NDOT Research Report

Report No. 295-10-803

I-15 North Project

November 2013

Nevada Department of Transportation
1263 South Stewart Street
Carson City, NV 89712

Disclaimer

This work was sponsored by the Nevada Department of Transportation. The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the data presented herein. The contents do not necessarily reflect the official views or policies of the State of Nevada at the time of publication. This report does not constitute a standard, specification, or regulation.

UNLV
UNIVERSITY OF NEVADA LAS VEGAS

Transportation Research Center

I-15 North Project

Author:

Himanshu Verma

vermah@unlv.nevada.edu, 702-267-8849

Student M.S. Electrical Engineering, and M.S. Mathematics

University of Nevada, Las Vegas

Principal Investigator:

Dr. Pushkin Kachroo, P.E.

http://faculty.unlv.edu/pushkin, pushkin@unlv.edu, 540-588-3142

Director, Transportation Research Center

Harry Reid Center for Environmental Studies

University of Nevada, Las Vegas

Prepared for:

Nevada Department of Transportation

1263 South Stewart Street, Carson City, NV 89712

November 5, 2013

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

GLOSSARY

Term	Definition
Sentiment Analysis	Extraction of sentiment from raw text data.
Content Similarity Measure	Measuring the similarity between two text data in semantic manner.
Concept Discovery or Context Recognition	Finding the concept/context in which the text data is about.
Word Sense Disambiguation	Determining the right sense of a word in text data.
Synset	A set of synonyms.
Semantic Network	A network of words in which they are arranged in order of their meaning not alphabetically.
Metric	It is a set of mathematical rules to find numerical distance between two concepts.
Corpus	Raw text database is generally known as corpus.
POS Tagging	Assigning part of speech to each word in a text.
Tokenize	Dividing sentence into words.
Function Words	Non informative words like prepositions.
Content Words	These are words which add information to the sentence.
Hamming Distance	It is the between two strings of equal length is the number of positions at which the corresponding symbols are different
Eucledian Distance	It is the distance between two points that one would measure with a ruler, and is given by the Pythagorean formula.
Normalization	It is a method used to standardize the range of independent variables.
Inner Product Space	It is a vector space with an additional structure of a scalar value known as inner product and defined over each pair of vectors.

EXECUTIVE SUMMARY

Thousands of people die every year in United States due to unfortunate accidents and crashes. This is one of the major concern which needs to be addressed by any means. Some of the facts about death tolls due to crashes are as follows:-

1. In 2007, 47.5 percent of people belonging to age group between 1-45 died by moving traffic[1].
2. In 2009, more than 2.3 million people were treated in emergency departments due to injuries caused by motor vehicle crashes[2].
3. In 2005, 70 billion of dollars was the added cost associated to deaths and injuries caused by crashes during transportation[2].

Like most of the other fields technology comes as the savior for this purpose also. And the way we can handle such situations are by making our transportation system intelligent and smarter. This gave birth to the current hot topic called Intelligent Transportation System (ITS). A lot of research is going on in several top-notch universities for this purpose.

These incidents are not only responsible for the death tolls and injuries but they also produce added burden on the economy, mental traums, added pain and high insurance premiums which is hard to afford for people. That is the reason why it

becomes a major concern to address. To improve the situation transportation department use lot of funds to aware people about use of safety belts and campaigning about obeying the traffic rules. Things improve with such campaigns a lot but there is still scope of improvements which can make the system much more safe for people if acknowledged in the right way.

We can make transportation safer with the use of technology. Technology makes things more efficient and reliable. It helps us in not only studying the artefacts of crashes or designing the highways but also provides a way deeper insight into the future. For this purpose, there is strong need to study and analyze the data received from various resources. This data can be incidental data, flow detector data or public outreach data.

Intelligent Transportation System deals with storage and retrieval of huge amount of traffic information efficiently and also analyzing it for the prediction of performance of the existing system. Analysis of such data helps not only in making transportation system more reliable but it also helps in setting future guidelines which in turn would reduce the traffic fatalities and financial burden created by them. Moreover, resources utilization can be optimized with help of transportation data analysis.

Nevada Department of Transportation(NDOT), Regional Transportation Center(RTC),Freeway & Arterial System of Transportation(FAST) and Transportation Research Center at UNLV have collaborated to address all such issues on the freeway I-15. Goals of the project are as follows:-

1. Conduct a comprehensive evaluation study on I-15 North Design Build Project

2. Analyze project implementation with respect to construction zone rules by

which contractor (NCC) had to abide

3. Analyze data collected from FAST for the various detour strategies

4. Study public outreach methods and their effectiveness

5. Conduct simulations of the network as well as major intersections

6. Develop guidelines for future major urban freeway detours.

Developing such guidelines are of major importance because over the next decade several major freeway reconstruction projects are planned in Nevada.

To achieve the goal we collected the data from various resources inlcuding the flow detector data provided by RTC-FAST. State-of-the-art techniques were applied to analyze the data which was in various formats. Statistical analysis techniques like wavelet analysis, logistic regression, T-Test, Anova, methods of moments were applied to analyze quantitative data. Several machine learning and NLP(Natural Language Processing) techniques were applied to qualitative data to predict the severity of incidents, extraction of sentiments in user comments and finding exact sense of words in which they were used in user comments or complaint. We developed an integrated software PAS(Performance Analysis System) for analysis in future.

CHAPTER 1

Introduction

1.1 Motivation

When the contractor for the I-15 Freeway North Design-Build project proposed a major lane reduction (three each way to two each way), many professionals were concerned about severe congestion on the freeway and local parallel arterials. However, with the efforts of the traffic experts, especially at FAST, sophisticated and innovative signal system patterns and detours were developed and implemented. Additionally, before the detour system was installed, an extensive public outreach was conducted. The resulting congestion increase was surprisingly less than anticipated. The detour system was dynamic so it could respond to incidents as well as changes to travel patterns as the construction work progressed. Effective study and analysis of past data can result in much efficient transportation system. Transportation Research Center, NDOT and RTC-FAST collaborated for the study of the past data and based on that system performance analysis was done.

This report has been divided into two parts. First part deals with the analysis of quantitative data. In second part, qualitative data analysis techniques and results have been described.

1.2 Quantitative Data Analysis

Quantitative data was gathered mostly from the flow detector sensors placed on various intersecions and highways. Flow detector data provides information about the volume of traffic and average speed of the traffic passing through the places where sensors are present.

We applied a lot of techniques for the analysis(Multivariate Analysis, Wavelet Analysis, Method of Moments, T-Test, ANOVA etc.) which are described in details in the next chapter of quantitative analysis.

1.3 Qualitatitve Data Analysis

Qualitatitve data was gathered from the complaint system of fast, various news resources and incident log provided by FAST. We applied state-of-the-art techniques to analyze the data and provide solutions to the problems. Here are some of the problems that were addressed:-

1. Extract sentiment from a given user comment or complaint.
2. Automatic prediction of severity(Negligible, Noticeable or Significant) based on some parameters(described later).
3. Automatically finding the exact sense of words that are used in a complaints or comments.
4. Labeling the data according to the information present in the statement(known as concept discovery).

These are some of the problems which are being focused upon. Detailed description to all the problems and solutions have been described in the chapter related to qualitative data analysis.

1.4 Conclusion

An integrated performance analysis tool was developed(named as PAS) for future. It addresses most of the issues described above. Based on the data analysis, performance of the system can be evaluated. Moreover, proper safety precautions and measures can be taken into account for future incidents and improving the system further.

CHAPTER 2

DATA SOURCES AND CHARACTERISTICS

Following is the list of major sources of various kinds of traffic and transportation related data that TRC acquires periodically.

- FAST- Freeway and Arterial System of Transportation
- LVMPD- Las Vegas Metropolitan Police Department
- NDOT- Nevada Department of Transportation
- NHP- Nevada Highway Patrol
- TRC, UNLV- Transportation Research Center, University of Nevada Las Vegas.
- UMC- University Medical Center

2.1 Data Sources and Characteristics

Following is the list and brief description of data available with TRC at present:

2.1.1 FAST

1. Flow Detector Data: The flow detector data is available to TRC as a live feed for two highway stretches at present; US-95 and I-15. It records the data every

minute about the occupancy of various lanes. Parameters that are available in the flow detector data are -

- Date and time stamps
- Detector IDs
- Lane wise vehicle count
- Occupancy
- Lane speed

2. SMS Data

    - Data and time
    - Location
    - Lane Blocked

3. AVL Data: The AVL data is recorded on the transit vehicles under Regional Transportation Commission (RTC) of Southern Nevada. The current data that is available is for routes 110 and 202. The route 110 (Eastern) functions between Cheyenne/Civic Center and Eastern/St. Rose Pkwy; while the route 202 (Flamingo) functions between Fort Apache/Flamingo and Harmon/Boulder Hwy. With the sensors in the vehicles, the time of arrival of the vehicle at various stops is noted and compared with the stipulated time of arrival at the stop. This allows calculating the delays at various stops which are saved in a data file. The parameters that are recorded into a data file are listed below -

    - Data and time

- Coach Number
- Block Number
- Trip Number
- Stop Locations (Names)
- Arrival Time at the stops
- Delay calculated from the stipulated time of arrival

4. Bluetooth Data: The Bluetooth data is collected by paired bluetooth detectors that are installed on the roads. The current paired bluetooth sensors are located between Hwy 93 & Lakeview Dr (u182) and Hwy 93 & West of Veterans Memorials Dr (u179). The bluetooth data is recording by getting the ID of any available Bluetooth device in a vehicle which crosses the sensor at a location. At the location of the next sensor, the ID is matched to calculate the time taken by the vehicle to traverse the distance between the sensors. Since the distance between the sensors is known, it can be used to calculate the speed of the vehicle. The data is publicly available on bluetoad.trafficcast.com. The data collected is filtered and has the following parameters -

   - Calculated Time
   - Last match Time
   - Travel Time
   - Speed

2.1.2 LVMPD

1. Arterial Incident Management (IM) Data

   - Event Number
   - Create Time
   - Arrival Primary Unit
   - Cleared Time
   - Code

2.1.3 NDOT

1. Crash Data: The accident data is compiled by NDOT, which includes very detailed information about the recorded crashes. The data covers various aspects ranging from the location and time of the accident to the number of fatalities and roadways conditions. A few major parameters that are recorded in the crash data are listed below

   - Date and Time stamp
   - Type of Accident (Hit and run/ vehicle collision)
   - Collision Description
   - Crash Time Origin and Clearance
   - Type of Damage
   - Distance from Street
   - Roadway type and number of lanes

- Lighting Conditions at time of accident
- Number of fatalities and injuries

Crash data is collected by various agencies like Las Vegas Metropolitan Police Department (LVMPD) for Las Vegas, National Highway Patrol (NHP), Sheriff Offices and other sources. Department of Motor Vehicles (DMV) collects data from all these sources and passes on to Nevada Department of Traffic (NDOT) which compiles the crash data happening around the Clark County. Office of Transportation Safety (OTS) Nevada along with University Medical Centre (UMC) Las Vegas and NDOT worked to link the NDOT crash data with the UMC Trauma data. This data is broken into 6 tables each covering an aspect of the crash like vehicle information, location information, crash information, etc. Some of these accidents generate a trauma support request i.e. a 911 call for emergency is made to support victims. It is this data which gets surfaced in trauma records of hospital. Using this information as the formal basis, both the sets of data were linked based on unique accident identification. The received data was then de-identified by UMC; that is all the personal information (like name and address) was removed before being handed over for analysis to TRC, UNLV. The linked data comprised of UMC trauma data for the period 2005-09 and NDOT crash data for 2005-08. Many hospitals in Clark County having the trauma department contributed with the data. Therefore not all the Clark county data is available. NDOT crash data was available for many accidents but since trauma data was the limiting factor, appropriate records were selected from it and linked to crash data. The final linked crash-trauma dataset had 4112 records for the period of 2005-08.

2.1.4 NHP

1. Freeway Incident Management (IM) Data

    - Date and time
    - Location
    - Place
    - Type Of Accident
    - Receive Time
    - Dispatch Time

2.1.5 TRC-UNLV

1. Construction Data

    - Date and time
    - Location
    - Closure point
    - End point
    - Detour
    - Scheduled Work

2. Seatbelt Data: Seatbelt data is collected by TRC every year as part of the statewide seatbelt usage surveys. The data is collected manually through data collection software which was designed in TRC on a PDA (Personal Digital

Assistant). According to the uniform criterion the study is conducted in two counties i.e. Clark and Washoe, selected based on population. There are 32 sites each in both the counties where data is collected. The parameters collected during the study are seatbelt status of the front seat occupants, age group, ethnicity, gender, vehicle type and license of registration. Survey provides us with an unweighted estimate of seatbelt usage. It has the following attributes:

- Gender
- Ethnicity
- Age
- Vehicle category
- State of registration

3. Travel Run Data

    - Date and time
    - Location
    - Speed
    - Traffic Light Status
    - Stopping time at red light
    - Proceeding time at light

4. iPhone Application Data The iphone application developed by TRC, UNLV is used to collect data on travel runs. The device can be simply be mounted in the

vehicle and it collects the data as you drive around. The iPhone can record the parameters at a pre-programmed time interval, which can be changed to allow the frequency of recording data. The various parameters that are recorded are listed below -

- Time Stamp
- Accelerometer data along the three axes
- Gyroscope data along the three axes
- Co-ordinates (Latitude and Longitude)
- Distance interval between recordings
- Total distance traveled
- Speed

2.1.6 UMC

1. Trauma Data

    - Accident Information
    - Patient Information
    - Vehicle Information
    - Traffic Conditions
    - Weather Information
    - People Information

2.2 Conclusion

This chapter listed the data sources and characteristics, which TRC manages and analyzes. Is is clear that the amount of data generated and used by ITS is huge. It requires special techniques to store the data efficiently.

CHAPTER 3

BASIC DATA ANALYSIS

3.1 Summary

Statistical Analysis is the study of collection and organization of data by which some serious conclusions can be drawn and further it can be used for future predictions. We have studied the effect of construction on I-15 freeway through different analysis techniques. The study was intended to observe the impact of construction events on traffic.

For this purpose, flow detector data has been studied before, during and after constructions on several construction sites. Three such events have been considered in the study and the statistical analysis has been applied to the volume of traffic passing through and on the average speed. Major effects were obsereved on the speed data while volume seems much less effected.

3.2 Analysis Techniques

We studied the effect of construction event by applying the following basic techniques:

1. Study of general statistics Mean, Standard Deviation, Higher Order Moments and Correlation between data.

2. T-Test.

3. ANOVA(Analysis of Variance) 4. Histogram and Entropy comparisons were studied to study the distribution and uncertainity in data.

T-Test and ANOVA have been described as follows.

3.2.1 T-Test

It is a statistical technique to estimate whether two groups of data have same statistical means or not. Assuming two groups of data to have same statistical means is called Null hypothesis. T-Test measures a ratio. The numerator is nothing but the difference between averages of two data set while the denominator is the measure of variability which is also known as **standard error of the difference**. The denominator is defined as follows:-

$$\boxed{N = X_1 - X_2}$$

$$\boxed{D = \sqrt{\frac{VAR_1}{N_1} + \frac{VAR_2}{N_2}}}$$

$$\boxed{t = \frac{N}{D}}$$

where, t = T-Test Value

X_1 = Mean of first dataset

$$X_2 = \text{Mean of second dataset}$$

$$VAR_1 = \text{Variance of first dataset}$$

$$N_1 = \text{Number of data values in first set}$$

$$VAR_2 = \text{Variance of second dataset}$$

$$N_2 = \text{Number of data values in second set}$$

We reject or fail to reject this Null hypothesis based on the value of t. If the t value is less than .05 (5% significant level) then we fail to reject the null hypothesis and thus deduce that both dataset have the same mean. If the value is more that .05 then we reject the null hypothesis. In Matlab T-Test returns the value 0 or 1 corresponding value within significance level or not. T-Test is used when there is need to compare only two datasets.

3.2.2 Analysis of Variance(ANOVA)

T-Test has its limitation. T-Test is used when only two datasets are compared while ANOVA comes into play when there is need to compare more than two dataset which in our case is three(data before, during and after construction). It is same as T-Test when used for two dataset.

In Matlab p = anova1(X) performs balanced one-way ANOVA for comparing the means of two or more columns of data in the matrix X, where each column represents an independent sample containing mutually independent observations. The function returns the p value under the null hypothesis that all samples in X are drawn from populations with the same mean.

If p is near zero, it casts doubt on the null hypothesis and suggests that at least one sample mean is significantly different than the other sample means. Common significance levels are 0.05 or 0.01. The anova1 function displays two figures, the standard ANOVA table and a box plot of the columns of X. The standard ANOVA table divides the variability of the data into two parts:

1. Variability due to the differences among the column means (variability between groups)

2. Variability due to the differences between the data in each column and the column mean (variability within groups)

The standard ANOVA table has six columns:

1. The source of the variability.
2. The sum of squares (SS) due to each source.
3. The degrees of freedom (df) associated with each source.
4. The mean squares (MS) for each source, which is the ratio SS/df.
5. The F-statistic, which is the ratio of the mean squares.
6. The p value, which is derived from the cdf of F.

3.3 Analysis of Data from Detector(Western Ave & West Wall St)

3.3.1 Volume Data Analysis

3.3.1.1 Data Plot

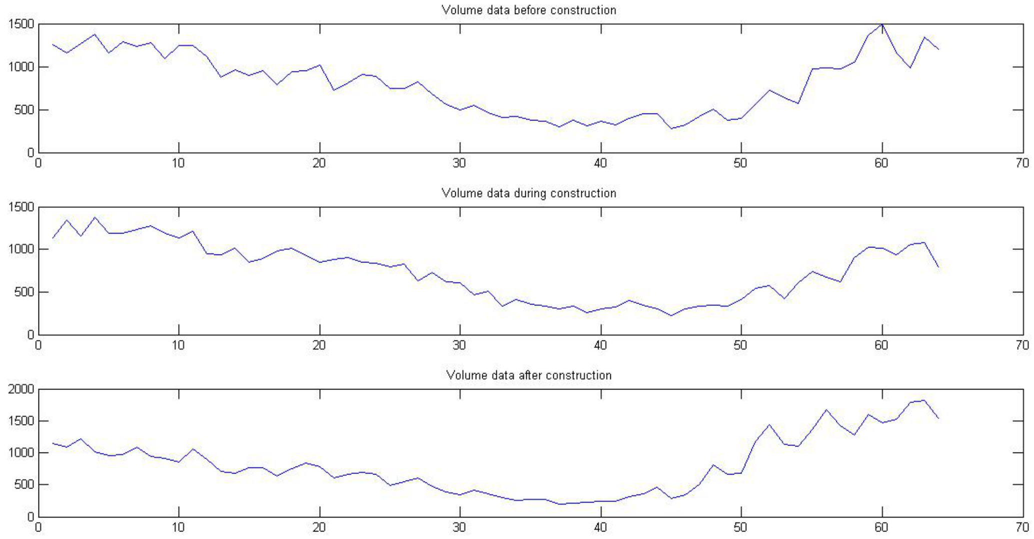


Figure 3.1: Volume data comparison before, during and after construction

3.3.1.2 Statistical Means

Volume Data	Statistical Mean
Before Construction	802.5781
During Construction	740.7656
After Construction	801.0469

3.3.1.3 Standard Deviations

Volume Data	Standard Deviation
Before Construction	349.6965
During Construction	331.2371
After Construction	443.8480

3.3.1.4 T-Test

T-Test result of volume ,data before and during construction = 0 (i.e. we fail to reject Null Hypothesis at the 5% significance level. So data can be considered from same means and same but unknown variance.)

