CE3020- Transportation Engineering-II
Term Project Report

Date- 13th May, 2013

A. Problem Statement-
Data mining and analysis

1. Learning to use big data sets that are both clean and unclean
2. Fitting appropriate statistical distributions and inferential analysis
3. Develop suitable stream models and check the validity of fundamental traffic flow relation
4. Compare results with other studies
5. Show one application using this data (Example: Incident detection)

B. The Data-set:

The given data-set consisted of 7 day traffic flow data on the IT-corridor near Indira Nagar. The traffic stream video was processed by an image-processing software canned Trazer which gave vehicle coordinates with time. The feed was processed to give the following traffic parameters for 1-minute intervals-

1. Count
2. Speed
3. Occupancy

Each of these parameters was available for 4 different types of vehicles- LMVs, HMVs, 3-wheelers and 2-wheelers.
Trazer collecting data by image processing of a traffic stream can be seen here.
The data set provided can be found here.

C. Data Munging:

In order to make the data-set usable, a lot of munging had to be performed. This is often the most tedious and time consuming step of any data analysis. Some of the processes involved were as follows-

1. Each variable should form a column, each observation should form a row
2. Providing meaningful names
3. Sub-setting to the required number of fields
4. Removal/completion of the incomplete records. In this case as there were sufficient number of observations, the incomplete observations were removed. Otherwise, they can also be replaced by the local or global averages.
5. Detection and removal of outliers. This was accomplished by simple sub-setting physically impossible values of parameters such as a speed greater than 200kmph, occupancy greater than 1 etc. This is shown in figure 1 below.

As this is a pro-active step, no fixed methodology is stipulated for it. It largely depends on the quality of the data-set and the goals of the data-analysis.

Apart from this, an important aspect is the addition of transformed variables that to the data-set. These variables are generally derived from the variables existing in the data-set.

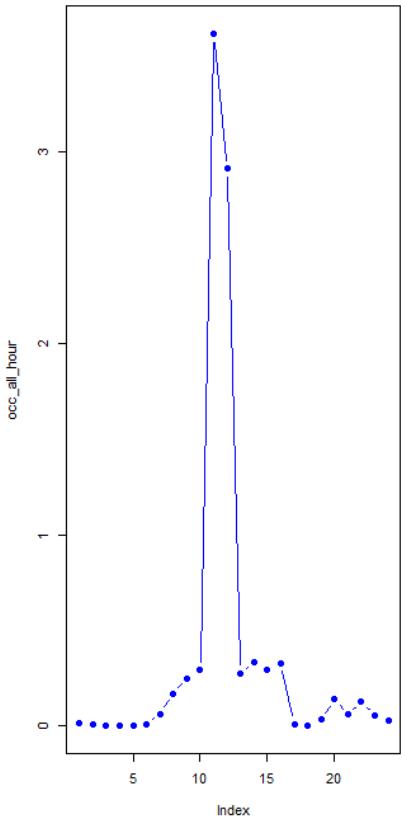


Fig. 1: An Occupancy outlier

Examples of Addition of transformed variables:

1. The occupancy observations given in the data-set were not based on the standard definition of occupancy- the fraction of time that a particular stretch of road is occupied in that particular interval. It was given as the total time a stretch was occupied in a 1 minute interval. This transformation was hence necessary.
2. Density is an important parameter in the modelling of any traffic flow. The relation between occupancy and density (assuming vehicle size is uniform) is-
$$o = (L_v + L_d) \times k$$
Here, L_v=Length of the vehicle, and L_d=Length of detector.
The average length of a passenger car has been taken to be 4m. The length of the section over which the occupancy was measured was 1.5m for the week of observation (note that this changes from time to time).

3. Passenger car equivalent: For the modelling of the different types of vehicles together, it was necessary to convert the vehicle count of each to the Passenger Car Equivalent (PCE). This was done as per the provisions in IRC106 shown in figure 3.

