

# Project 2: Forecasting Highway Car Volumes

Chen Nan

Due date: November 17, 2019

## 1 Introduction

This project studies traffic volumes on an express way connecting two metropolitan cities. A measuring station is set up to record the number of cars passing the station (one direction). The data is organized as an hourly data, which consists of the traffic volumes, as well as weather information. The description of the variables are summarized below:

1. **IsHoliday**: a categorical variable to indicate whether it is a public holiday or not
2. **Temp**: a numeric variable, average temperature during the hour in Kelvin
3. **Rain1h**: a numeric variable, amount in mm of rain that occurred in the hour
4. **Snow1h**: a numeric variable, amount in mm of snow that occurred in the hour
5. **CloudsAll**: a numeric variable, percentage of cloud covering the sky
6. **WeatherMain**: a categorical variable, short textual description of the current weather
7. **WeatherDescription**: a categorical variable, longer textual description of the current weather
8. **Time**: DateTime Hour of the data collected in local time
9. **TrafficVolume**: the numerical **response** variable, hourly reported westbound traffic volume

In the datafile “**P2train.csv**”, 40,000 hourly readings are included. Each row includes information for one hour. The testing dataset “**P2test.csv**” has the same data structure. However, some response columns are “missing”, and you need to forecast the hourly traffic volumes based on information before each respective row. Note that not all rows in “**P2test.csv**” have missing response variables. The row indices (starting from 1) with missing response values are saved in “**P2test\_index.csv**”. You can use provided responses in the testing file to make predictions. The accuracies of your prediction will be evaluated based on the numbers you provided.

## 2 Project Task

### 2.1 Regression on Time

In this task, you need to build a regression model only using the variables `Time` and `TrafficVolume` for prediction. You can consider what are the necessary components to include in the “regression on time” model, and make diagnostic check and model interpretation.

### 2.2 Exponential Smoothing

In this task, you can apply appropriate exponential smoothing to make forecasting. Please explore and verify your choice of model, and elaborate potential pros and cons of the model.

### 2.3 Free form forecasting

In this task, you can select or build your own model to predict the traffic volumes. You can only use methods discussed in the module (regression, exponential smoothing, Box-Jenkins methodologies, etc). Elaborate your choice and potential room for improvement. For this task, you need to submit your forecasting results for accuracy evaluation.

## 3 Submission

1. A report not more than 10 pages with 1.5 spacing (soft copies and hard copies), which documents the methods using, main findings, and interpretations. Codes and software printouts should NOT be included in the report.
2. Complete codes used for the analysis, with reasonable details of comments in Jupyter Notebook (Soft copies only). **Attention:** Make sure your results are replicable by the codes you submitted. Unreproducible results are considered cheating/plagiarism.
3. Forecasting results on the test dataset in a “csv” file with a single column, as shown in the following example

```
A0001124H
10.31
8.5
20.1
...
11.5
```

Only the Task 3 require forecasting results submission.