T-Test result of data before and after construction = 0 (i.e. we fail to reject Null Hypothesis at the 5% significance level. So data can be considered from same means and same but unknown variance.)

3.3.1.5 Correlation

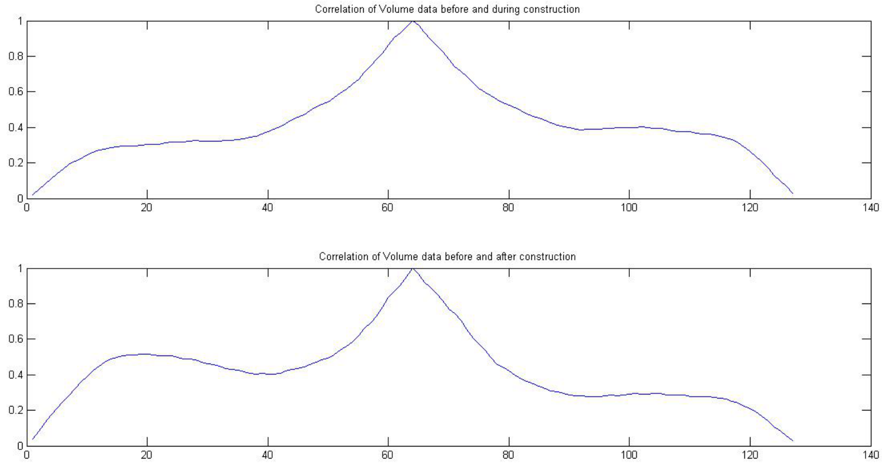
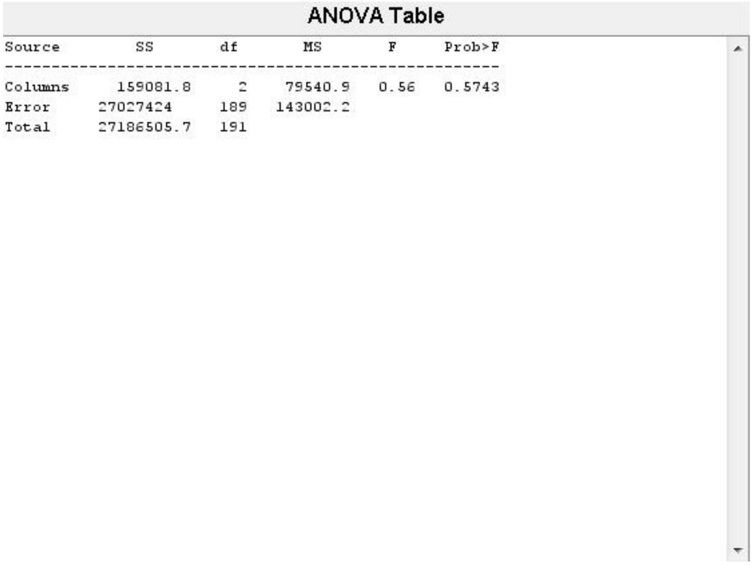


Figure 3.2: Correlation between volume data

3.3.1.6 Higher Order Moments Comparisons

Volume Data	Second Order Moment	Third Order Moment
Before Construction	1.2038e+05	4.3558e+06
During Construction	1.0800e+05	2.2330e+06
After Construction	1.9392e+05	4.4995e+07

3.3.1.7 Result of ANOVA Test



ANOVA Table

```
Source        SS        df       MS        F     Prob>F
-----------------------------------------------------
Columns    159081.8      2    79540.9    0.56    0.5743
Error     27027424     189   143002.2
Total     27186505.7   191
```

Figure 3.3: Analysis of Variance on Volume Data

As in our case it can be seen that the p value is .5743 which is much greater than common significance levels, we fail to reject Null Hypothesis.

3.3.1.8 Histogram

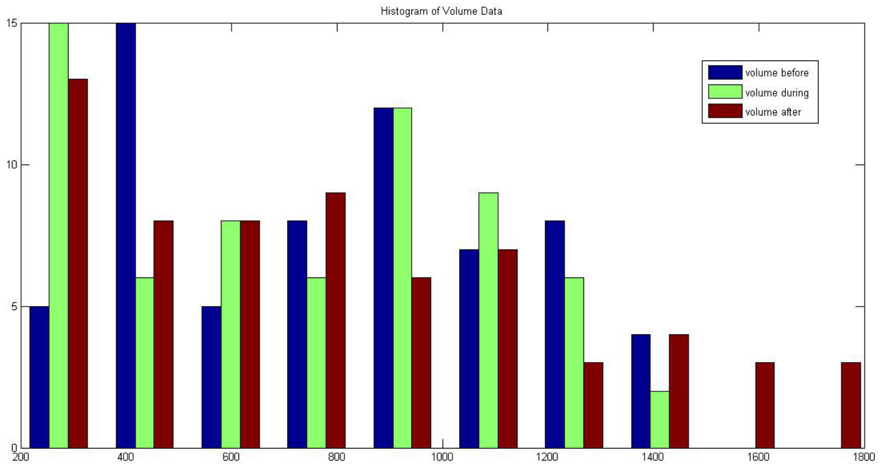


Figure 3.4: Histogram of Volume Data

3.3.1.9 Entropy

From the histogram shown above we estimated the entropy of volume data before, during and after construction.

Volume Data	Entropy
Before Construction	3.1505
During Construction	3.1587
After Construction	3.1553

3.3.2 Speed Data Analysis

3.3.2.1 Data Plot

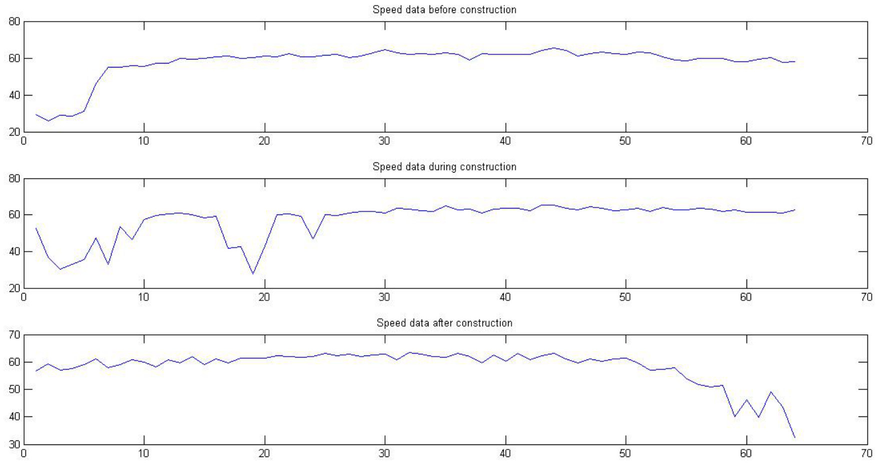


Figure 3.5: Speed data comparison before, during and after construction

3.3.2.2 Statistical Means

Speed Data	Statistical Mean
Before Construction	57.9336
During Construction	57.3398
After Construction	58.4219

3.3.2.3 Standard Deviations

Speed Data	Standard Deviation
Before Construction	9.0052
During Construction	9.7284
After Construction	6.2389

3.3.2.4 T-Test

T-Test result of speed data before and during construction = 0 (i.e. we fail to reject Null Hypothesis at the 5% significance level. So data can be considered from same means and same but unknown variance.)

T-Test result of speed data before and after construction = 0 (i.e. we fail to reject Null Hypothesis at the 5% significance level. So data can be considered from same means and same but unknown variance.)

3.3.2.5 Correlation

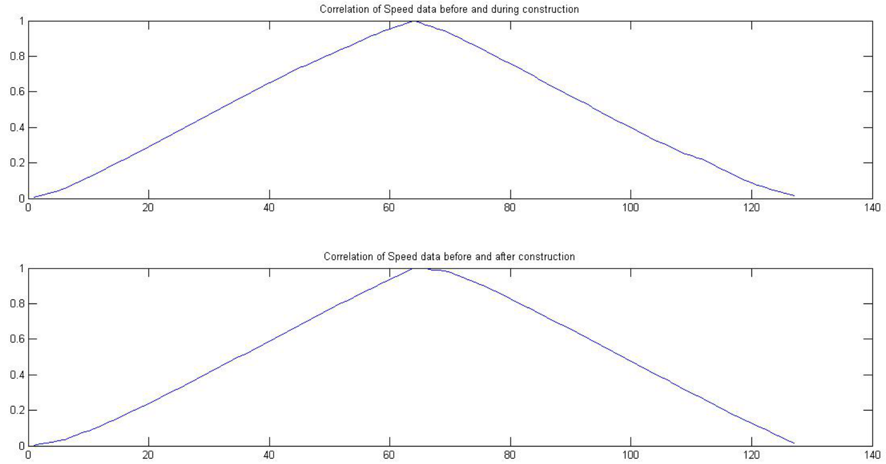
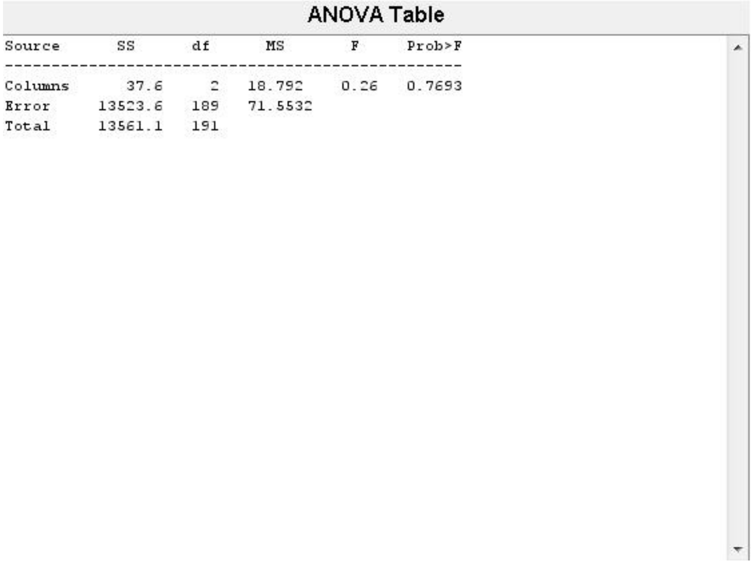


Figure 3.6: Correlation between speed data

3.3.2.6 Higher Order Moments Comparisons

Speed Data	Second Order Moment	Third Order Moment
Before Construction	79.8266	-1914.4
During Construction	93.1628	-1575.2
After Construction	38.3162	-554.7882

3.3.2.7 Result of ANOVA Test



```
Source       SS       df       MS        F     Prob>F
-------------------------------------------------------
Columns      37.6      2    18.792     0.26    0.7693
Error     13523.6    189    71.5532
Total     13561.1    191
```

Figure 3.7: Analysis of Variance on Speed Data

As in our case it can be seen that the p value is .7495 which is much greater than common significance levels, we fail to reject Null Hypothesis.

3.3.2.8 Histogram

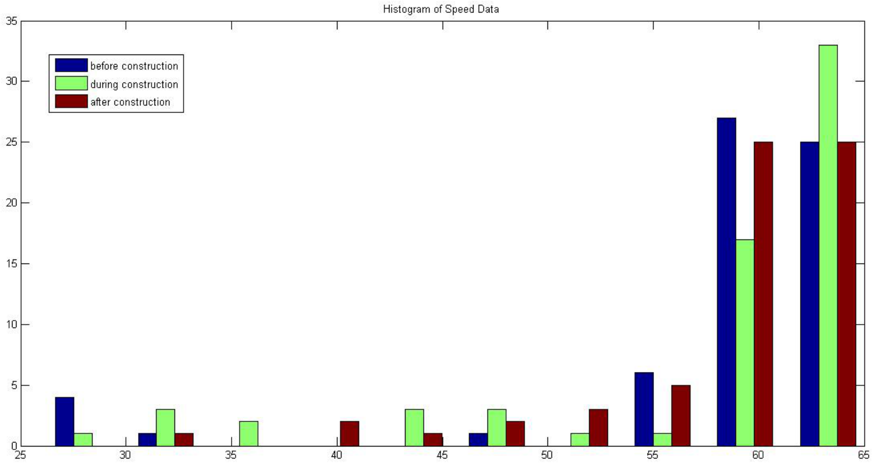


Figure 3.8: Histogram of Speed Data

3.3.2.9 Entropy

From the histogram shown above we estimated the entropy of speed data before, during and after construction.

Speed Data	Entropy
Before Construction	1.8127
During Construction	2.1684
After Construction	2.0355

3.4 Analysis of Data from Detector(E Carey Ave & 5th St

3.4.1 Volume Data Analysis

3.4.1.1 Data Plot

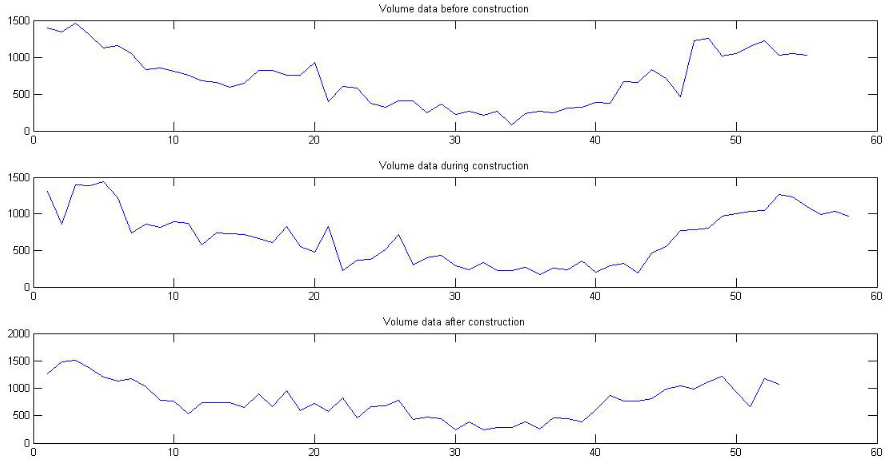


Figure 3.9: Volume data comparison before, during and after construction

3.4.1.2 Statistical Means

Volume Data	Statistical Mean
Before Construction	708.9636
During Construction	680.8793
After Construction	767.6226

3.4.1.3 Standard Deviations

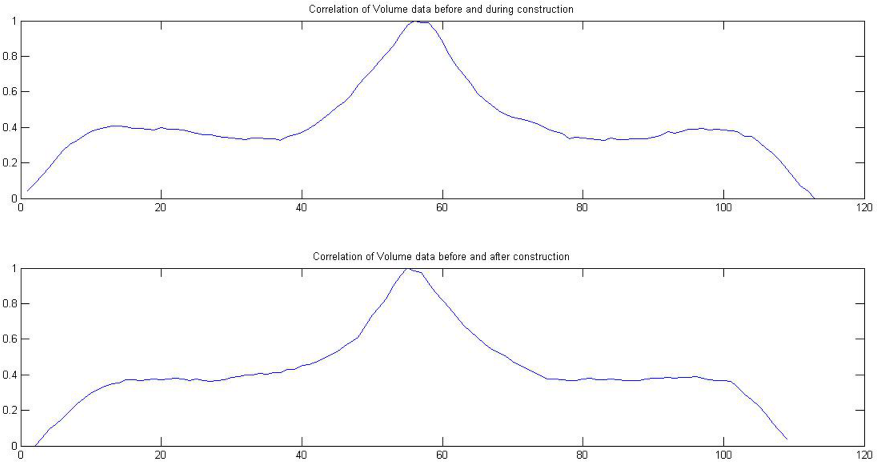
Volume Data	Standard Deviation
Before Construction	370.1794
During Construction	362.0547
After Construction	328.0738

3.4.1.4 T-Test

T-Test result of volume ,data before and during construction = 0 (i.e. we fail to reject Null Hypothesis at the 5% significance level. So data can be considered from same means and same but unknown variance.)

T-Test result of data before and after construction = 0 (i.e. we fail to reject Null Hypothesis at the 5% significance level. So data can be considered from same means and same but unknown variance.)

3.4.1.5 Correlation


Correlation of Volume data before and during construction

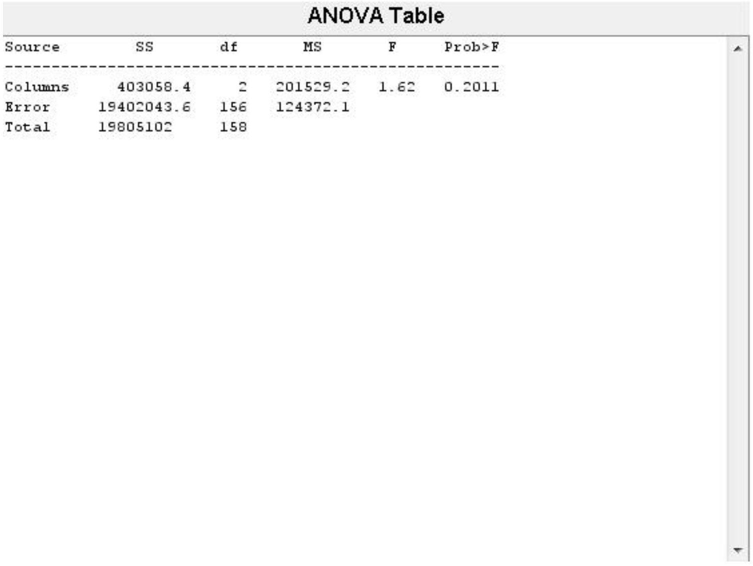
Correlation of Volume data before and after construction

Figure 3.10: Correlation between volume data

3.4.1.6 Higher Order Moments Comparisons

Volume Data	Second Order Moment	Third Order Moment
Before Construction	1.3454e+05	1.1815e+007
During Construction	1.2882e+05	1.5513e+007
After Construction	1.0560e+05	1.0907e+007

3.4.1.7 Result of ANOVA Test



ANOVA Table

```
Source         SS       df       MS        F     Prob>F
-------------------------------------------------------
Columns     403058.4     2    201529.2   1.62   0.2011
Error     19402043.6   156    124372.1
Total     19805102     158
```

Figure 3.11: Analysis of Variance on Volume Data

As in our case it can be seen that the p value is .2011 which is much greater than common significance levels(.05), we fail to reject Null Hypothesis.