TABLE 1. OBSERVED VEHICULAR DIMENSIONS

Vehicle type (1)	Overall dimension (m)	
	Length (2)	Breadth (3)
Buses	10.3	2.5
Trucks	7.5	2.5
LCV	5.0	1.9
Cars	4.0	1.6
M.Th.W.	2.6	1.4
M.T.W.	1.8	0.6
Bicycles	1.9	0.5
Tricycles	2.5	1.3

Fig. 2: Standard vehicle dimensions

TABLE 1. RECOMMENDED PCU FACTORS FOR VARIOUS TYPES OF VEHICLES ON URBAN ROADS

Vehicle Type	Equivalent PCU Factors	
	Percentage composition of Vehicle type in traffic stream	
	5%	10% and above
Fast Vehicles		
1. Two wheelers Motor cycle or scooter etc.	0.5	0.75
2. Passenger car, pick-up van	1.0	1.0
3. Auto-rickshaw	1.2	2.0
4. Light commercial vehicle	1.4	2.0
5. Truck or Bus	2.2	3.7
6. Agricultural Tractor Trailer	4.0	5.0
Slow Vehicles		
7. Cycle	0.4	0.5
8. Cycle rickshaw	1.5	2.0
9. Tonga (Horse drawn vehicle)	1.5	2.0
10. Hand cart	2.0	3.0

Fig. 3: Recommended PCU factors for different types of vehicles on urban roads

In addition to the above the following logistics also had to be considered-

1. **Software sensitivity:**

Trazer as an image processing software yields reliable results only in the time periods with sufficient daylight.
The time period of observation for the given dataset is from 8th Nov. 2012 to 14th Nov. 2012. During this time, the average sunrise and sunset times in Chennai were 6:05am and 5:41pm. Hence, the dataset had to be subset from 6am to 6pm.

2. **Aggregation to an interval with consistent values of the flow parameters- Conversion from 1 minute intervals to 15 minute intervals-**
   One-minute intervals are too short to yield consistent values of flow parameters. This high variability may lead to inaccuracies in the interpretations drawn from them. Hence we need to combine these discrete 1 minute intervals into larger continuous intervals.
   Choosing the length of the larger interval has been considered to be an optimization of two factors:
   i. The extent to which it can be considered to be representative of the actual flow conditions such as count, occupancy, speed and density.
   ii. The number of observation that it results in.
   Based on this, fifteen minute intervals were chosen to model the flow.

3. **Power-cuts in Chennai:**
   The interval under consideration, 6 am to 6 pm, each day, is randomly interspersed with intervals with power failures. During these intervals, the video streaming terminates. These intervals were identified and removed as shown in figure 4.

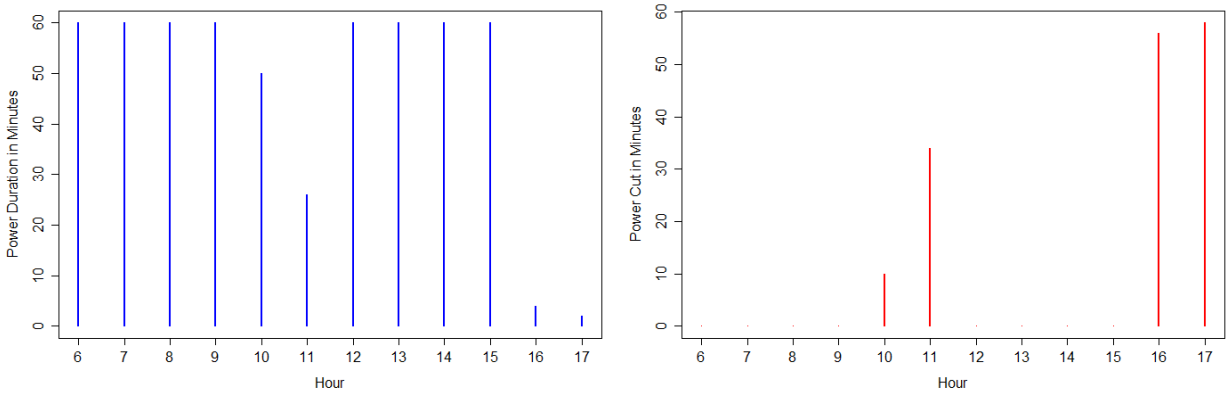


Fig. 4: Power supply/cut diagram showing details for each hour under consideration on 14th November, 2012 for the given data-set

Hence, the data was subset considering the following two factors:

1. The daylight hours (6am to 6pm)
2. The power-cuts (randomly interspersed)

After accounting for these 2 factors, we were left with 283- 15 minute intervals, which amounts to 70 hours and 45 minutes for the week.

