | MHDTHI | 13000 | 1168 |
| 70 | HHDDT_S | 13000 | 738 |

It is important to notice that there is some variability in the way the experiments were applied and measured. In particular, we observe that

- Some cycles were not applied to all trucks.
- The time length among similar cycles is not the same.
- Two different weight loads were applied to the trucks.

The variability generated by the first two points is not of relevance since the convolution model accounts for this naturally by assuming that the data is stationary. However, we plan to consider the weight load as fixed effect in future work.

Moreover, Figure 7.1 shows the autocovariance of two vehicle residuals. As discussed in previous chapter, we can see that they have similar pattern. In particular, we can see that they get close to zero after lag 20. However, we can see in the y-axis that the

values of the autocovariance between these two vehicles differ. Similar behavior happens to the other vehicles. This implies that we have to account for a heterogeneous variance among all the medium heavy-duty trucks of this case study.

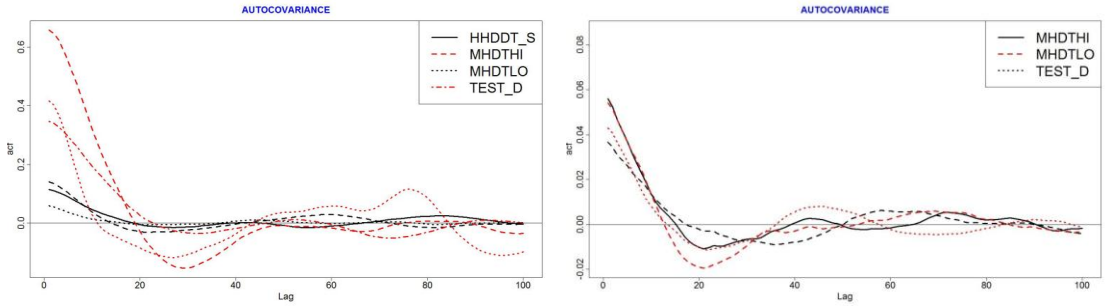



**Figure 7.1.** Autocovariance of the samples of two different vehicles

For further work, we plan to estimate a model such as the following expression:

$$PM_{ij}(t_i) = \alpha + \int_{t_{ij}-\delta_{vel}}^{t_{ij}} \beta_{vel}(w)\xi_{ijvel}(w)dw + \int_{t_{ij}-\delta_{accel}}^{t_{ij}} \beta_{accel}(w)\xi_{ijaccel}(w)dw$$

$$+weight_{ij} + \gamma_j + \varepsilon_{ij}(t)$$

$$\varepsilon_j(t) \sim N(0, \sigma_j^2(|t-s|))$$

$$\gamma_j \sim N(0, \sigma_\gamma^2)$$

where $i = 1, \ldots, 4$ (number of the driving cycles) and $j = 1, \ldots, 11$ (number of the trucks) and

$PM_{ij}(t_i)$ is the response or dependent variable

$\xi_{ijvel}(w)$ and $\xi_{ijaccel}(w)$ are the stimuli

$\beta_{vel}(w)$ and $\beta_{accel}(w)$ are the coefficient of the convolution

$\delta_{vel}$ and $\delta_{accel}$ are the length of the convolution

$weight_{ij}$ is the fixed effect of the weight load

$\gamma_j$ is the random effect of the $j^{th}$ vehicle

$\varepsilon_{ij}(t)$ is the random error

$\sigma_j^2$ is the variance given by a particular set of data

$\sigma_\gamma^2$ the variance of vehicle random effect

We will consider this and other alternatives of this model with the objective to find one that takes care of the heterogeneity of the variance and behavior of the coefficient of the convolution. That is, we plan to introduce an interaction term between the weight factor and the covariates $\xi_{ijvel}(w)$ and $\xi_{ijaccel}(w)$ to see how the change in driving behavior variables and weight load impacts the production of particulate matter. We also will look at different estimates of the variance $\sigma_j^2(|t-s|)$ by considering either the samples of one specific vehicle or the different driving cycles. Furthermore, we will explore the effect of introducing a random effect in the variance. We use cross-validation procedure similar to the one discussed in chapter 2 to estimate this random. In conclusion, we expect to find a general mean model that provides informative and valid estimates for this case study.