3.4.1.8 Histogram

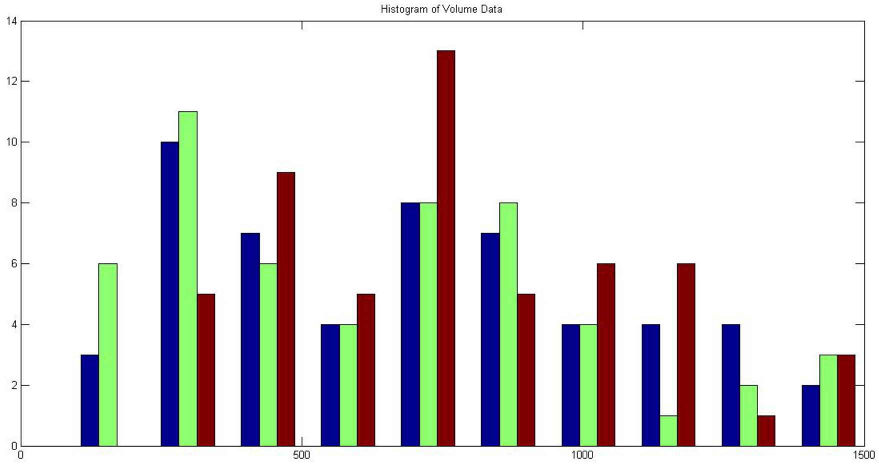


Figure 3.12: Histogram of Volume Data

3.4.1.9 Entropy

From the histogram shown above we estimated the entropy of volume data before, during and after construction.

Volume Data	Entropy
Before Construction	3.1634
During Construction	3.1280
After Construction	3.1725

3.4.2 Speed Data Analysis

3.4.2.1 Data Plot

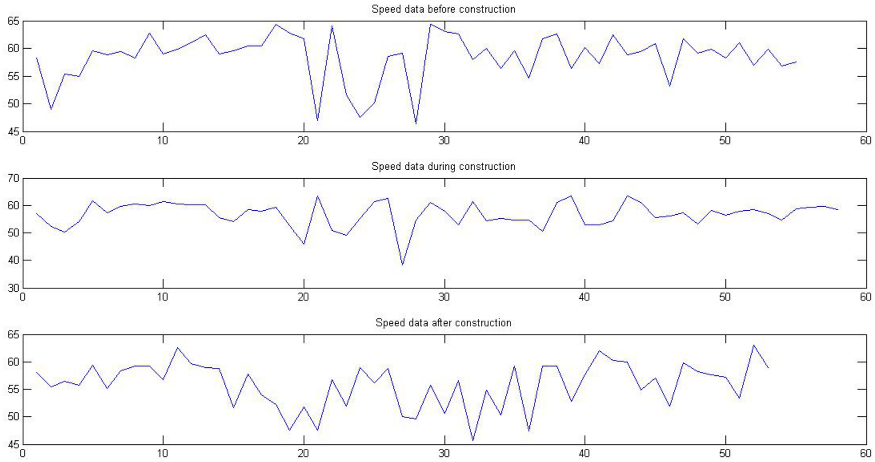


Figure 3.13: Speed data comparison before, during and after construction

3.4.2.2 Statistical Means

Speed Data	Statistical Mean
Before Construction	58.4621
During Construction	56.6655
After Construction	55.9257

3.4.2.3 Standard Deviations

Speed Data	Standard Deviation
Before Construction	4.2647
During Construction	4.9020
After Construction	4.1389

3.4.2.4 T-Test

T-Test result of speed data before and during construction = 1 (i.e. we reject Null Hypothesis at the 5% significance level. So data can be considered from same means and same but unknown variance.)

T-Test result of speed data before and after construction = 1 (i.e. we reject Null Hypothesis at the 5% significance level. So data can be considered from same means and same but unknown variance.)

3.4.2.5 Correlation

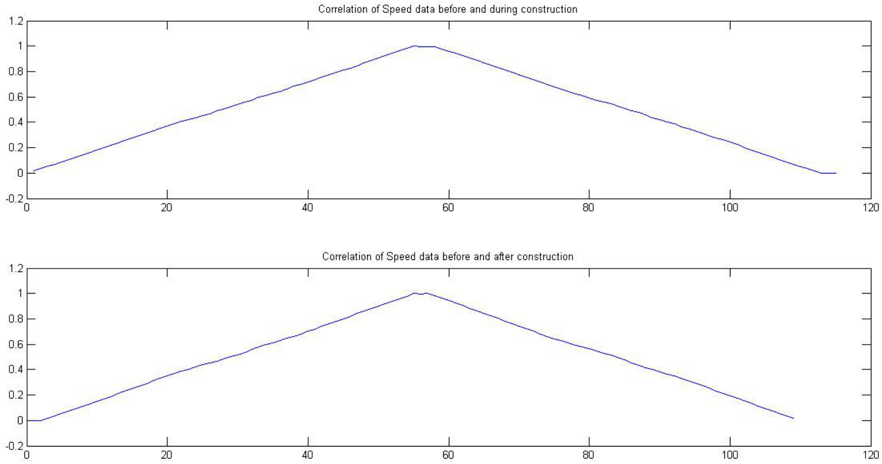
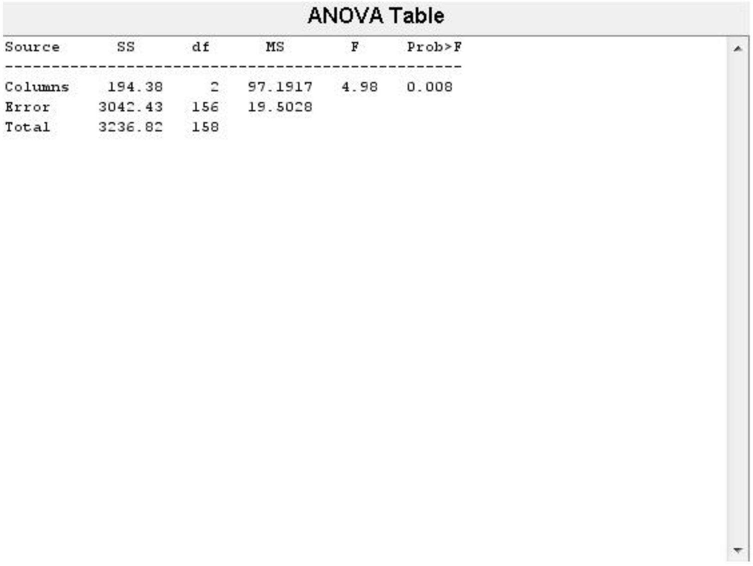


Figure 3.14: Correlation between speed data

3.4.2.6 Higher Order Moments Comparisons

Speed Data	Second Order Moment	Third Order Moment
Before Construction	17.8566	-94.7114
During Construction	20.7067	-118.7843
After Construction	16.8069	-45.3780

3.4.2.7 Result of ANOVA Test



ANOVA Table

```
Source       SS      df      MS        F     Prob>F
-------------------------------------------------
Columns    194.38     2   97.1917   4.98   0.008
Error     3042.43   156   19.5028
Total     3236.82   158
```

Figure 3.15: Analysis of Variance on Speed Data

As in our case it can be seen that the p value is .008 which is much smaller than common significance levels, we reject Null Hypothesis.

3.4.2.8 Histogram

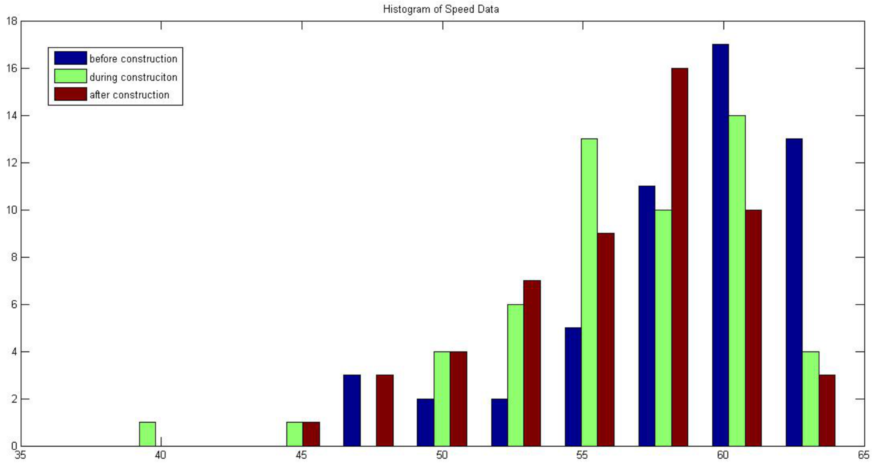


Figure 3.16: Histogram of Speed Data

3.4.2.9 Entropy

From the histogram shown above we estimated the entropy of speed data before, during and after construction.

Speed Data	Entropy
Before Construction	2.8908
During Construction	2.6527
After Construction	3.0012

3.5 Analysis of Data from Detector(E Evans St & Bulloch St

3.5.1 Volume Data Analysis

3.5.1.1 Data Plot

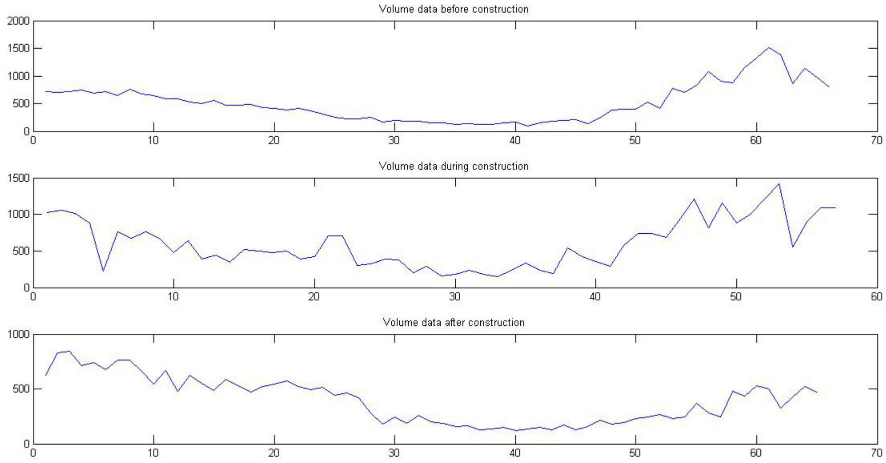


Figure 3.17: Volume data comparison before, during and after construction

3.5.1.2 Statistical Means

Volume Data	Statistical Mean
Before Construction	515.0909
During Construction	596.2807
After Construction	394.9692

3.5.1.3 Standard Deviations

Volume Data	Standard Deviation
Before Construction	344.4260
During Construction	327.0426
After Construction	208.3354

3.5.1.4 T-Test

T-Test result of volume data before and during construction = 1 (i.e. we reject Null Hypothesis at the 5% significance level. So data can be considered from same means and same but unknown variance.)

T-Test result of volume data before and after construction = 0 (i.e. we fail to reject Null Hypothesis at the 5% significance level. So data can be considered from same means and same but unknown variance.)

3.5.1.5 Correlation

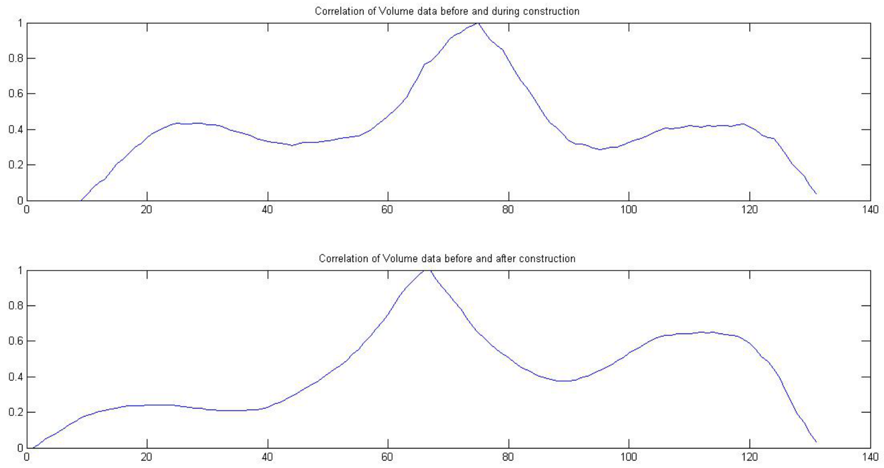
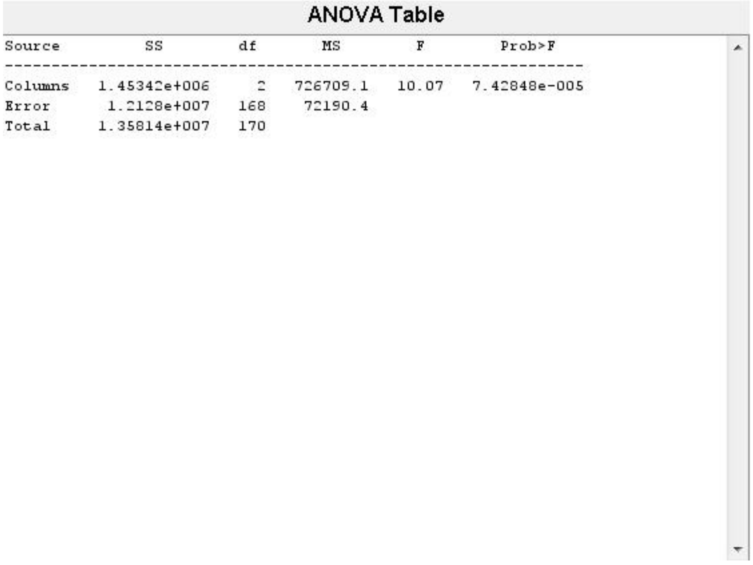


Figure 3.18: Correlation between volume data

3.5.1.6 Higher Order Moments Comparisons

Volume Data	Second Order Moment	Third Order Moment
Before Construction	1.1683e+005	3.4977e+007
During Construction	1.0508e+005	1.8440e+007
After Construction	4.2736e+004	2.8781e+006

3.5.1.7 Result of ANOVA Test



ANOVA Table

```
Source          SS         df        MS         F        Prob>F
----------------------------------------------------------------
Columns   1.45342e+006      2    726709.1    10.07   7.42848e-005
Error      1.2128e+007    168     72190.4
Total     1.35814e+007    170
```

Figure 3.19: Analysis of Variance on Volume Data

As in our case it can be seen that the p value is much smaller than common significance levels(.05), we reject Null Hypothesis.

3.5.1.8 Histogram

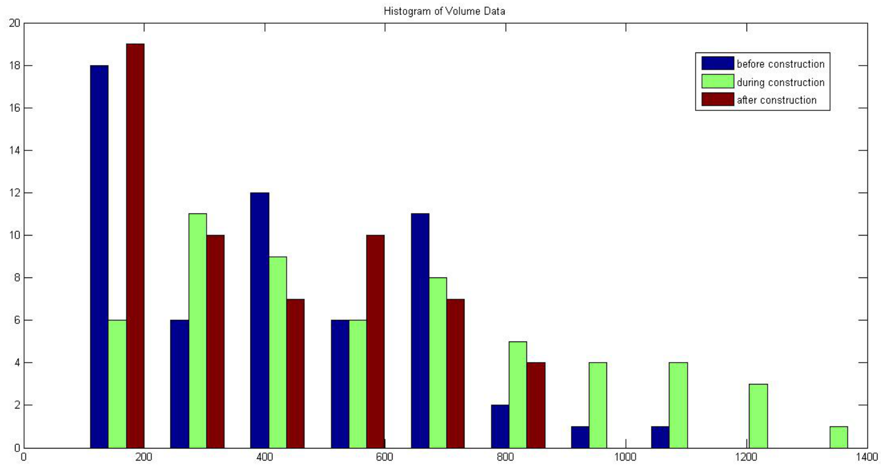


Figure 3.20: Histogram of Volume Data

3.5.1.9 Entropy

From the histogram shown above we estimated the entropy of volume data before, during and after construction.

Volume Data	Entropy
Before Construction	2.8391
During Construction	3.0644
After Construction	2.9225

3.5.2 Speed Data Analysis

3.5.2.1 Data Plot

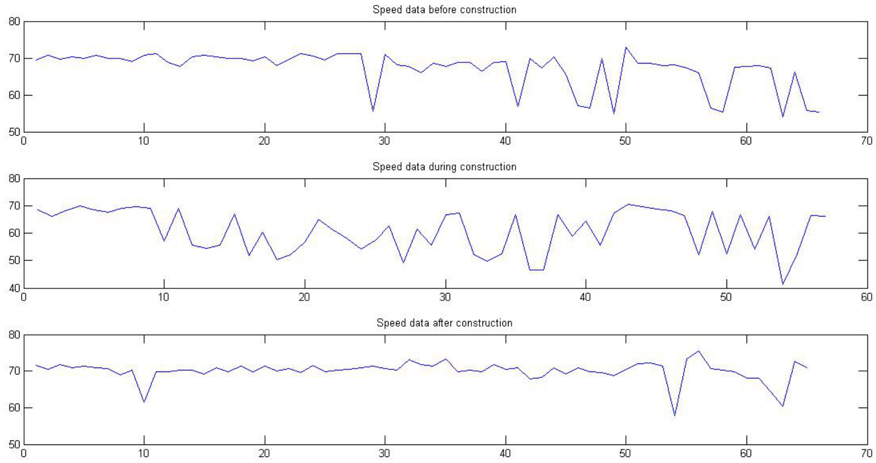


Figure 3.21: Speed data comparison before, during and after construction

3.5.2.2 Statistical Means

Speed Data	Statistical Mean
Before Construction	66.9176
During Construction	60.5826
After Construction	70.0421

3.5.2.3 Standard Deviations

Speed Data	Standard Deviation
Before Construction	5.0730
During Construction	7.6934
After Construction	2.7335

3.5.2.4 T-Test

T-Test result of speed data before and during construction = 1 (i.e. we reject Null Hypothesis at the 5% significance level. So data can be considered from same means and same but unknown variance.)

T-Test result of speed data before and after construction = 1 (i.e. we reject Null Hypothesis at the 5% significance level. So data can be considered from same means and same but unknown variance.)