Application Development:

The data-munging process can be very tedious. The process has been developed into an R-based application and is available free of cost here.
The application can be seen in action here.

D. Validating the fundamental Traffic Flow equation:
The values for flow observed in the field was compared with the values obtained using the fundamental Traffic Flow Equation-

$$q = ku$$

The results for the same are shown in figure 6.
From this, it can be seen that the trends are similar but the deviations are significantly high. The variation can be explained by-

1. Uncertainty in the length of the sensing area on the road (it is *approximately* 1.5m)
2. The space-mean speed is better modeled as the harmonic mean rather than the arithmetic mean as it has been done here.

The process of verifying the relationship can be improved by incorporating the changes listed above.

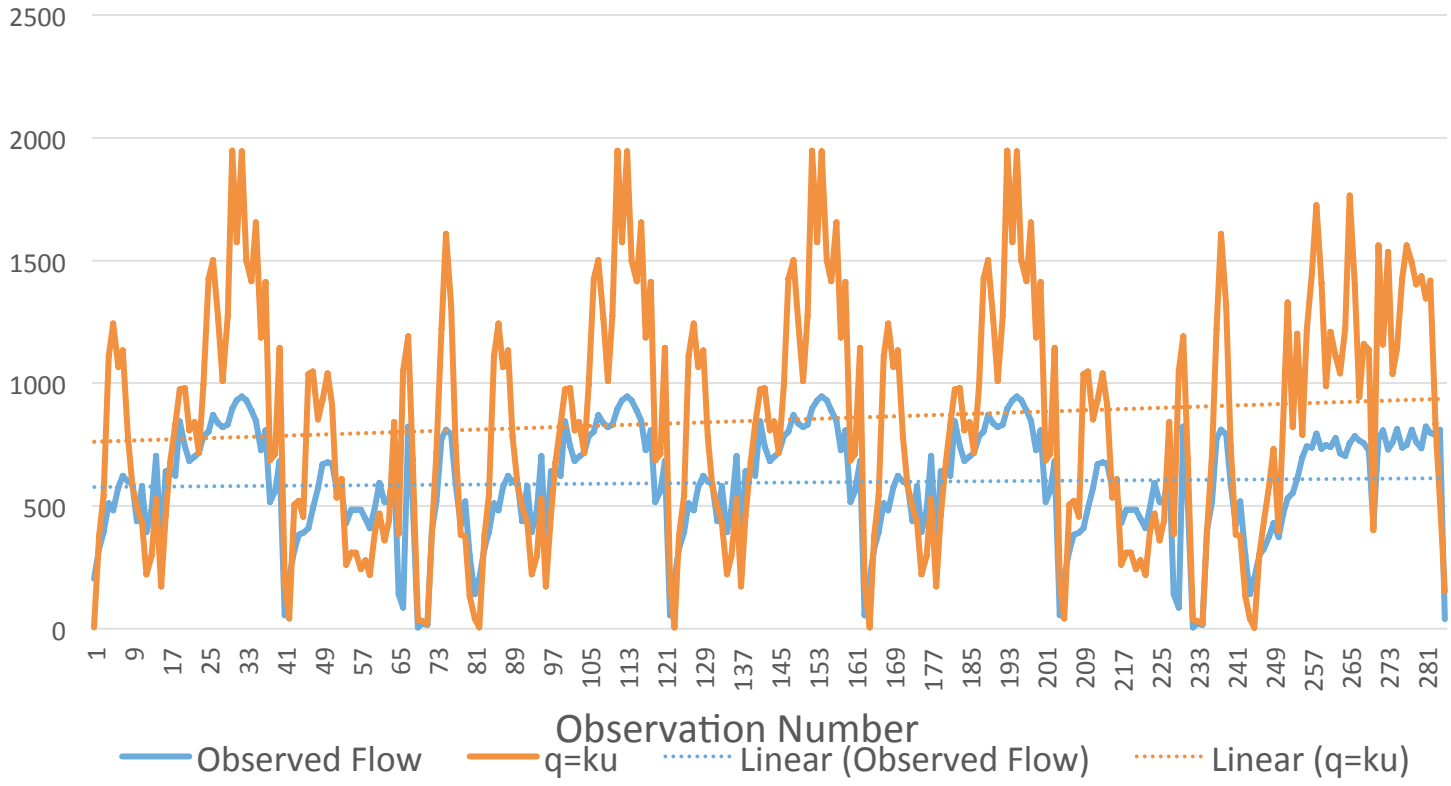


Fig. 6: Verification of the Fundamental Traffic Flow Equation

E. Development of stream flow models:
From the derived values of the density and speed for the 15-minute intervals, it is possible to generate an appropriate stream flow model.

Firstly we attempt to generate a fitting model for the derived data-set and then we will compare it to the existing models namely Greenshield, Greenberg and Underwood's models

From the plot of the data-points, it was observed that a multi-regime or a polynomial fit would be the most appropriate for the given data-set.
A quadratic model for the data-set is shown in figure 7.

From the fitted curve, it can be seen that this is a reasonable fit for the given data-set.