# REFERENCES


1. Cuevas, A.; Febrero, M.; Fraiman, R. (2002) Linear functional regression: The case of fixed design and Functional Response. *The Canadian Journal of Statistics* **30,** 285-300
2. Malfait, N., Ramsay, J. O. (2003) The Historical Functional Linear Model. *The Canadian Journal of Statistics* **31,** 115-128
3. Ajtay, D., Weilenmann, M., Soltic, P. (2005) Towards accurate instantaneous emission models. *Atmospheric Environment* **39**, 2443–2449.
4. Ramsay, J. O., Silverman, B. W. (2006) Functional Data Analysis. Second edition. New York: Springer.
5. Capiello, A., Chabini, I., Nam, E., Lue, A., Zeid, M. (2002) A statistical model of vehicle emissions and fuel consumption. *Ford-MIT Alliance*
6. Ahn, K., Rakha, H., Trani, A., Aerde, M. (2002) Estimating vehicle fuel consumption and emissions base on instantaneous speed and acceleration levels. *Journal of Transportation Engineering* **128**, 182-190
7. Weilenmann, M., Soltic, P., Ajtay, D. (2003) Describing and compensating gas transport dynamics for accurate instantaneous emission measurement. *Atmospheric Environment* **37**, 5137–5145.
8. Laden, F, Schwartz J, Dockery DW, Neas LM. (2000) Association of Fine Particulate Matter from Difference Sources with Daily Mortality in Six U. S. *Environmental Health Perspectives* **108**, 941–947.
9. McCreanor, J., Cullinam, P., Nieuwenhuijsen, M. J. (2007) Respiratory Effects of Exposure to Diesel Traffic in Persons with Asthma. *The New England Journal of Medicine* **357**, 2348-2358.

10. Janssen, N, Mansom, D. F, Jagt, K, Harssema, H, and Hoek, G. (1997) Mass Concentration and Elemental Composition of Airborne Particulate Matter at Street and Background Locations. *Atmospheric Environment* **31**, 1185-1193.
11. Smit, R., Poelman, M., Schrijver, J. (2008) Improved road traffic emission inventories by adding mean speed distributions. *Atmospheric Environment* **42**, 916-926.
12. Schwartz J and Neas LM. (2000) Fine Particles Are More Strongly Associated Than Coarse Particles with Acute Respiratory Health Effects in Schoolchildren. *Epidemiology* **11**, 6-10.
13. Schwartz, J. (2000) Assessing Confounding, Effect Modification, and Thresholds in the Association between Ambient Particles and Daily Deaths. *Environmental Health Perspectives*, **108** 563-568.
14. Schwartz, J., Laden, F., Zanobetti, A. (2002) The Concentration-Response Relation between PM 2.5 and Daily Deaths. *Environmental Health Perspectives* **110**, 1025-1029.
15. Lahiri, S. (2003) Resampling Methods for Dependent Data. New York: Springer
16. Efron, B. (1979) Another Look at the Jackknife. *The Annals of Statistics* **7**, 1-26.
17. Efron, B. (1985) Confidence Intervals for a Class of Parametric Problems. *Biometrika* **72**, 45-58.
18. Carlstein, E. (1986) The use of subseries methods for estimating the variance of general statistic from a stationary time series. *The Annals of Statistics* **14**, 1171-1179.
19. Cook, D. R. (1982) Residuals and Influence in Regression. New York: Chapman and Hall

20. Christensen, R. (2002) Plane Answers to Complex Questions: The Theory of Linear Models. Third Edition. New York: Springer.
21. Weisberg, S. (1985) Applied Linear Regression. Second Edition. New York: Wiley and Sons.
22. Pankratz, A. (1991) Forecasting with Dynamic Regression Models. New York: Wiley and Sons.
23. Stulajter, F. (2002) Predictions in Time Series Using Regression Models. New York: Springer.
24. Belsley, D., Kuh, E., Welsh, R. E. (1980) Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. New York: Wiley and Sons
25. Kedem, B. , Fokianos, K. (2002) Regression Models for Time Series Analysis. New Jersey: Wileysand Sons
26. Ramsay, J. O., Silveman, B. W. (2002) Applied Functional Data Analysis: Methods and Case Studies. New York: Springer
27. Ramsay, J. O., Hooker, G., Graves, S. (2009) Functional Data Analysis with R and MATLAB New York: Springer