3.5.2.5 Correlation

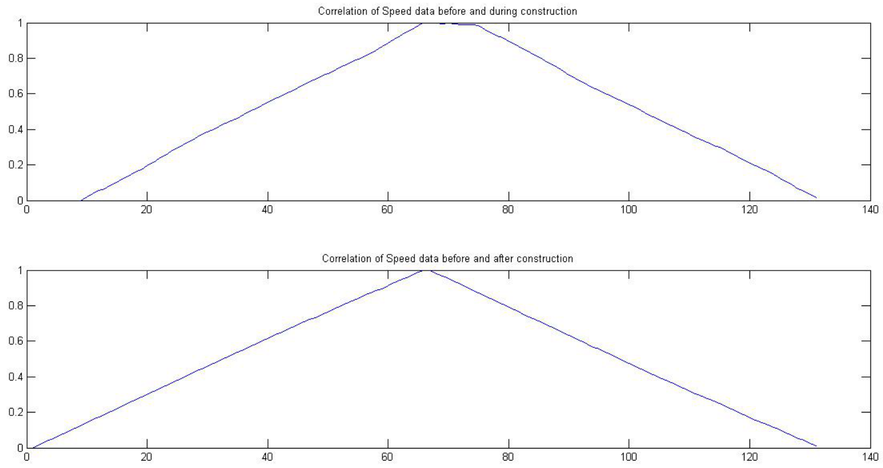
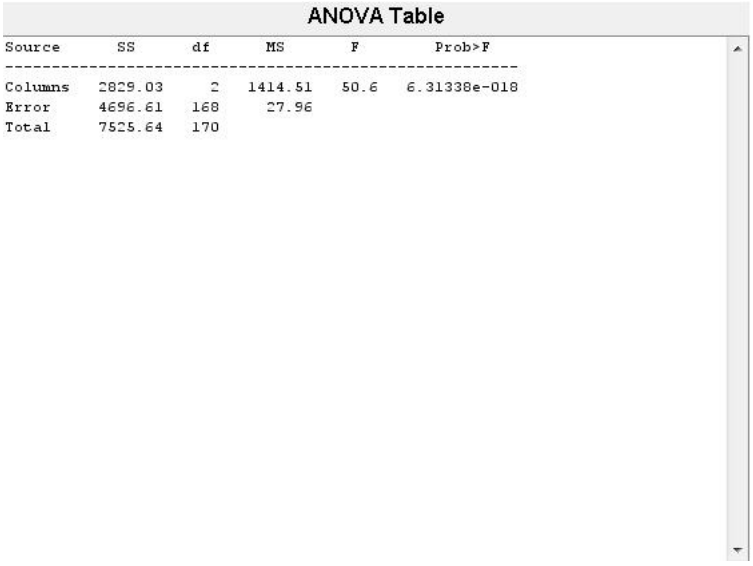


Figure 3.22: Correlation between speed data

3.5.2.6 Higher Order Moments Comparisons

Speed Data	Second Order Moment	Third Order Moment
Before Construction	25.3451	-207.3147
During Construction	58.1500	-219.2646
After Construction	7.3572	-49.5824

3.5.2.7 Result of ANOVA Test



ANOVA Table

```
Source       SS      df       MS        F        Prob>F
-------------------------------------------------------
Columns   2829.03     2    1414.51   50.6   6.31338e-018
Error     4696.61   168      27.96
Total     7525.64   170
```

Figure 3.23: Analysis of Variance on Speed Data

As in our case it can be seen that the p value is much smaller than common significance levels, we reject Null Hypothesis.

3.5.2.8 Histogram

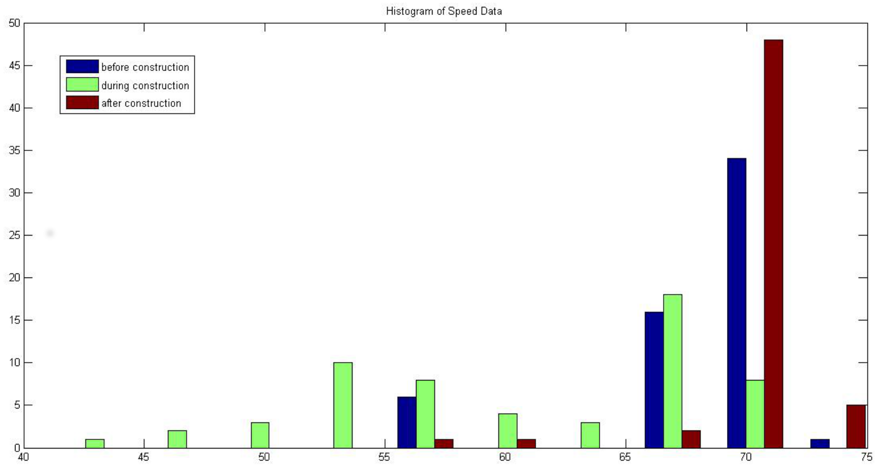


Figure 3.24: Histogram of Speed Data

3.5.2.9 Entropy

From the histogram shown above we estimated the entropy of speed data before, during and after construction.

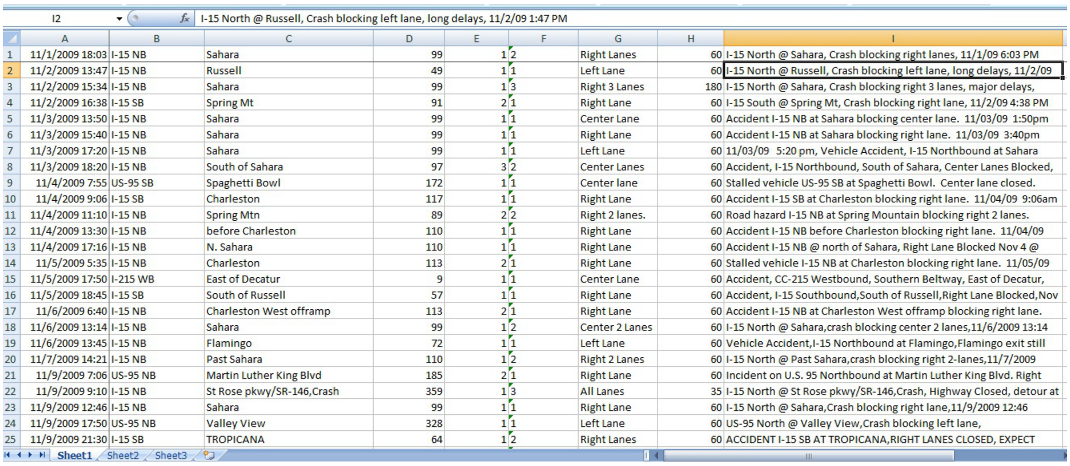
Speed Data	Entropy
Before Construction	2.2903
During Construction	2.8956
After Construction	2.0149

3.6 Incident Data Analysis

RTC-FAST provided TRC-UNLV the data related to incident on which basic analysis was done to extract vital information and to predict the severity of an incident based on other relevant data like number of lanes closed, blockage duration, total time taken to clear the lanes and truck involvement. We integrated these schemes in the Performance Analysis System(PAS) software.

3.6.1 Data Representation

Incident data has the following representation:-



I2 | I-15 North @ Russell, Crash blocking left lane, long delays, 11/2/09 1:47 PM

	A	B	C	D	E	F	G	H	I
1	11/1/2009 18:03	I-15 NB	Sahara	99	1	2	Right Lanes	60	I-15 North @ Sahara, Crash blocking right lanes, 11/1/09 6:03 PM
2	11/2/2009 13:47	I-15 NB	Russell	49	1	1	Left Lane	60	I-15 North @ Russell, Crash blocking left lane, long delays, 11/2/09
3	11/2/2009 15:34	I-15 NB	Sahara	99	1	3	Right 3 Lanes	180	I-15 North @ Sahara, Crash blocking right 3 lanes, major delays,
4	11/2/2009 16:38	I-15 SB	Spring Mt	91	2	1	Right Lane	60	I-15 South @ Spring Mt, Crash blocking right lane, 11/2/09 4:38 PM
5	11/3/2009 13:50	I-15 NB	Sahara	99	1	1	Center Lane	60	Accident I-15 NB at Sahara blocking center lane. 11/03/09 1:50pm
6	11/3/2009 15:40	I-15 NB	Sahara	99	1	1	Right Lane	60	Accident I-15 NB at Sahara blocking right lane. 11/03/09 3:40pm
7	11/3/2009 17:20	I-15 NB	Sahara	99	1	1	Left Lane	60	11/03/09 5:20 pm, Vehicle Accident, I-15 Northbound at Sahara
8	11/3/2009 18:20	I-15 NB	South of Sahara	97	3	2	Center Lanes	60	Accident, I-15 Northbound, South of Sahara, Center Lanes Blocked,
9	11/4/2009 7:55	US-95 SB	Spaghetti Bowl	172	1	1	Center lane	60	Stalled vehicle US-95 SB at Spaghetti Bowl. Center lane closed.
10	11/4/2009 9:06	I-15 SB	Charleston	117	1	1	Right Lane	60	Accident I-15 SB at Charleston blocking right lane. 11/04/09 9:06am
11	11/4/2009 11:10	I-15 NB	Spring Mtn	89	2	2	Right 2 lanes.	60	Road hazard I-15 NB at Spring Mountain blocking right 2 lanes.
12	11/4/2009 13:30	I-15 NB	before Charleston	110	1	1	Right Lane	60	Accident I-15 NB before Charleston blocking right lane. 11/04/09
13	11/4/2009 17:16	I-15 NB	N. Sahara	110	1	1	Right Lane	60	Accident I-15 NB @ north of Sahara, Right Lane Blocked Nov 4 @
14	11/5/2009 5:35	I-15 NB	Charleston	113	2	1	Right Lane	60	Stalled vehicle I-15 NB at Charleston blocking right lane. 11/05/09
15	11/5/2009 17:50	I-215 WB	East of Decatur	9	1	1	Center Lane	60	Accident, CC-215 Westbound, Southern Beltway, East of Decatur,
16	11/5/2009 18:45	I-15 SB	South of Russell	57	1	1	Right Lane	60	Accident, I-15 Southbound,South of Russell,Right Lane Blocked,Nov
17	11/6/2009 6:40	I-15 NB	Charleston West offramp	113	2	1	Right Lane	60	Accident I-15 NB at Charleston West offramp blocking right lane.
18	11/6/2009 13:14	I-15 NB	Sahara	99	1	2	Center 2 Lanes	60	I-15 North @ Sahara,crash blocking center 2 lanes,11/6/2009 13:14
19	11/6/2009 13:45	I-15 NB	Flamingo	72	1	1	Left Lane	60	Vehicle Accident,I-15 Northbound at Flamingo,Flamingo exit still
20	11/7/2009 14:21	I-15 NB	Past Sahara	110	1	2	Right 2 Lanes	60	I-15 North @ Past Sahara,crash blocking right 2-lanes,11/7/2009
21	11/9/2009 7:06	US-95 NB	Martin Luther King Blvd	185	2	1	Right Lane	60	Incident on U.S. 95 Northbound at Martin Luther King Blvd. Right
22	11/9/2009 9:10	I-15 NB	St Rose pkwy/SR-146,Crash	359	1	3	All Lanes	35	I-15 North @ St Rose pkwy/SR-146,Crash, Highway Closed, detour at
23	11/9/2009 12:46	I-15 NB	Sahara	99	1	1	Right Lane	60	I-15 North @ Sahara,Crash blocking right lane,11/9/2009 12:46
24	11/9/2009 17:50	US-95 NB	Valley View	328	1	1	Left Lane	60	US-95 North @ Valley View,Crash blocking left lane,
25	11/9/2009 21:30	I-15 SB	TROPICANA	64	1	2	Right Lanes	60	ACCIDENT I-15 SB AT TROPICANA,RIGHT LANES CLOSED, EXPECT

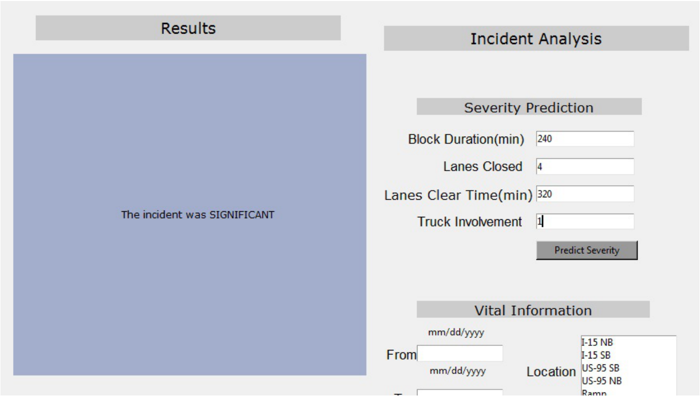
Figure 3.25: Incident Log

3.6.2 Severity Prediction

FAST maintains the incident log as shown in figure 3.25. Incident are tagged with significant, noticeable and negligible severity based on human decision. We provided a noval approach to automatically assign such severity to a given incident based on the data monitored. For this purpose, we extracted the historic tagged data and used it as a training dataset. Training dataset contains a list of features which are tagged with significant, noticeable or negligible severity. These features are nothing but the following four data values which are extracted from the incident log.

1. Block duration
2. Number of lanes closed
3. Lanes clear time
4. Truck involvement

We trained a classifier with this training set which generates a probabilistic model based on which future data or untagged data is assigned the severity based above four data values. The system is shown in the following figure:-



The incident was SIGNIFICANT

Block Duration(min) 240
Lanes Closed 4
Lanes Clear Time(min) 320
Truck Involvement 1

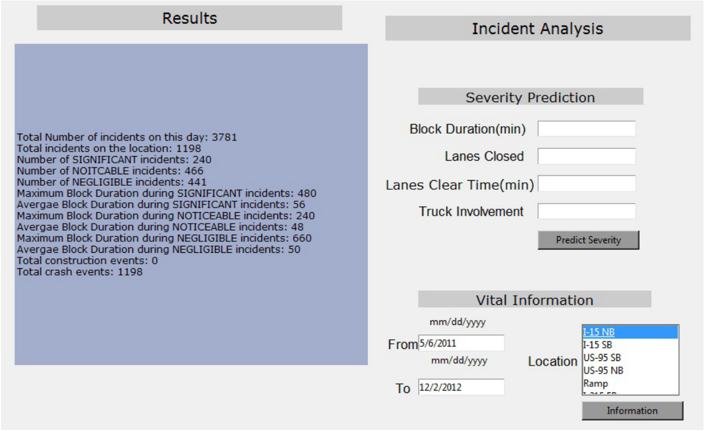
Figure 3.26: Severity Prediction

3.6.3 Information Extraction

We analyzed the incident log to extract vital information regarding incidents respective to freeways in selected time interval. This study could help in managing such mishappenings in future. The system extracts the following information:-

1. Total number of incidents.
2. Total number of incidents on selected freeway.
3. Number of significant incidents.
4. Number of noticeable incidents.
5. Number of negligible incidents.
6. Maximum and average block duration during significant incidents.
7. Maximum and average block duration during noticeable incidents.
8. Maximum and average block duration during negligible incidents.
9. Total construction events in the duration.
10. Total crash events.

The information extraction system is shown in the following figure:-



```
Total Number of incidents on this day: 3781
Total incidents on the location: 1198
Number of SIGNIFICANT incidents: 240
Number of NOITCABLE incidents: 466
Number of NEGLIGIBLE incidents: 441
Maximum Block Duration during SIGNIFICANT incidents: 480
Avergae Block Duration during SIGNIFICANT incidents: 56
Maximum Block Duration during NOTICEABLE incidents: 240
Avergae Block Duration during NOTICEABLE incidents: 48
Maximum Block Duration during NEGLIGIBLE incidents: 660
Avergae Block Duration during NEGLIGIBLE incidents: 50
Total construction events: 0
Total crash events: 1198
```

mm/dd/yyyy
From 5/6/2011
mm/dd/yyyy
To 12/2/2012
Location I-15 NB

Figure 3.27: Information Extraction

CHAPTER 4

MULTIVARIATE DATA ANALYSIS

4.1 Summary

Statistics play an important and major role in the process of development and assesment of models. Although, statistics and data mining share a common goal which is to find some sort of correlation and structure in data yet some factors may be considered as a dissociation between them. Data mining methodology uses not only the concept of database technology but also inherits the concepts from machine learning which are not considered in the domain of statistics from a broader perspective but deep down machine learning methodologies, statistics play the major role. Learning algorthims generally use statistical tests during construction of rules and model formation for the decision making.

Multivariate analysis follows the concepts of multivariate statistics involving observation and analysis of more than one outcome variable at a time. In design and analysis, the technique is used to perform trade studies across multiple dimensions while taking into account the effects of all variables on the responses of interest. It is mostly used for analyzing the concepts with respect to changing scenarios.

We applied the analysis on incident data gathered from 10 years ranging from 1999-2009. Based on the data collected, a hypothesis was formed respected to which we performed certain tests. According to our hypothesis, irrespective of the sex of driver, a traffic crash (which may or may not lead to fatality) is due to young people driving drunk at certain day of the week and at certain hour. To test our hypothesis multivariate analysis was performed. The parameters which were chosen by us were age, sex, accident time, atmospheric conditions, blood alcohol content, travel speed, and day of the week.

4.2 Application of Analysis

The analysis was done in SPSS which a software package used for statistical analysis developed by IBM and the data taken into account is the accident data obtained from Fatality Analysis Reporting System(FARS). The variables included are:

- Age: age of the driver.
- Alcohol: the amount of alcohol consumed by the driver.
- Travel speed: the speed of the vehicle at the time of the accident.
- Atmospheric conditions the atmospheric conditions present at the time of the accident, for example: rainy, cloudy, sunny, etc.
- Day of the week: denotes the day of the week on which the accident occurred. It has been noticed that the number of accidents increases on weekends.
- Sex: sex is inferred biological sex from physical appearance.

4.2.1 Cluster Analysis

All variables were standardized to mean 0 and variance 1. The data set was subdivided into 3 clusters, based on the atmospheric conditions. Based on this analysis method, the atmospheric conditions were grouped under three visibility conditions - poor, average and good visibilities. The poor visibility group consisted of snow, sleet and fog/smog/smoke conditions; average visibility consisted of rain and light snowy conditions while good visibility consisted of clear/cloudy condition (no adverse condition). The results suggest that in case of such atmospheric conditions, when we have blowing snow or fog or sleet - people in higher age group tend to lose control of their vehicle more often and hence have higher speed and accident rates. This group of people has lower blood alcohol content, and hence, the external factor contributing to such incidences is the atmospheric condition.

Based on the analysis of variance(ANOVA), we see from the following figures, that changing pattern of variables with atmospheric conditions the number of incidence of accidents are less in case of good visibility and lower age, as compared to poor visibility/higher age. Sober drivers tend to overcome driving conditions (poor/average) which consists of low snow and rain. But as the alcohol content goes up, the number of incidence goes up, even in cases of good visibility. Drivers at higher speed tend to have more accidents than those at lower speed.