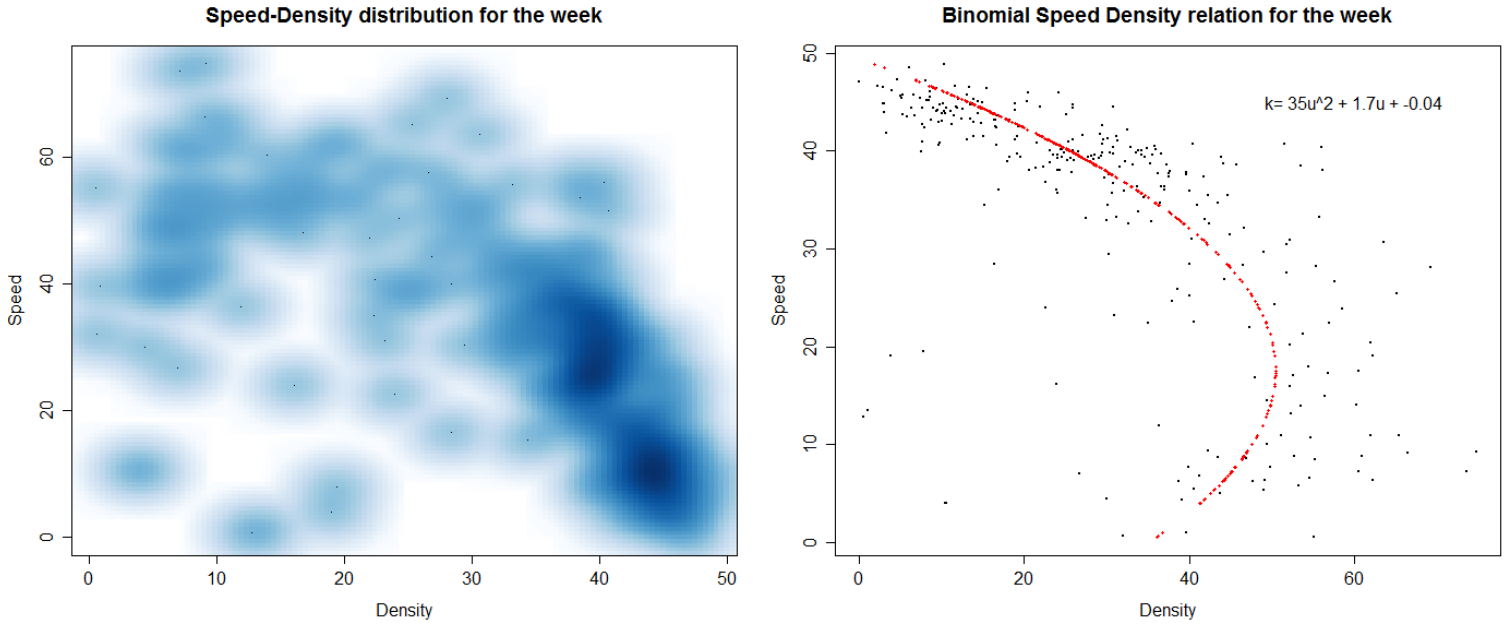
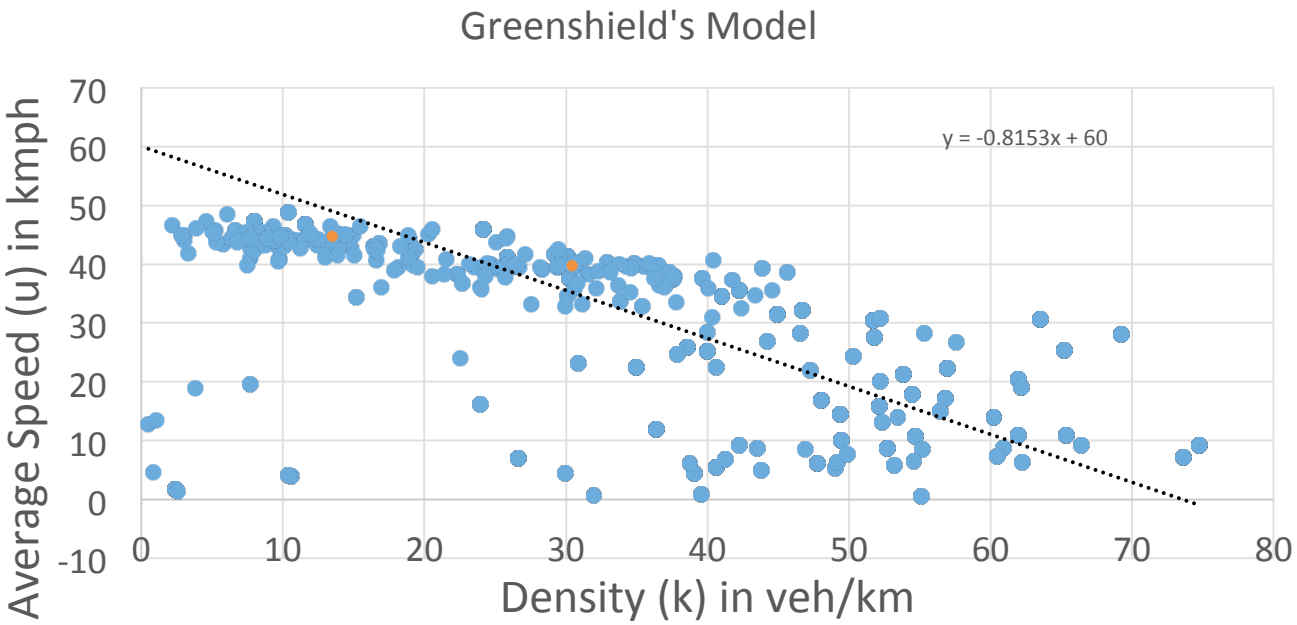


Fig. 7: Fitting a quadratic polynomial through the given data-set

The speed and the density values were then output to a file and the fitting of the other flow-models as done in MS-Excel is shown in figures 8 through 10.