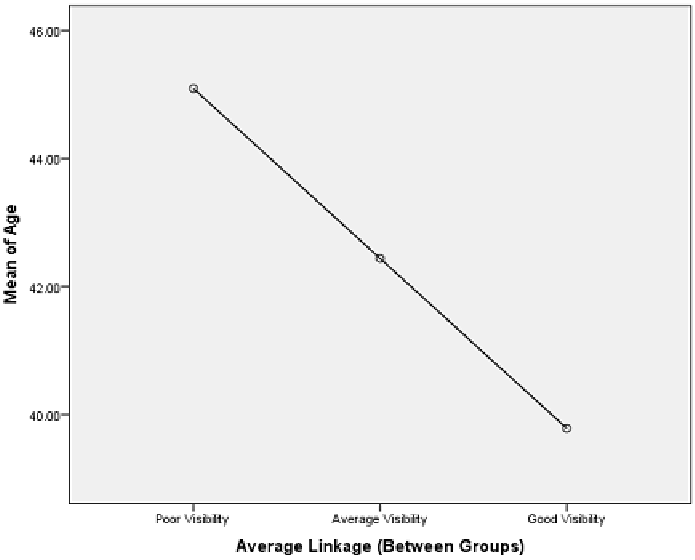


Figure 4.1: Mean plot of age Vs visibility for three cluster solution

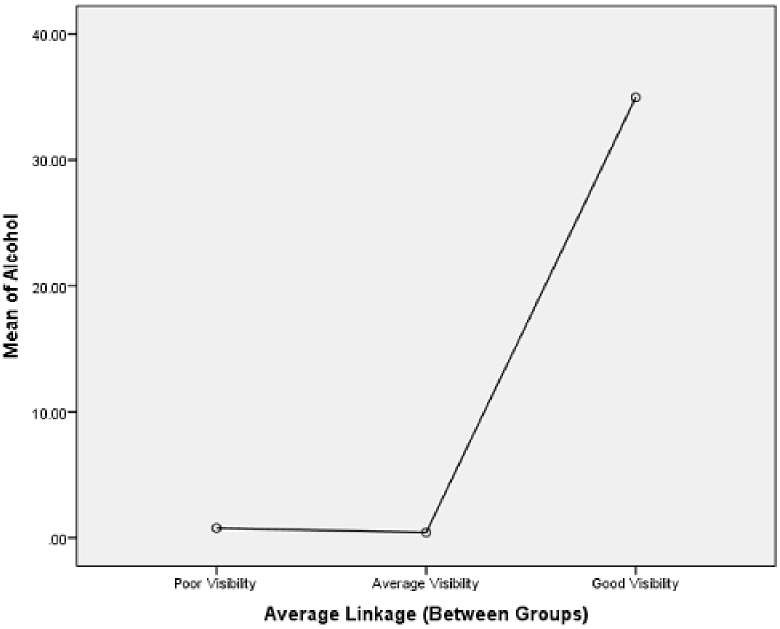


Figure 4.2: Mean plot of blood alcohol content Vs visibility for three cluster solution

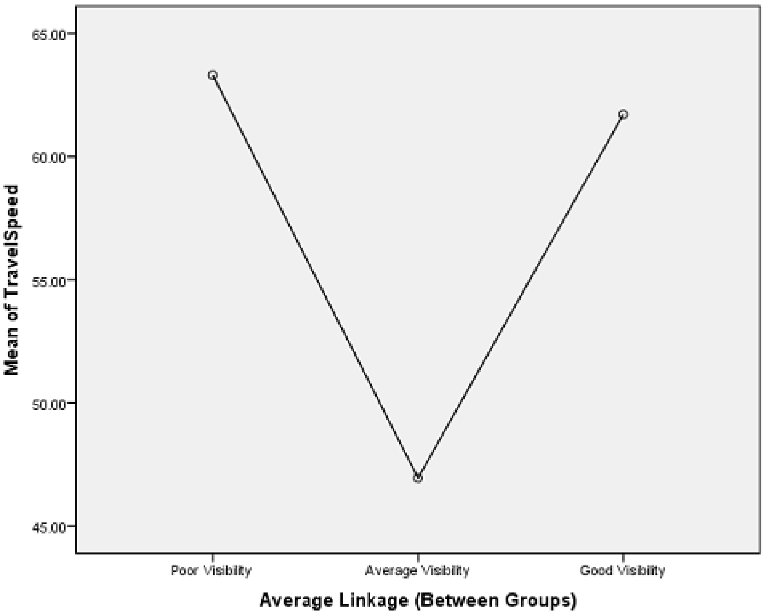


Figure 4.3: Mean plot of Travel Speed Vs visibility for three cluster solution

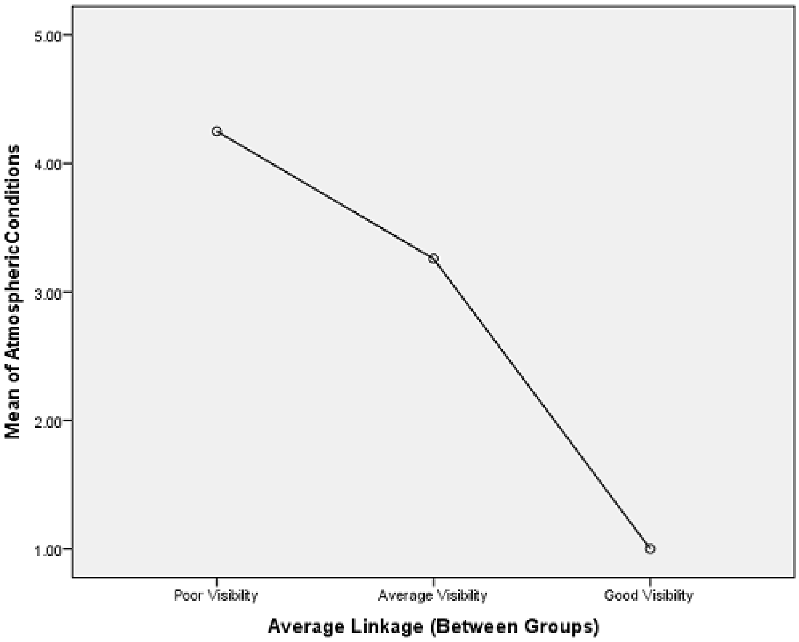


Figure 4.4: Mean plot of Atmospheric Condition Vs visibility for three cluster solution

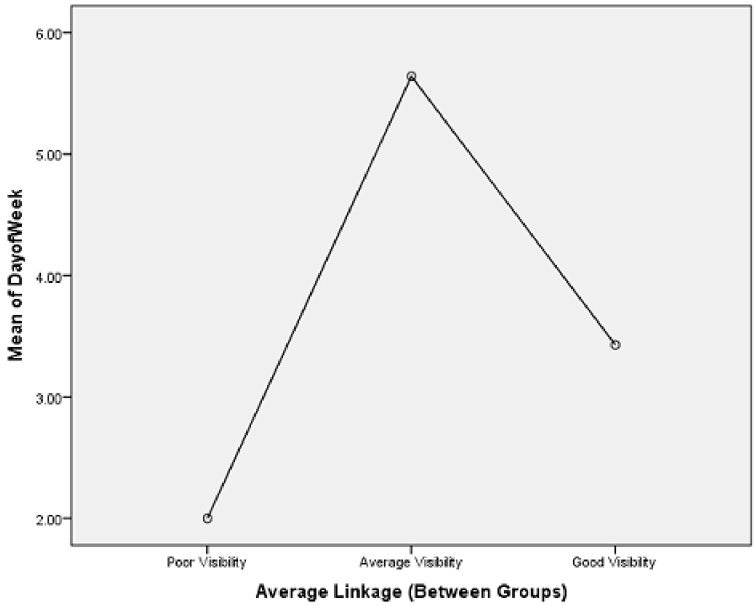


Figure 4.5: Mean plot of Day of the week Vs Visibility for three cluster solution

Based on atmospheric conditions, it is evident that the accident rate decreases with better visibility. Comparison of day of the week and visibility suggests that most accidents took place on/around weekend (Friday), even though the visibility was not poor, but not good.

4.2.2 Multivariate Regression Analysis

Based on this analysis, it is deduced that atmospheric conditions and time of the day matter the most in case of accidents. Alcohol content, age of the driver and day of the week are the factors which follow them. Based on the experience, this result seemed quite logical as adverse weather conditions and high volume traffic in rush hours have led to higher incidences of accidents.

Further we grouped our database to based on sex to understand the effect of age on accidents. It is interesting to note that male drivers outnumber in accidents that their female counterparts, as the regression tend towards a positive trend for male than female. Male drivers, it seems are more prone to drive drunk on high traffic days. The time of accident has less partial effect for both groups.

However, we also see that incidence among the female group is more significant when we use other predictors like accident time, day of the week, age, atmospheric conditions and alcohol. Regression analysis of both groups also shows that atmospheric condition has higher effect on the number of incidence for both groups. For males, time of the day is bigger killer. For female group, alcohol content and age are the next significant factors.

4.2.3 Hierarchical Multiple Regression

We have decided to use age and alcohol as the priority in hierarchy, based on above results which all show that alcohol have significant effect on the number of incidences. The results obtained in this method suggest, that alcohol, age and atmospheric condition are more significant factors in contributing to the number accidents which were caused by high travel speed. Among these three, atmospheric condition plays a bigger role, with high value of significance.

The partial effects of all the factors being used here on the dependent variable, travel speed, suggests irrespective of number of factors used, atmospheric condition continues to be a deciding factor.

4.2.4 Discriminant Analysis

This analysis suggests that accident time, and of the driver may be good discriminator as the separations are large. The results obtained in this method suggest that our data holds good and hence, is adequate for analysis of the hypothesis stated above, as there are significant difference between male and female group for all independents.

4.2.5 Conclusions

Based on our analysis of the data from 1999 to 2009, for all counties in Nevada using different methods we come to following conclusions:

- Adverse atmospheric conditions has more effect and is probably bigger causal factor in accidents than other factors.
- Alcohol content is next contributing factor as sober drivers tend to overcome poor to average atmospheric conditions.
- Accident rate decreases with visibility.
- Accident rates increased as the week came to an end.
- Male drivers outnumber in accident incidences than their female counterparts.
- For male drivers, time of the day has more effect.
- For Female drivers alcohol content and age are the next significant factors after the time of the day.

We have also tested our data for reliability and we can say the following about our data:

- Data being used has no multicollinearity issue.
- Our assumption that the various factors being used in this research has some kind of effect on travel speed is correct based on sphericity test.
- KMO test value is lower than 0.5, which is the general acceptance level for data.

However, it is very close to 0.5, and hence we have continued with the dataset without any alteration. A possible future work could be incorporating more data and testing it initially so that the KMO value is significant, before proceeding with analysis.

CHAPTER 5

WAVELET ANALYSIS

5.1 Summary

Research show that wavelet analysis can be quite efficiently used for denoising and features extraction for any given signal. In our analysis, we explored wavelets in intelligent transportation system for data management, knowledge discovery, incident detection and compression of traffic data. It allows us to separate the common patterns throughout the traffic data and deviations from the average flow by capturing differences from the average flow.

The study was done in two parts. In the first part, we focused on data management, data compression and data visualization techniques using wavelets while in the second section, we have focused upon the automation of incident detection from spatio-temporal(space and time) traffic data.

For example consider traffic flow data which is expected to be more or less same at a particular point during peak hours on weekdays. It might be different on weekends or other hours of the day. Hence instead of storing whole data there is a need to capture the essence of the data. There are two types of data compression schemes:-

- Lossy compression:- Where there is some kind of data loss.

- Lossless compression:- Where there is no data loss.

Wavelet analysis provides us a lossy compression. Although, it looses some data but at the same time it provides more compression therefore less resource is needed to store the data as there is always a trade off between data distortion and compression ratio.

5.2 Wavelet

A wavelet is a piece of wave. We can define a wavelet, simply as a function of a short duration in time which has exactly the same area as above and below x- axis. Fourier transforms uses an infinitely repeating sinusoidal wave whereas a wavelet exists only within finite time duration and is zero elsewhere else. A wavelet transform is performed by convolving the signal against particular instances of the wavelet at various time scales and positions. By modeling changes in frequency (by adjusting the time scale) and modeling time changes (by shifting the position of the wavelet), we can model both frequency changes and location of the frequency.

5.3 Wavelet Analysis

In the wavelet transform we get information about when certain features occurred and information about the scale characteristics of the signal. Scale can be described as analogous to frequency, and is a measure of the amount of detail present in the given signal. Scale is a number related to the number of coefficients and is counter-intuitive to the level of detail. Small scale generally means gross details, and large scale means fine details. In wavelet analysis we first represent our signal as a linear combination of wavelets. This linear combination is formed by using translations and scalings of only one 'mother wavelet'. Coefficients corresponding to the wavelets indicate the significance of that wavelet. These coefficients are called wavelet transform of the data and they highlight local features of the data. As per needs further processing of these coefficients can be done, this may include a de-noising feature extraction, clustering or compression problems. Finally, inverse wavelet transform is computed to reconstruct back the data in the original domain.

5.4 Data Compression and Visualization

5.4.1 One Dimensional Data Compression

The flow detector data that we receive is updated at every five minutes. Flow detectors are installed at freeways. Data contains time stamp, traffc volume, average speed, detector ID and lane occupancy. Data is extracted for a particular detector, for a single day. Here is the representation:-

Time(in minutes)	Volume(Veh/Hour)
2	144
7	84
12	180
17	84
22	132
27	108
32	108
37	96
42	72
47	72
52	48
..	..

Table 5.1: Data from Flow Detector

Table 5.2 and 5.3 show the compression ratio and percentage recovery of the data for two thresholding methods(level independent(hard) and level dependent(soft)). We observe that in case of level independent thresholding percentage recovery is above 99%(only 1% data loss) for all the levels and wavelets, while the compression ratio ranges from 25.35% to 54.52%. Whereas in case of level dependent thresholding compression ratio ranges from 25.00% to 49.48%

Level Independent(Hard) Thresholding				
Wavelet	Level	Recovery (%)	Compression Ratio (%)	Threshold (absolute value)
db3	1	99.84	27.05	25.45
	2	99.83	26.35	25.45
	3	99.83	27.00	25.45
	4	99.89	25.35	25.45
	5	99.82	29.31	25.45
	6	99.76	42.57	25.45
db5	1	99.76	41.78	25.45
	2	99.75	41.99	25.45
	3	99.81	40.28	25.45
	4	99.77	43.69	25.45
	5	99.73	48.84	25.45
	6	99.73	47.76	25.45
db7	1	99.72	49.85	25.45
	2	99.78	46.53	25.45
	3	99.73	49.49	25.45
	4	99.72	53.11	25.45
	5	99.71	51.09	25.45
	6	99.71	51.78	25.45
haar	1	99.77	48.96	25.45
	2	99.72	52.86	25.45
	3	99.73	54.05	25.45
	4	99.74	52.28	25.45
	5	99.73	54.00	25.45
	6	99.76	49.31	25.45
sym2	1	99.72	54.52	25.45
	2	99.77	53.50	25.45
	3	99.79	52.07	25.45
	4	99.79	53.99	25.45
	5	99.77	49.48	25.45
	6	99.74	54.30	25.45

Table 5.2: Results-Level Independent Thresholding

Level Dependent(Soft) Thresholding				
Wavelet	Level	Recovery (%)	Compression Ratio (%)	Threshold (absolute value)
db3	1	99.87	25.00	23.47
	2	99.85	25.00	24.67
	3	99.86	25.00	23.19
	4	99.89	25.35	25.46
	5	99.89	25.17	20.80
	6	99.84	37.50	21.81
db5	1	99.84	37.17	23.17
	2	99.83	37.18	21.60
	3	99.85	37.50	24.00
	4	99.86	37.54	20.74
	5	99.81	43.52	21.82
	6	99.81	43.27	23.06
db7	1	99.82	42.77	20.48
	2	99.83	43.75	24.00
	3	99.83	43.73	21.21
	4	99.82	46.56	21.62
	5	99.79	46.11	23.27
	6	99.81	45.56	21.48
haar	1	99.79	47.57	24.00
	2	99.82	46.80	21.25
	3	99.82	47.90	22.00
	4	99.81	47.42	23.45
	5	99.82	47.14	21.36
	6	99.78	48.61	24.00
sym2	1	99.82	48.16	22.11
	2	99.83	48.73	22.79
	3	99.83	48.22	23.50
	4	99.85	47.66	21.61
	5	99.77	49.48	25.46
	6	99.82	49.01	22.65

Table 5.3: Results-Level Dependent Thresholding

5.4.2 Two Dimensional Data Compression

We arranged the data in 2 dimensions to study the wavelet compression for data collected over a week. Here is the data representation:-

Days	Time								
Day 1	144	84	180	84	132	108	108	96	...
Day 2	72	24	84	36	120	60	48	12	...
Day 3	36	12	96	48	12	48	72	48	...
Day 4	24	36	72	36	12	12	12	60	...
Day 5	96	12	60	60	48	12	36	36	...
Day 6	48	96	60	72	60	48	24	60	...
Day 7	90	60	48	108	60	120	60	60	...
...	...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...	...

Table 5.4: 2-D Data Arrangement

Our idea was to treat such data as a gray scale image and then apply image compression techniques over it. Here is the visualization of traffic volume data for 16 days.

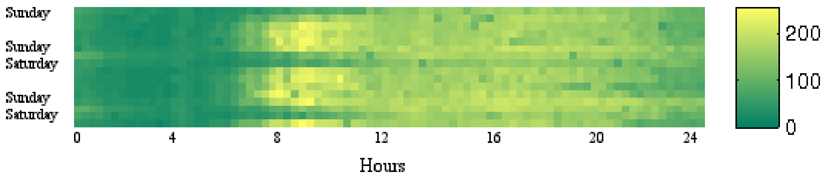


Figure 5.1: 2D-visualization

A measure of achieved compression is given by the compression ratio (CR) and the Bit- Per-Pixel (BPP) ratio. CR and BPP represent equivalent information. CR indicates that the compressed image is stored using CR % of the initial storage size while BPP is the number of bits used to store one pixel of the image. Reduction Ratio indicates the reduction in percentage, of the initial storage size.

We applied the same technique by arranging the data in 3-D fashion. Multiple 2-D data matrices were combined to do so. We observed that the compression performance for 3D was slightly better than for the 2D matrix arrangement.

5.5 Incident Detection

Traffic incidents are non periodic and pseudo random events causing traffic jam and hitting the overall performance of the road network. Probability of traffic incidents is higher during the peak hours. Many major cities in US have a traffic management system which includes traffic characteristic detectors and a centralized operations center for monitoring. These detectors comprise of video cameras, bluetooth sensors, ow detector sensors etc. They can capture traffic characteristics, such as traffic speed, occupancy and volume. However automatic incident detection techniques using these data are not widely used yet. Reliable and quick detection of incidents can prove very useful in incident management on roadways. Emergency crew can be sent on the incident location for obstruction clearance and medical help. It will also help to manage detour efficiently and better management of traffic and road network in case of an incident.

In this chapter we will focus on regression model for dichotomous data, i.e. logistic regression. This model is suitable when the outcome can takes only limited number of values, in our case only two, presence or absence of an incident. We presented a framework to use logistic regression after denoising the data by wavelet transform to predict incident in transportation systems. We tested the model on the historical traffic data and analyzed the reliability and robustness of the system.

Following figure shows the locations from where data was collected and analyzed.

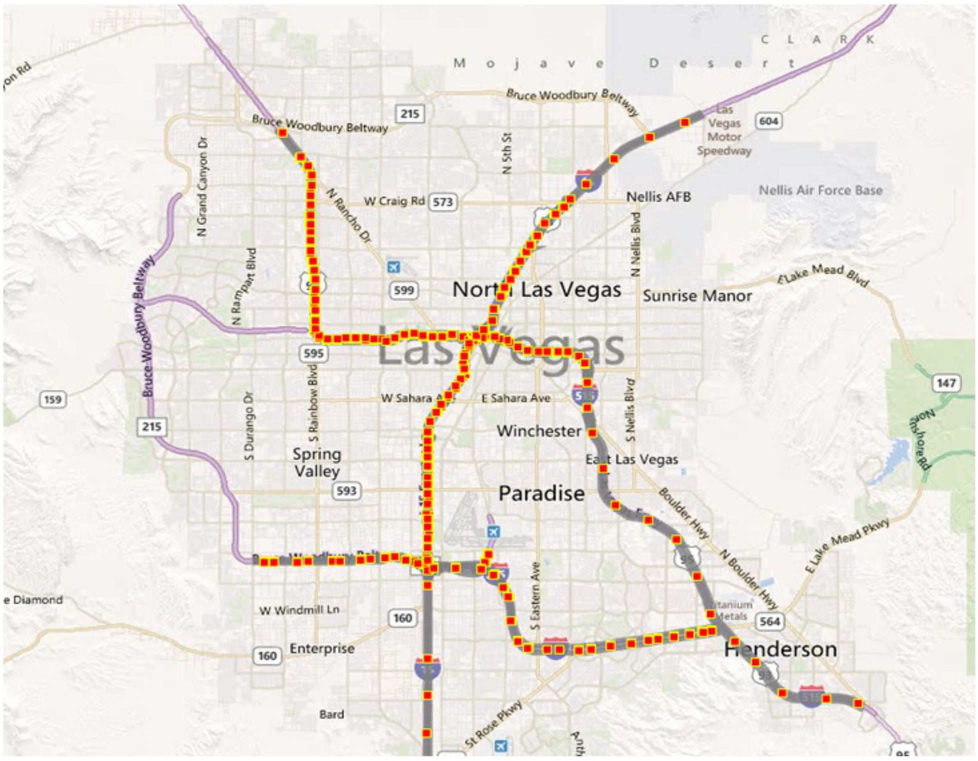


Figure 5.2: Traffic Sensors on freeway (Las Vegas Area)

Another database on FAST's website gives details about the time and location of the traffic incidents on the freeway. We analyzed the data and identified the location with maximum number of incidents as I-15 North bound, past Sahara (fig 5.3). We downloaded two separate files containing traffic parameters and traffic incidents during April 2012. We then combined the two data sets by matching the time stamps and formed one single database that looked like table 5.5.

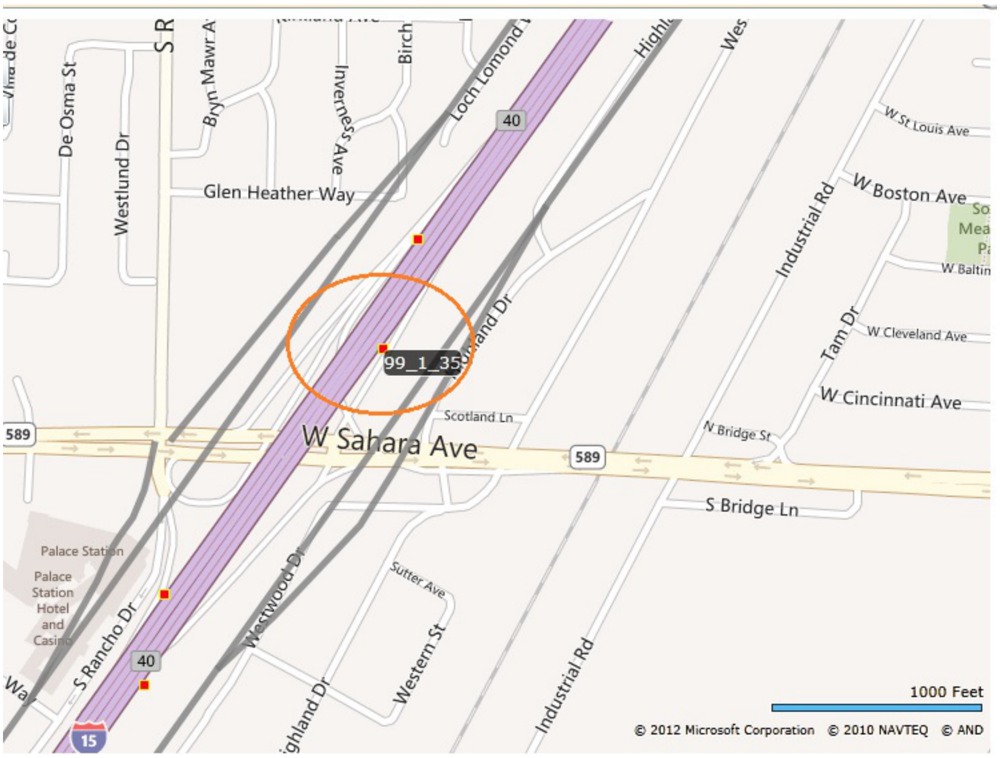


Figure 5.3: Identified Crash Site

Occupancy	Volume	Avg Speed	Incident
5	1848	61.6	0
5.8	1607	60.2	0
5.2	1840	63	0
5.2	1805	63.6	0
3.8	1945	36	1
4	1652	21.2	1
3.6	1744	23.8	1
3	1649	21	1
2.6	1770	22.4	1
3.2	1770	25.6	0
2.8	1866	24.6	0
2.4	1206	25.8	0
2.4	1474	28.6	0
2.4	1845	25.8	0
2	1971	27.8	0
2.4	1825	24.8	0
2.4	1858	24.6	0
2.6	1794	28.2	0
2.2	1919	26.2	0
1.6	1768	45.8	0
5	1959	49.8	0
5.8	1727	60	0
5.2	1692	65.6	0

Table 5.5: Dataset for Incident Detection

Table 5.6 gives frequencies of actual and predicted outcomes. Here, prediction value 1 means that system predicted the incident and value 0 means incident was not predicted.

	Predicted		
Actual	0	1	Total
0	78	6	84
1	4	11	15
Total	82	17	99

Table 5.6: Frequencies of actual and predicted outcomes

Table 5.7 and 5.8 give prediction success and prediction failure of the model.

Parameter	Description	Value (%)
Sensitivity	actual 1s correctly predicted	73.33
Specificity	actual 0s correctly predicted	92.85
Positive predictive value	predicted 1s that were actual 1s	64.70
Negative predictive value	predicted 0s that were actual 0s	95.12
Correct prediction	actual 1s and 0s correctly predicted	89.89

Table 5.7: Prediction Success

Parameter	Description	Value (%)
False pos. for true neg.	actual 0s predicted as 1s	7.14
False neg. for true pos.	actual 1s predicted as 0s	26.66
False pos. for predicted pos.	predicted 1s actual 0s	35.29
False neg. for predicted neg.	predicted 0s actual 1s	4.87
False predictions	actual 1s and 0s incorrectly predicted	10.10

Table 5.8: Prediction Failure

Table 5.9 gives the incident detection rate and false alarm rate for different of traffic parameters (using threshold probability as 0.5). We observe highest detection rate of 64.15% while using a combination of Volume + Speed + Occupancy.

Traffic Data	Incident Detection Rate (%)	False Alarm Rate (%)
Volume (Veh/h)	18.87	0.99
Occupancy (%)	50.94	5.59
Avg Speed (miles/h)	58.49	5.26
Avg Speed + Occupancy	60.38	4.61
Avg Speed + Volume	64.15	5.26
Occupancy + Volume	54.72	4.61
Avg Speed + Occupancy + Volume	64.15	5.26

Table 5.9: Incident Detection Results using Logit Models

Figure 5.4 shows the variation of incident detection rate against False alarm rate, as we vary the threshold probability from 1 to 0.

Table 5.10 shows incident detection results using logistic regression after preprocessing of the data by DWT.

It can be observed that the new hybrid model combining DWT and Logistic Regression, yields a better incident detection rate of 75% as compared to 64.15% using only logistic regression model. False alarm rate in this hybrid model is also on the lower side 4.14%, as compared to 5.26% in previous case. These results are for the combination all three traffic parameters i.e. Traffic Volume + Avg Speed +

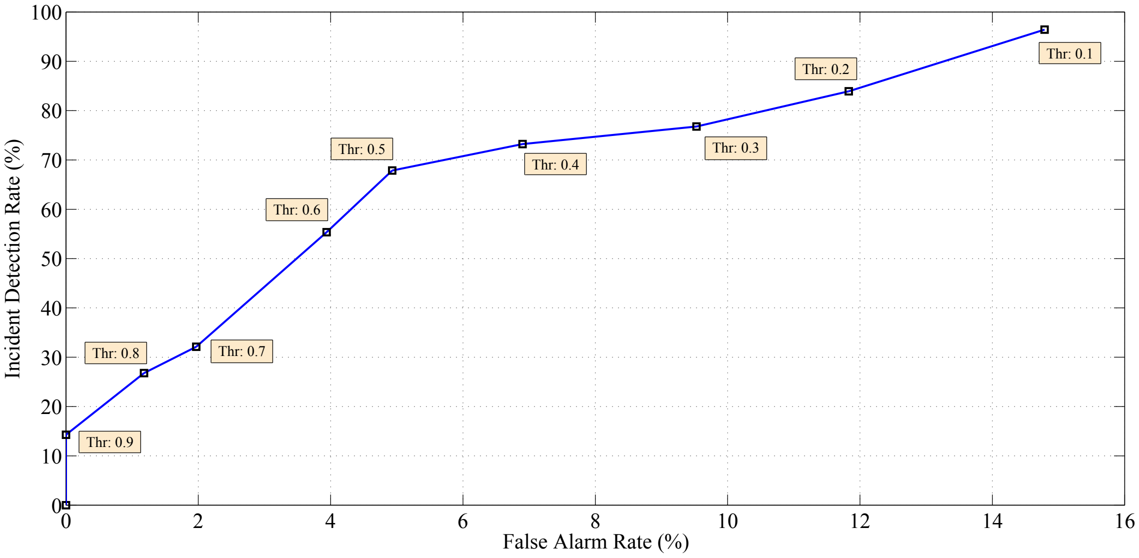


Figure 5.4: Incident Detection vs False Alarm Rate

Traffic Data	Incident Detection Rate (%)	False Alarm Rate (%)
Volume(Veh/h)	32.14	3.45
Occupancy (%)	60.71	6.90
Avg Speed (miles/h)	53.57	6.21
Avg Speed + Occupancy	71.43	4.14
Avg Speed + Volume	71.43	4.83
Occupancy + Volume	67.86	4.83
Avg Speed + Occupancy + Volume	75.00	4.14

Table 5.10: Incident Detection Results using DWT and Logit Models

Occupancy. However as clear from Tables 5.9 and 5.10, for each of the parameters and their combinations, preprocessing the data using DWT yields in better detection rate and lesser false alarm.

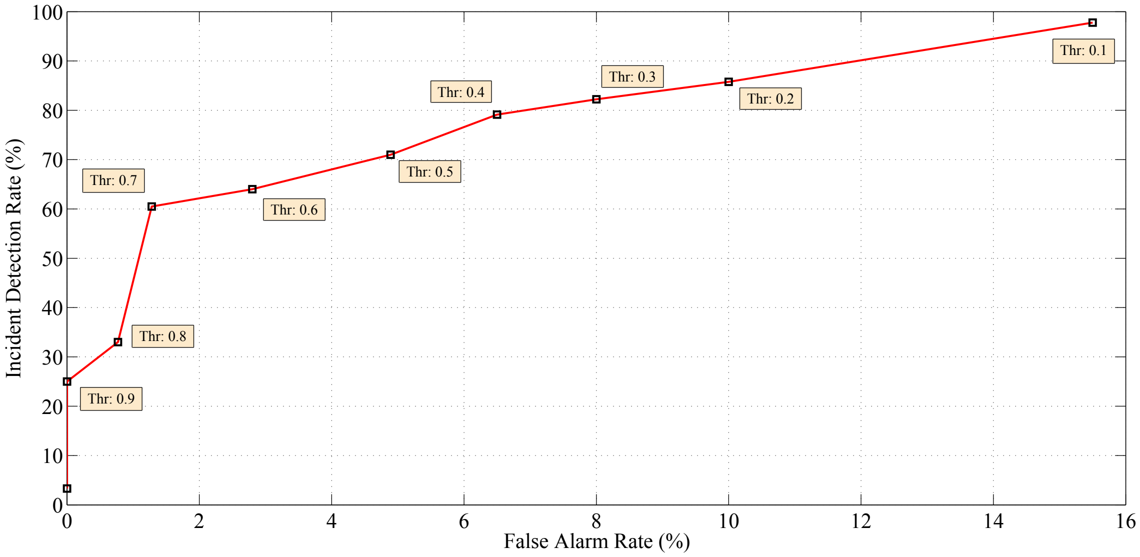


Figure 5.5: Incident Detection vs False Alarm Rate for filtered data

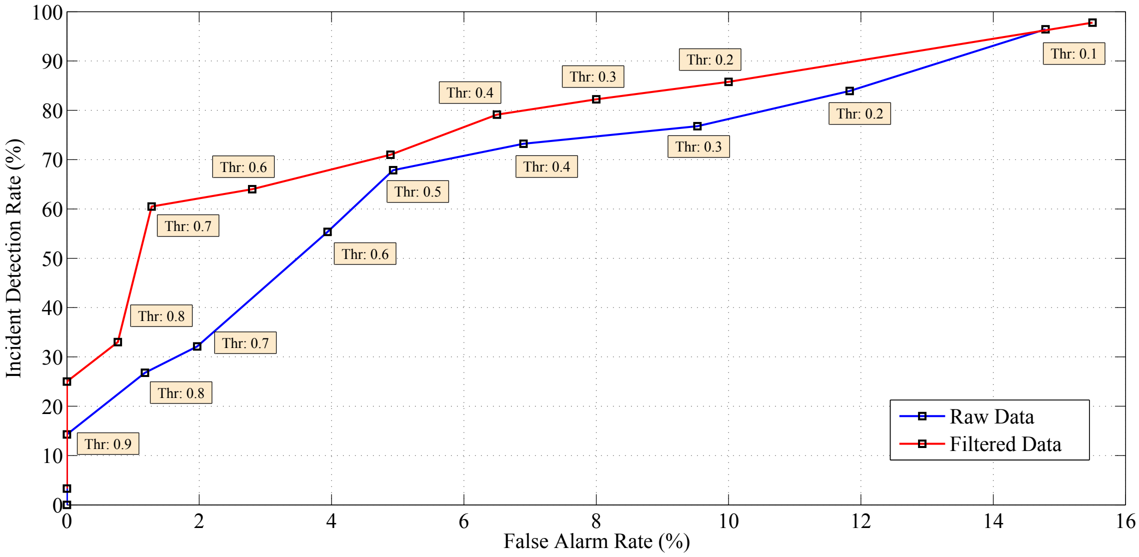


Figure 5.6: Comparison of Raw and Filtered Data

5.6 Conclusion

Logistic Regression technique was discussed and used for traffic incident detection. Various combinations of traffic parameters were tested and best detection rate of 64.15% (using 0.5 as default threshold) was observed for the combination of Traffic Volume + Occupancy + Avg Speed. Receiver Operating Characteristic (ROC) curves were plotted by varying the threshold, which showed a maximum of 95% detection rate for a 14% false alarm rate.

A new hybrid model was proposed combining two different computational approaches: wavelet transform and logistic regression. It was observed that detection rate improved with the new model (75%) and false alarm rate was also reduced to 4.14%. Receiver Operating Characteristic (ROC) curves were plotted by varying the threshold, and the curve was compared with the ROC curve of unfiltered data. It was observed that at each fixed false alarm rate, hybrid model gave a better incident detection rate.

CHAPTER 6

Qualitative Analysis

6.1 Summary

With enormous amount of linguistic data present on web, text analysis has become one of the major fields of interest today. This field includes sentiment analysis, information retrieval, text document classification, knowledge based modeling, content similarity measure, data clustering, words prediction/correction, decision making etc. Managing and processing such data has vital importance.

For this purpose, we collected the traffic related data from various sources including news articles and user comments and complaints which are in the form of raw text. The field being quite broad, our focus was mainly on text sentiment analysis and categorization/classification which can be further sub-divided in concept discovery and word sense disambiguation. Concept discovery is basically the method of extracting the actual concept/context in which the text is about. Word sense disambiguation is to find the correct sense in which a word is used. It is the basic necessity for discovering the concept.

A lot of research has been done in this field with major improvements. However, when it comes to short texts like user comments, the field still seems in nascent stage. Moreover most of the methods today require huge amount training corpus(database) arranging which is definitely a cumbersome task. We present a novel approach for word sense disambiguation which in turn allow us to find the context of the text. For this purpose we use the existing knowledge based semantic dictionary called WordNet. We have also proposed a model which provides us with a method to avoid the use of huge corpus and works for general context recognition.

6.2 Discovering Signal Related Issues

User complaint analysis is one of the best way of measuring the performance of any product. Same thing applies for our transportation complaint management system from which we can identify the performance of several traffic equipments and signals by analyzing the complaint data received by people. These measurements act like a feedback to the system and help in improving the performance in future. Complaints can be in the form of text data (example: sms or e-mails) or voice messages. In this report we have proposed a model for the implementation of such performance measurement system (PMS), which will be able to extract important information(features) from the complaint received by any user. From these features, complaint will be classified to a particular type, for example whether the complaint is signal related or not.

Our purpose is to make the system automatic such that it can extract the information by semantic analysis. We will also look for the emotion involved in a particular complaint. So we have also proposed a model which can be implemented to extract the information semantically.

6.2.1 Proposed Model

This model is based on supervised classification technique. Supervised classification requires an initial training data, from which machine learning algorithm generates a probabilistic model for decision making. Size of initial set of training data actually determines the efficiency of the model. To avoid this trade-off, we can make this model adaptive so that the efficiency of the model can improve with time.