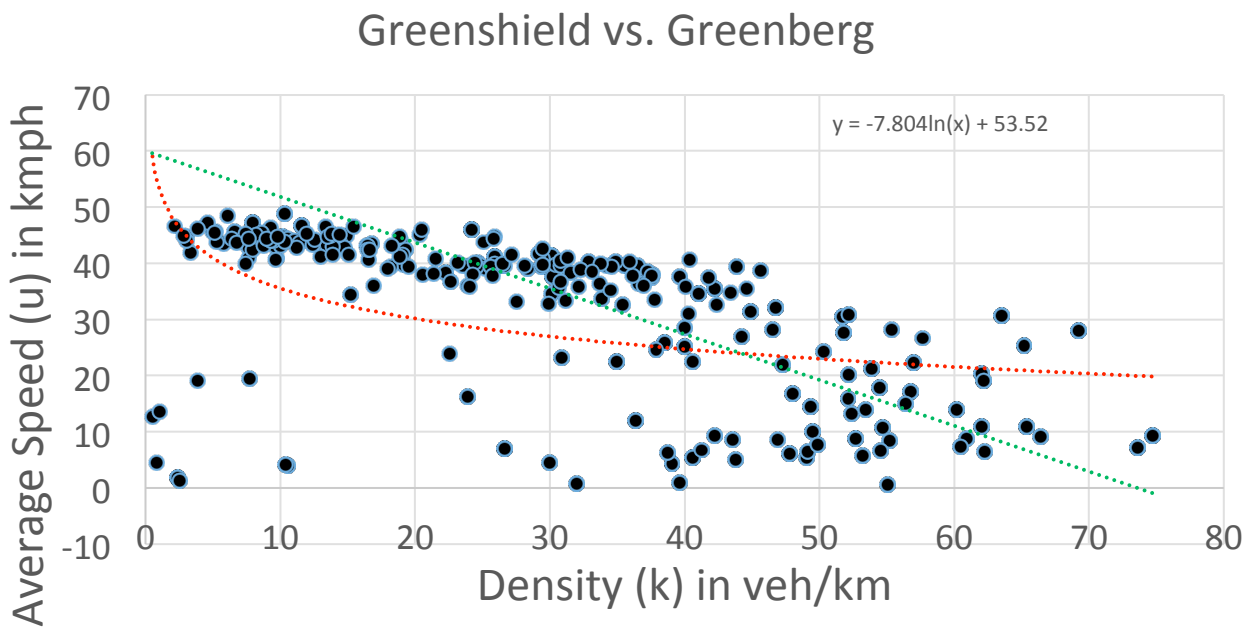


Fig. 8: Greenshield's Model applied to the given data-set

Fig. 9: Greenberg's Model applied to the given data-set and comparison to the Greenshield's model

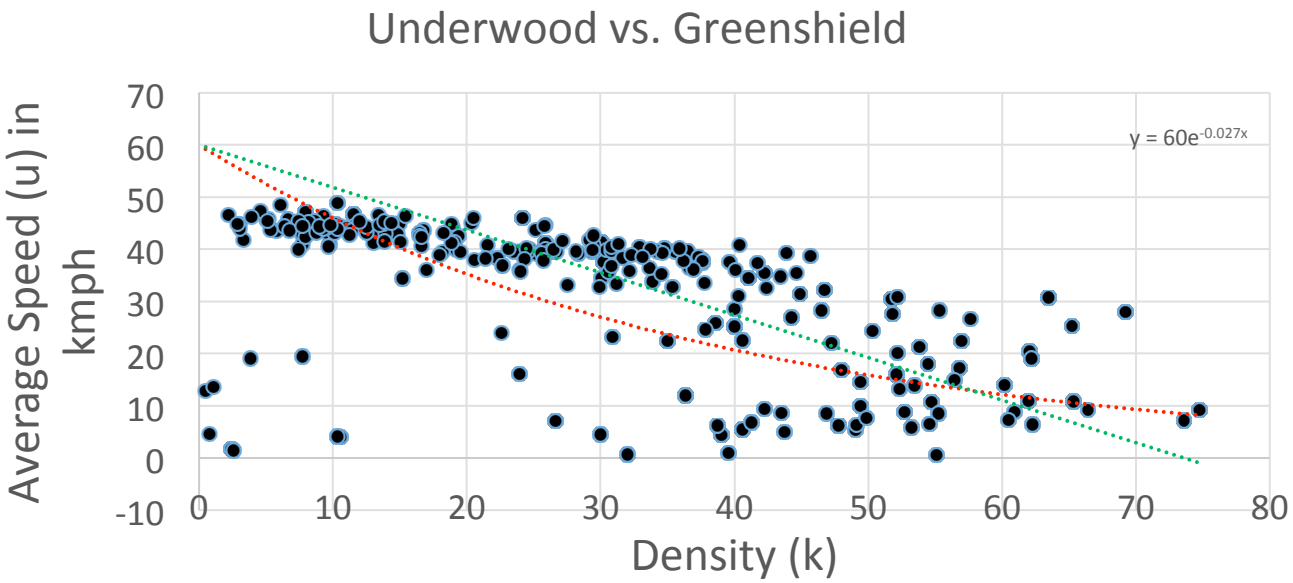


Fig. 10: Underwood's Model applied to the given data-set and comparison to the Greenshield's model