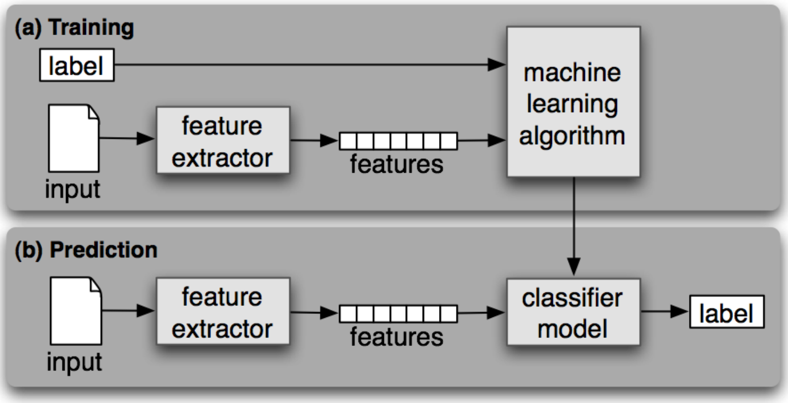


Figure 6.1: Complaint Analysis System Model (Copyright 2009 Steven Bird, Ewan Klein, and Edward Loper)

To implement this model, feature extractor needs to be defined first. Feature extractor is nothing but a codebook generator. This codebook consists of values of different features. For signal related issues, basic proposed features are as follows:-

1. Signal related words(example: Detection, Light, Green, Clock etc.)

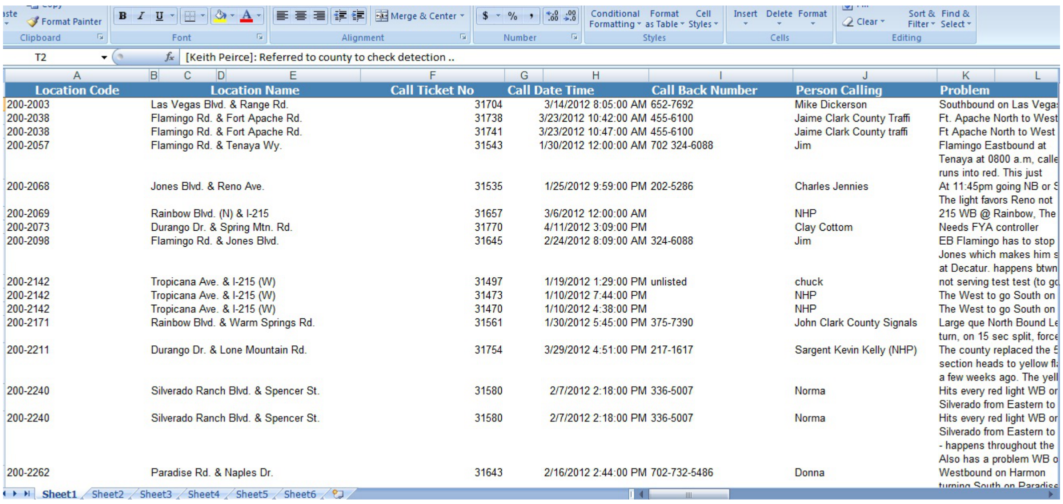
2. Nouns(this would allow us to extract any nouns within the sentence. Examples: Place name, Person Name etc.)

3. Timing Information

As we efficiently increase the number of features, system will become more robust. The other important part is the classifier, which needs to be trained with the help of a machine learning algorithm. After this training, classfier is ready to make decisions based on this prior information.

6.2.2 Data Description

The data representation of the complaint system is shown in the following figure:-



T2 [Keith Peirce]: Referred to county to check detection ..

Location Code	Location Name	Call Ticket No	Call Date Time	Call Back Number	Person Calling	Problem
200-2003	Las Vegas Blvd. & Range Rd.	31704	3/14/2012 8:05:00 AM	652-7692	Mike Dickerson	Southbound on Las Vega
200-2038	Flamingo Rd. & Fort Apache Rd.	31738	3/23/2012 10:42:00 AM	455-6100	Jaime Clark County Traffi	Ft. Apache North to West
200-2038	Flamingo Rd. & Fort Apache Rd.	31741	3/23/2012 10:47:00 AM	455-6100	Jaime Clark County traffi	Ft Apache North to West
200-2057	Flamingo Rd. & Tenaya Wy.	31543	1/30/2012 12:00:00 AM	702 324-6088	Jim	Flamingo Eastbound at Tenaya at 0800 a.m, calle runs into red. This just
200-2068	Jones Blvd. & Reno Ave.	31535	1/25/2012 9:59:00 PM	202-5286	Charles Jennies	At 11:45pm going NB or S The light favors Reno not
200-2069	Rainbow Blvd. (N) & I-215	31657	3/6/2012 12:00:00 AM		NHP	215 WB @ Rainbow, The
200-2073	Durango Dr. & Spring Mtn. Rd.	31770	4/11/2012 3:09:00 PM		Clay Cottom	Needs FYA controller
200-2098	Flamingo Rd. & Jones Blvd.	31645	2/24/2012 8:09:00 AM	324-6088	Jim	EB Flamingo has to stop Jones which makes him s at Decatur. happens btwn
200-2142	Tropicana Ave. & I-215 (W)	31497	1/19/2012 1:29:00 PM	unlisted	chuck	not serving test test (to go
200-2142	Tropicana Ave. & I-215 (W)	31473	1/10/2012 7:44:00 PM		NHP	The West to go South on
200-2142	Tropicana Ave. & I-215 (W)	31470	1/10/2012 4:38:00 PM		NHP	The West to go South on
200-2171	Rainbow Blvd. & Warm Springs Rd.	31561	1/30/2012 5:45:00 PM	375-7390	John Clark County Signals	Large que North Bound Le turn, on 15 sec split, force
200-2211	Durango Dr. & Lone Mountain Rd.	31754	3/29/2012 4:51:00 PM	217-1617	Sargent Kevin Kelly (NHP)	The county replaced the 5 section heads to yellow fl a few weeks ago. The yell
200-2240	Silverado Ranch Blvd. & Spencer St.	31580	2/7/2012 2:18:00 PM	336-5007	Norma	Hits every red light WB or Silverado from Eastern to
200-2240	Silverado Ranch Blvd. & Spencer St.	31580	2/7/2012 2:18:00 PM	336-5007	Norma	Hits every red light WB or Silverado from Eastern to - happens throughout the Also has a problem WB o
200-2262	Paradise Rd. & Naples Dr.	31643	2/16/2012 2:44:00 PM	702-732-5486	Donna	Westbound on Harmon turning South on Paradise

Figure 6.2: Complaint Data provided by FAST

6.2.3 Methodology

We have implemented our model in Python. We have used Naive Bayesian Classifier, which generates the probabilistic model based on conditional probabilities of different features.

In the phase of implementation, our analysis is based on single feature which is the extracted signal problem related word from a complaint. To implement this feature carefully, a frequency distribution analysis was done on 5 months complaint data. This analysis allowed us to prepare a list of words that are most frequent in signal related issues. So, classification of a complaint has been done on the basis on this feature. Frequenct distribution analysis was done for three sets of data(Signal Problems, Problem Comments and Work Performed). We obtained following distributions for 5 months of data-

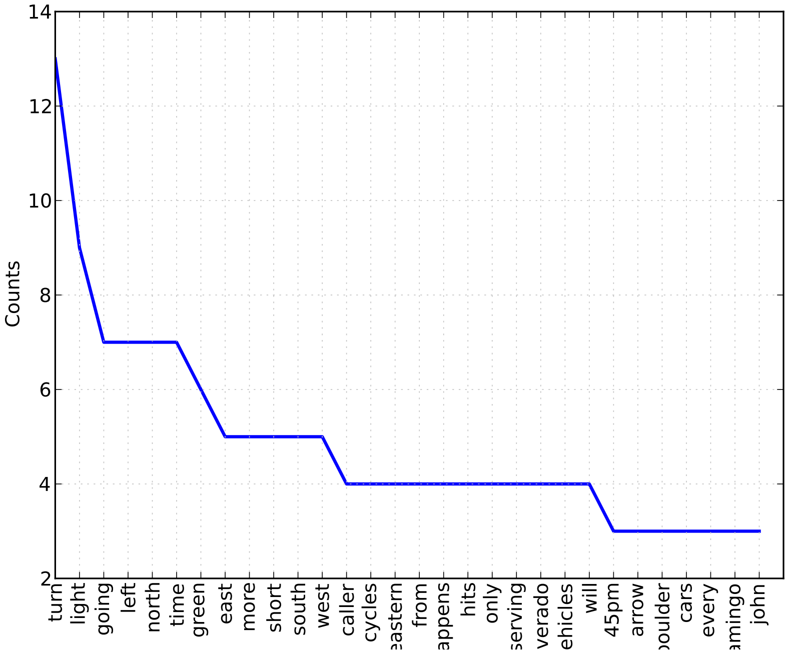


Figure 6.3: Words frequency distribution for Signal Related Problems

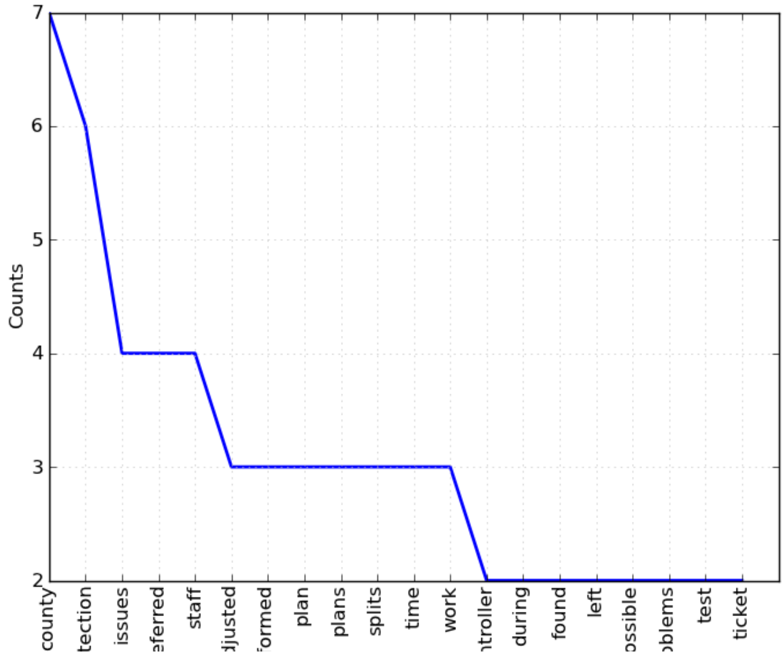


Figure 6.4: Words frequency distribution for Work Performed

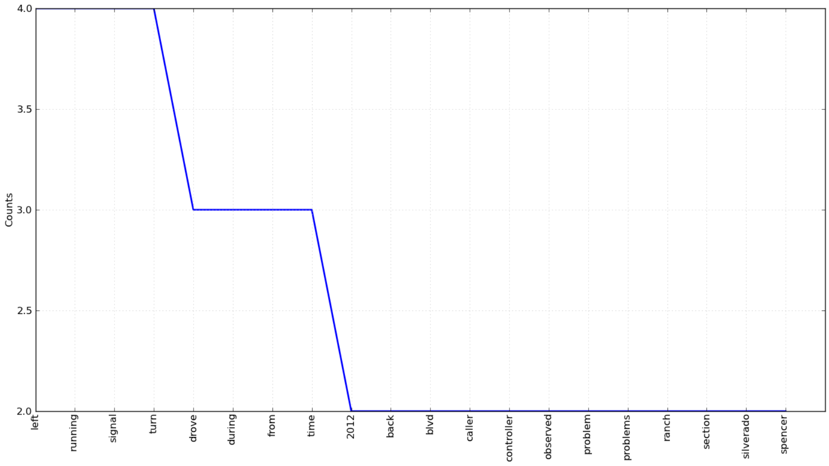


Figure 6.5: Word frequency distribution for Problem Comments

Besides, we also performed frequency analysis to find the distribution of number of days to acknowledge complaints and number of problems at different locations. Results are shown in the following figures:-

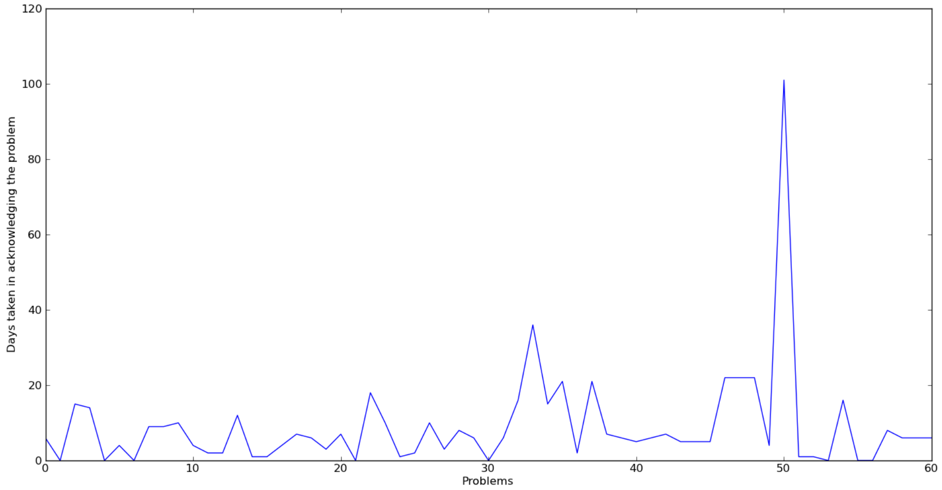


Figure 6.6: Number of days it took to acknowledge the issues

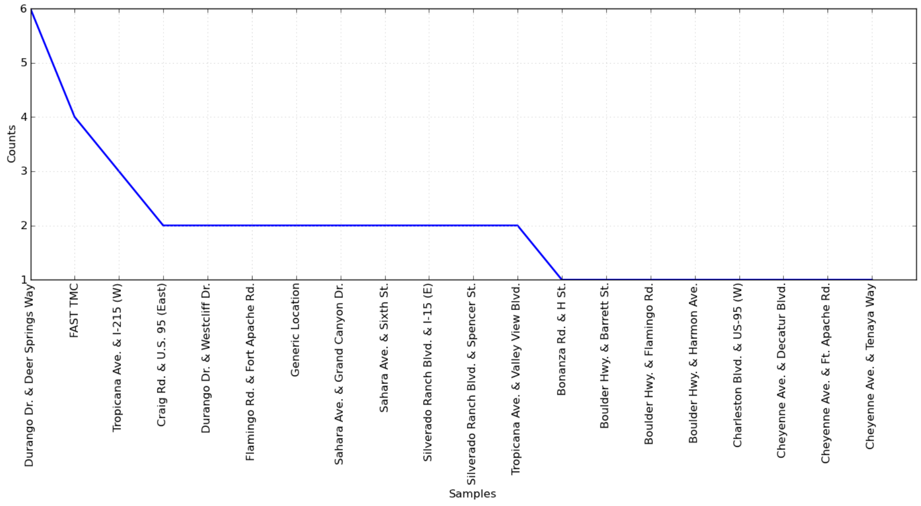
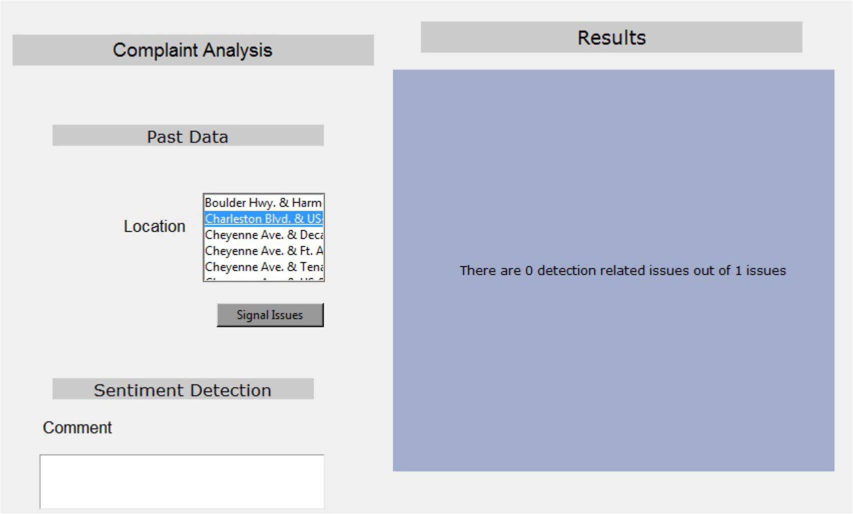


Figure 6.7: Number of Problems at a particular location

1. Maximum number of days to acknowledge an issue is 101.
2. Average number of days to acknowledge an issue is 9.
3. Minimum number of days to acknowledge an issue is 0.
4. Total number of problems are 74 from August 2011 to April 2012.
5. Durango Dr.and Deer springs way is the location which has come accross most number of issues.

After the basic frequency analysis, a training set was prepared from the data provided by RTC-FAST, and we made a dictionary of these sets with a label stating whether a complaint is signal related or not. After this training, we tested this model for both signal related and non-signal related issues.

We developed the Performance Ananlysis System(PAS) software which utilizes the methodology described above and tag the complaints as signal related or non signal related. Selecting a particular location we can find how many complaints at that location were actually signal related issues. The snapshot of the system is displayed in the following figure:-



There are 0 detection related issues out of 1 issues

Figure 6.8: Signal Related Issues

6.3 Sentiment Analysis

Sentiment analysis is done to extract the emotion in user comments or complaints. We performed this analysis by two methods and based on the efficiency of the methodology we integrated one of them in PAS.

First method utilizes the concepts of machine learning as described in the section above. As mentioned in the previous section, an efficient supervised machine learning decision making algorithm requires training database but unfortunately maintaining such tagged database requires higher cost and a lot of human intervention. Therefore, we used an existing twitter training database which is freely available for sentiment analysis. It has thousands of tweets which are tagged by one of these four sentiments:-

1. neutral

2. irrelevant

3. positive
4. negative

We trained the classifier with this dataset and tested the system on our data but unfortunately the system turned out to be quite inefficient the reason being the absence of any traffic related training data in the training set.

The second method utilized the list of positive and negative words. We built a list of all positive and negative sentiment words in English language and based on their presence we tagged the sentence with positive or negative emotion. We kept count of total positive words and total negative words in a sentence and based on the that gave a score(0.5 for each count) to both these emotions. If the score of positive emotion turned out be more than negative emotion we assigned the sentence positive and if it is less we assigned to it negative emotion. In case of equal(or 0) scores we tagged the statement as neutral.

6.4 Word Sense Disambiguation

In English language, almost every word has multiple senses in which it can be used. For example, a word traffic can be used in trafficking sense as well as in the sense of movement of vehicles. With huge amount of linguistic data present on web it is of utmost importance to automatically extract the relevant data for analysis. Here we provide a noval approach to analyze the raw text data semantically which means the process of analyzing text by its meaning. We used a lexical semantic dictionary known as WordNet for our algorithm.

6.4.1 WordNet

WordNet was created under the direction of Professor George A. Miller at the Cognitive Science laboratory in Princeton University[6]. It is lexical database which basically groups English language words in the sets of synonyms called **synsets**. A word can be in multiple synsets depending on the variation of meanings it has. It also keeps track of the various relationships which these synsets have semantically. For example, some of synstes of word **book** are the following:-

1. **Synset('book.n.01')**- a written work or composition that has been published (printed on pages bound together).

2. **Synset('book.n.06')**- a collection of playing cards satisfying the rules of a card game.

3. **Synset('book.n.07')**- a collection of rules or prescribed standards on the basis of which decisions are made.

4. **Synset('book.v.01')**- engage for a performance.

Now, these are just four of the fifteen synsets the word **book** has. It can be easily figured out that the letter **n** in the synset refers to **Noun** and the letter **v** refers to **Verb**. So, it distinguishes between noun, verb, adjective and adverbs. Moreover, the dictionary provides the short definition or sense in which the word is kept in that synset.

Story does not end here. All these synsets are arranged in a network such that they follow an IS-A hierarchy (similar to the hierarchy which is followed in object oriented programming language like JAVA). For example, everything is an **object** and **object** can further be classified into **abstract** and **physical object**. So it is a tree of concepts. A **dog** is an **animal** and **cat** is also an **animal** but **cat** is not a **dog**. So, **cat** and **dog** do not follow IS-A hierarchy but they follow the same with **animal**.

This kind of network is quite useful in estimating a semantic distance between two words. Question to think:- does it really help in estimating the distance between meanings ? The answer to this question is yes because, it helps in estimating the distance between synsets not the words. So, now the main problem is to correctly find the most relevant synset for a word in a given sentence.

Besides this IS-A hierarchy WordNet also maintains a number of other semantic relations. Some of which are described here:-

1. **Hypernyms** - All parent synsets are called hypernyms of child synsets in IS-A hierarchy.
2. **Hyponyms** - All child synsets are called hyponyms of parent synsets in IS-A hierarchy.
3. **Holonyms** - A synset **A** is a holonym of synset **B** if **B** is a part of **A** (Tree is a holonym to branch).
4. **Meronym** - A synset **A** is a holonym of synset **B** if **A** is a part of **B** (Branch is a meronym to tree).

Although all these informations in the WordNet make it quite powerful even then it has lot of limitations. The first and foremost being the non-relatedness between

different POS. WordNet does not have very strong relations between different parts of speeches. For example:- Adverb **costly** and Noun **cost** do not have any network bond that is we cannot find similarity value between these these words, although they seem quite similar to each other.
The other problem with WordNet is it incompleteness. Incompleteness in the sense of etymological information which most of the dictionaries poses. It also does not include the information about pronunciation and forms of irregular verbs.

6.4.2 Proposed Method

We propose a completely unsupervised hypothesis for context recognition for short text(sentence) by word sense disambiguation which doesn't require any prior corpus knowledge. Our main idea is that in general statements the informative words have strong relatedness. For example: In sentence, **He is sitting on the river bank**, the information word river will be closely related to the word bank which has **river_bank** sense not a **financial_company** sense. Wordnet which is the lexical knowledge based semantic network comes out as a savior in this task. It has ISA(is-a) hierarchy of Nouns(which actually are concepts) and hierarchy for Verbs as well. We utilize power of this semantic relation in our task. Based on the recognition of context for short text we have also proposed the context recognition for documents, in which we parse the document in short texts or sentences, determine the context of each of this chunk and then look for the strongest relationship between these contexts. Here is the propesd architecture from a high level :-

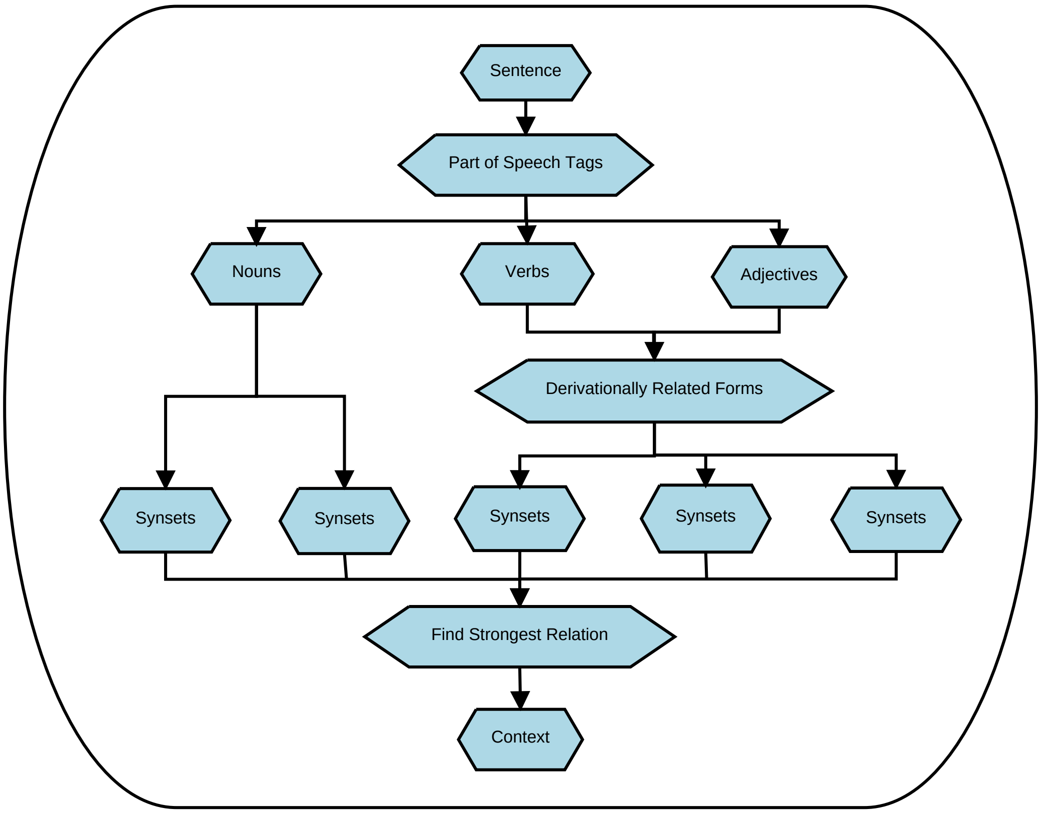


Figure 6.9: Context Determination Model for a Sentence

The following figure explains the method in detailed manner. Each word has many synsets(meanings) and in a sentence we are trying to find the closest relation between all combinations of these meanings. Closest relation gives the exact meaning in which the words are used.

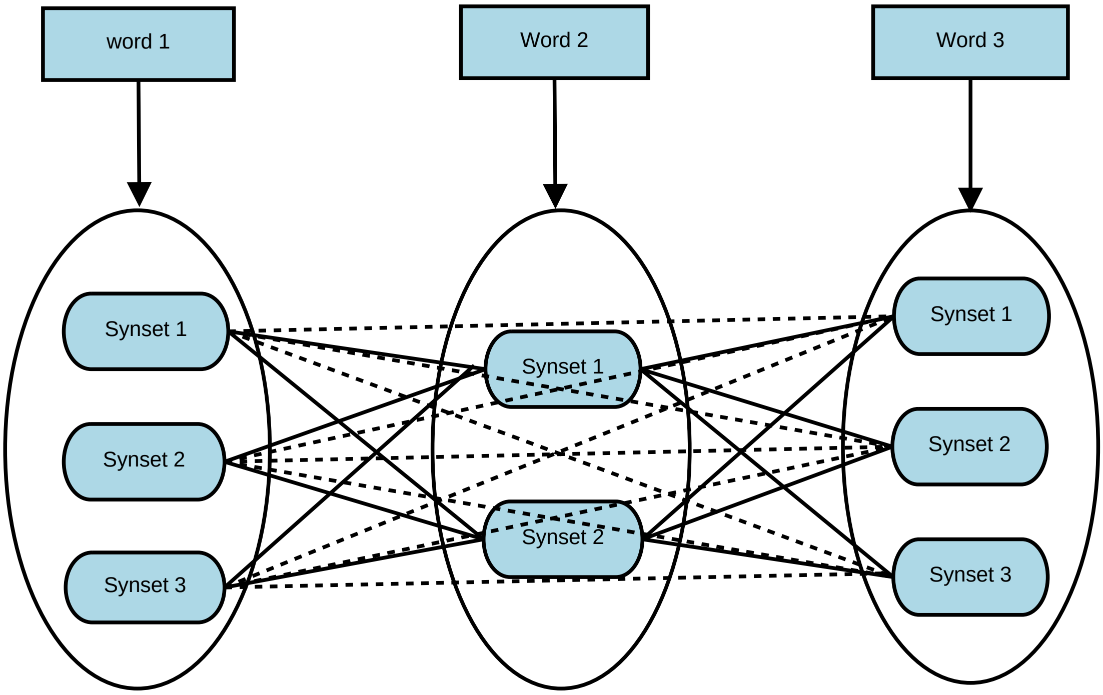


Figure 6.10: Synsets Network

This architecture is based on Wordnet network which has few issues. For example, Verbs hierarchy is not very well formed in Wordnet. Wordnet also does not have any semantic hierarchy for Adjectives, Adverbs or other parts of speech. So, it becomes difficult for context recognition of sentences like **This thing is costly**, where actual context is an adverb. So, to eliminate such issue we have used a technique which estimates **Derviationally Related Forms** of such POS. So, for example, word **costly** has the derivationally related form **Cost** or **Price**.

So, we convert all words which are not Nouns into it's closest derivationally related Noun and then estimate the closeness between those Nouns or contexts. Here are the steps that are involved:

6.4.2.1 Step 1

This is the preprocessing step. Each word of the sentence should be tagged with its POS. We used Python NLTK for this purpose. Then non informative function words like prepositions should be removed. So that we have a list of tuples having information words and their POS.

6.4.2.2 Step 2

In this step, we convert all non-Noun information words to their respected derivationally related Noun synset. WordNet provides list of all lemmas with POS tags which are derivationally related to the input word. For convenience, we chose the first Noun lemma that occurs in the list.

6.4.2.3 Step 3

Following Step 2, we have the list of all noun words and synsets of other derivationally related forms of different POS. Now, a synset list is made corresponding to each word. Then WUP similarity is calculated between synsets of different words. We made all combinations of synsets from different words and calculated WUP similarity for each combination. Then we find the summation of similarity measures between each synset in a combination. The combination which has the maximum value for this summation disambiguate the word sense and thus we find the exact concept for that sentence.

6.4.3 Test Setup

Here we describe the test setup we developed for testing our algorithm. We chose five words at random having disambiguous senses depending the sentences. Words chosen for this purpose were:-

1. Turn
2. Traffic
3. State
4. Draw
5. Film

For each word in the list above, we randomly make a dataset of few sentences in which they have different meanings. We apply our algorithm for each sentence having that word and give a score(0 or 1) manually. If the definition of selected synset is relevant to the sentence we give a score 1 if not we give a score 0.

We applied the algorithm over all sentences. Based on the score we estimate the percentage of 1's which is nothing but the percentage of accuracy of the algorithm. Here are the example dataset developed:-

SENTENCE	SYNSET	DEFINITION OF SYNSET	POS	SCORE
That was his turn to spin the wheel in the game.	Synset('turn.n.03')	the activity of doing something in an agreed succession	Noun	1
Please make a left turn from next signal.	Synset('turn.n.02')	the act of changing or reversing the direction of the course	Noun	1
Turn right at the junction to cross the bridge.	Synset('bend.n.01')	a circular segment of a curve	Noun	1
Turned out to be worse than any of us imagined.	Synset('turning.n.04')	a movement in a new direction	Verb	1
It is the turn of the next player to bat.	Synset('turn.n.03')	the activity of doing something in an agreed succession	Noun	1
Drive slow on the next turn.	Synset('turn.n.09')	a division during which one team is on the offensive	Noun	0
Next team will score higher when their turn comes.	Synset('turn.n.06')	the act of turning away or in the opposite direction	Noun	0
Please turn off the lights.	Synset('turn.n.09')	a division during which one team is on the offensive	Noun	0
I turn_down the job offer.	Synset('nonacceptance.n.01')	the act of refusing an offer	Verb	1
They turn_off the lights of the room.	Synset('turnoff.n.02')	a side road where you can turn off	Verb	0
There is sharp turn ahead on the highway.	Synset('bend.n.01')	a circular segment of a curve	Noun	1

Table 6.1: Sense Disambiguation for word TURN

SENTENCE	SYNSET	DEFINITION OF SYNSET	POS	SCORE
Web links communication generate traffic for your site.	Synset('traffic.n.03')	the amount of activity over a communication system during a given period of time	Noun	1
Human traffic is the illegal trade of people.	Synset('traffic.n.02')	buying and selling; especially illicit trade	Verb	0
Network choked because of huge data traffic.	Synset('traffic.n.01')	the aggregation of things (pedestrians or vehicles) coming and going in a particular locality during a specified period of time	Adj	0
Data pipelining is important for managing traffic in communications.	Synset('traffic.n.03')	the amount of activity over a communication system during a given period of time	Adj	1
Traffic in rush hours cause major delays.	Synset('traffic.n.03')	the amount of activity over a communication system during a given period of time	Adj	0
Travel time analysis is important to maintain traffic.	Synset('traffic.n.03')	the amount of activity over a communication system during a given period of time	Adj	0
Car accident caused huge traffic jam on the highway.	Synset('traffic.n.03')	the amount of activity over a communication system during a given period of time	Noun	0
Traffic is the major problem in big cities.	(Synset('traffic.n.01')	the aggregation of things (pedestrians or vehicles) coming and going in a particular locality during a specified period of time	Adj	1
Communication traffic is the major problem in big cities.	Synset('traffic.n.02')	buying and selling; especially illicit trade	Verb	0
These two lanes merge on next traffic signal.	Synset('traffic.n.03')	the amount of activity over a communication system during a given period of time	Noun	0
Traffic jam caused by all the solitary car drivers.	Synset('traffic.n.01')	the aggregation of things (pedestrians or vehicles) coming and going in a particular locality during a specified period of time	Adj	1
Web links communication generate traffic for your site.	Synset('traffic.n.03')	the amount of activity over a communication system during a given period of time	Noun	1

Table 6.2: Sense Disambiguation for word TRAFFIC

SENTENCE	SYNSET	DEFINITION OF SYNSET	POS	SCORE
His mental state was not good when he decided to leave.	Synset('state.n.02')	the way something is with respect to its main attributes	Noun	1
Nation states grew into empires, with ever more power.	Synset('state.n.04")	a politically organized body of people under a single government	Noun	0
Water can be found in all the three states of matter.	Synset('state_of_matter.n.01')	(chemistry) the three traditional states of matter are solids (fixed shape and volume) and liquids (fixed volume and shaped by the container) and gases (filling the container)	Noun	1
States of consciousness.	Synset('state.n.02')	the way something is with respect to its main attributes	Noun	1
State legislature should pass these law also.	Synset('state.n.04')	a politically organized body of people under a single government	Noun	1
State of serious disrepair over the last few decades .	Synset('state.n.02')	the way something is with respect to its main attributes	Noun	1
State of the art labeling can place in a single process.	Synset('state.n.02')	the way something is with respect to its main attributes	Noun	1
This is not a citizenship bequeathed to people by a sovereign state.	Synset('state.n.02')	the way something is with respect to its main attributes	Noun	0
The patient's mental state was not improved in the hospital.	Synset('state.n.02')	the way something is with respect to its main attributes	Noun	1
Nevada state's population is almost equal to Wyoming's.	Synset('state.n.01')	the territory occupied by one of the constituent administrative districts of a nation	Noun	1

Table 6.3: Sense Disambiguation for word STATE

6.4.4 Conclusions

Our Tests resulted in 66.03% of 1's as compared to score 0 which is nothing but the success rate of our algorithm. Thus we can predict the concept of any sentence with the information present in it taking the help of semantic dictionary like WordNet. Furthermore, it does not require training dataset to disambiguate word senses in a sentence.

BIBLIOGRAPHY

[1] WISQARS, *Leading Causes of Death Reports*, (2007), <http://webappa.cdc.gov/sasweb/ncipc/leadcaus10.html>,(2011).

[2] *Injury Prevention & Control: Motor Vehicle Safety*, <http://www.cdc.gov/motorvehiclesafety/>,(07/10/2011).

[3] Kai-Bo Duan and S. Sathiya Keerthi, *Which Is the Best Multiclass SVM Method? An Empirical Study*, BioInformatics Research Center, Nanyang Technological University.

[4] Yun Qian Miao and Mohamed Kamel, *Pairwise Optimized Rocchio Algorithm for Text Categorization*, Pattern Recognition Letter 32:375-382, (2011)

[5] Manuel de Buenaga Rodriguez, Jose Mar?a Gomez-Hidalgo, Belen D?az-Agudo, *Using WordNet to Complement Training Information in Text Categorization*, Departamento de Sistemas Informaticos y Programacion, Universidad Complutense de Madrid, (Septemeber 1997).

[6] G. Miller, *Special Issue, WordNet: An on-line lexical database.* International Journal of Lexicography, 3(4) (1990)

[7] E. Agirre, G. Rigau, *Word sense disambiguation using conceptual density.* Proc. of COLING (1996)

[8] R. Richardson, A. Smeaton,*Using wordnet in a knowledge-based approach to information retrieval.* Proc. of the BCS-IRSG Colloquium, Crewe (1995)

[9] E. Agirre, G. Rigau, L. PADRO and J. ATSERIAS, *Combining Supervised and Unsupervised Lexical Knowledge Methods for Word Sense Disambiguation*, Computers and the Humanities, Vol. 34, No. 1/2, (2000)

[10] Rocchio, J.,*Relevance feedback in information retrieval.* In: Salton, G. (Ed.), The Smart Retrieval System-Experiments in Automatic Document Processing. Prentice Hall, pp. 313323. (1971)

[11] Ittner, D.J., Lewis, D.D., Ahn, D.D.,*Text categorization of low quality images.* In: Proc. SDAIR-95, 4th Annual Symposium on Document Analysis and Information Retrieval, pp. 301315. (1995)

[12] Landauer, T.K. and Dumais, S.T. .*Latent semantic analysis and the measurement of knowledge.* In R. M. Kaplan and J. C. Burstein (Eds) Educational Testing Service Conference on Natural Language Processing Techniques and Technology in Assessment and Education. Princeton, Educational Testing Service (1994)

Nevada Department of Transportation
Rudy Malfabon, P.E. Director
Ken Chambers, Research Division Chief
(775) 888-7220
kchambers@dot.nv.gov
1263 South Stewart Street
Carson City, Nevada 89712