From the above graphs, it can be seen that a polynomial model or a multi-regime model would be the best fits for the given data-sets. This can be verified by defining a common cost function for all the models and selecting the model with the least cost. Note that higher order polynomials will definitely have lower costs but suffer from the hazard of over-fitting.

F. Applications:

1. Incident detection: Changes in the traffic flow when incidents such as accidents occur can be modeled and these can be used to respond more proactively to such incidents. Figure 11, though not an incident in itself shows how incident detection can be applied. Further, Machine Learning algorithms can be utilized for complete automation of the incident detection system.

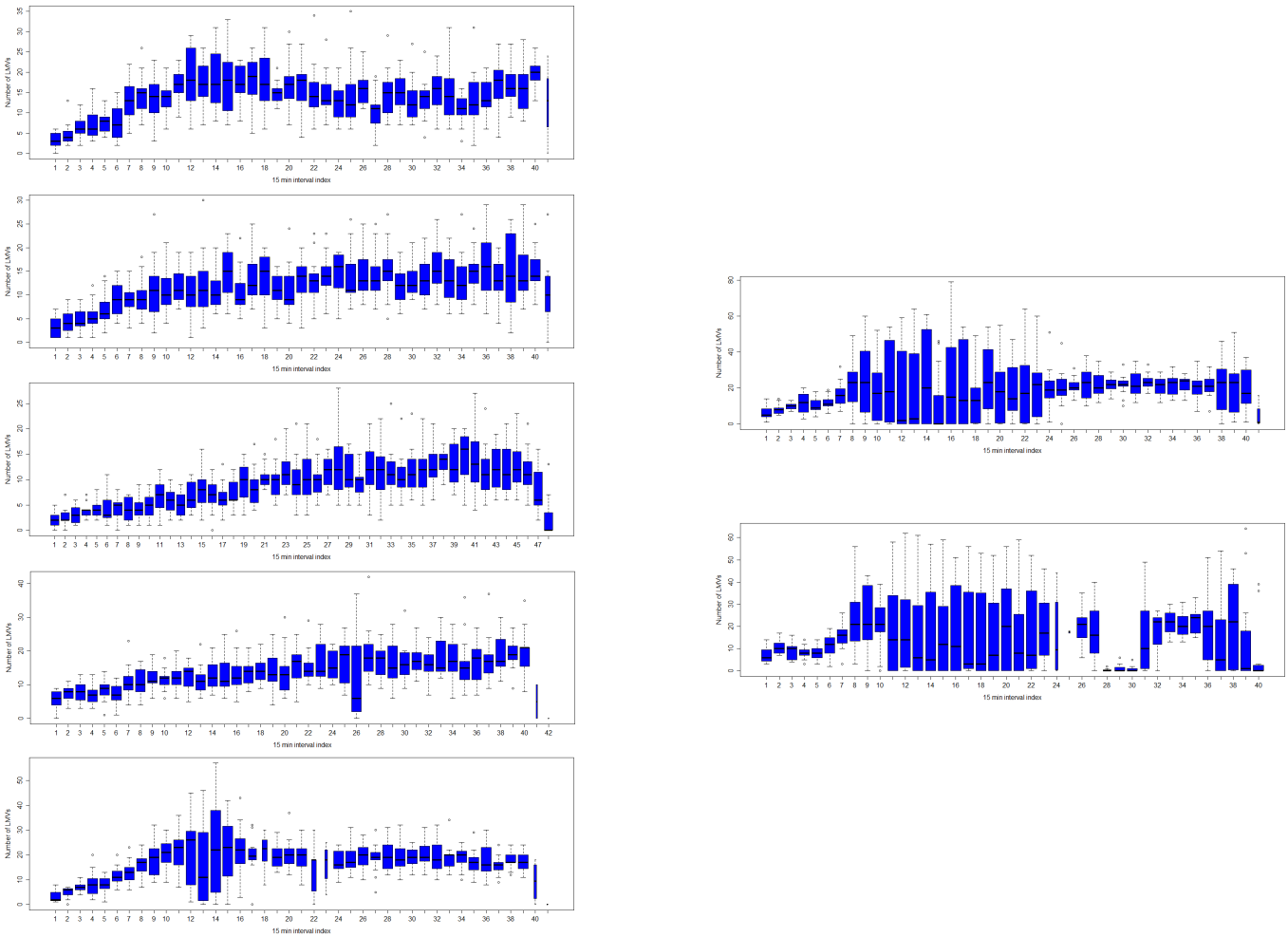


Fig. 11- The variation in the vehicle count over the interval of a day for each of the seven days under consideration. Two of the days show distinctly dis-similar plots as compared to the remaining 5 days.

2. Studying traffic flow change between the weekends and the weekdays. Roads are typically designed for the peak flows which usually occurs on weekdays. This can be used to develop a better utilization plan for roads for the weekends. This can be seen in figure 12.

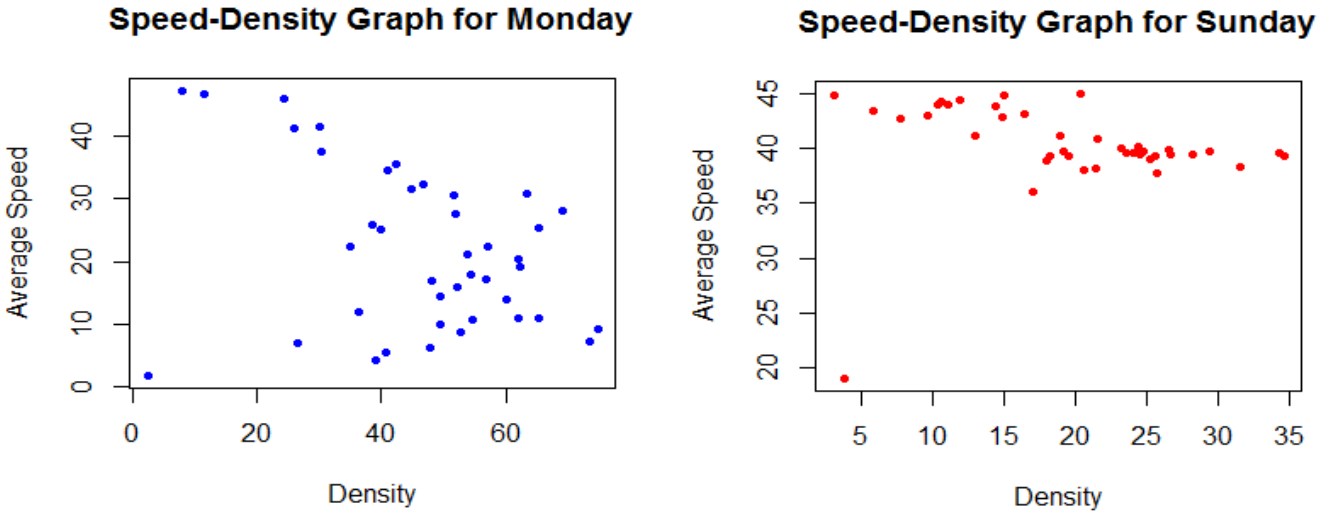


Fig. 12: Speed-Density variation between the weekends and weekdays

G. Conclusion:

The given data-set has been analyzed for the required parameters.
The entire project is reproducible and hence can be improved as and when required.

H. Resources-

1. Paper No. 542: Study of the effect of traffic volume and road width on PCU value of vehicles using microscopic simulation, by V. Thamiz Arasan and K. Krishnamurthy.
2. IRC 106
3. www.timeanddate.com

Special Thanks: Dr. Lelitha Devi and Ramesh (ITS Lab, IIT-M)

The presentation used in class is available here.