



UNIVERSITY OF THESSALY  
SCHOOL OF ENGINEERING  
DEPARTMENT OF MECHANICAL ENGINEERING

Master Thesis

**CRIME STATISTICAL ANALYSIS AND PREDICTIVE POLICING:  
THE CASE STUDY OF VOLOS**

by

**AFRODITI TEMOURTZIDOU  
GARYFALLIA TZANIDAKI  
SOKRATIS CHARISIS**

Submitted to fulfill part of the requirements  
for the acquirement of the Master Diploma of Mechanical Engineer

© 2018 Afroditi Temourtzidou, Garyfallia Tzanidaki, Sokratis Charisis

The approval of the thesis by the Department of Mechanical Engineering, School of Engineering, University of Thessaly does not imply the acceptance of the authors' opinions (Law 5343/32 Article 202 § 2).

**Approved by the Members of the Examination Committee:**

First Examiner      Dr. Athanasios Ziliaskopoulos  
(Supervisor)      Professor, Department of Mechanical Engineering, University of  
Thessaly

Second Examiner      Dr. George Liberopoulos  
Professor, Department of Mechanical Engineering, University of  
Thessaly

Third Examiner      Dr. Dimitris Pantelis  
Associate Professor, Department of Mechanical Engineering,  
University of Thessaly

## **Acknowledgements**

First of all, we would like to thank our supervisors, Professor Dr. Athanasios Ziliaskopoulos and Adjunct Lecturer Dr. Athanasios Lois, for their valuable guidance and support during the research for this thesis. We would also like to thank especially the Police Office Director of the Police Department of Magnesia Mr. Vasilios Markogiannakis and Police Office Sub-director of the Police Department of Magnesia Mr. Vasilios Karaiskos for their reliance on our integrity and their valuable input by providing confidential police data and information for the completion of this thesis. Moreover, we would like to thank all the Police Officers of the Police Department of Volos for their assistance and cooperation, and especially Police Sergeant Christos Platis, Police Sergeant Stavros Kampouranis and Police Warrant Officer Dimitrios Karagiorgis for their participation and assistance in the real life case scenario that was filmed for the purpose of this thesis. To conclude, we would like to express our gratitude to all professors of the department of Mechanical Engineering in University of Thessaly for the knowledge that they provided with during the postgraduate studies in «State-of-the-Art Design and Analysis Methods in Industry» program.

Above all, we are grateful to our families for their wholehearted love, ethical and financial support during our postgraduate studies. We dedicate this thesis to ourselves and our families.

Afroditi Temourtzidou

Garyfallia Tzanidaki

Sokratis Charisis

# **CRIME STATISTICAL ANALYSIS AND PREDICTIVE POLICING: THE CASE STUDY OF VOLOS**

AFRODITI TEMOURTZIDOU

GARYFALLIA TZANIDAKI

SOKRATIS CHARISIS

University of Thessaly, Department of Mechanical Engineering, 2018

Supervisor: Dr. Athanasios Ziliaskopoulos, Professor in Optimization of Production and  
Transportation Systems

## **Summary**

Law enforcement work is frequently reactive. Police officers respond to calls for service, investigate crime incidents and make arrests. Today more than ever, law enforcement work needs to be also proactive.

In proactive policing, law enforcement uses historical data and analyzes patterns to understand the nature, the occurrence and the frequency of crime incidents. Police officers devise strategies and tactics to prevent or mitigate future occurrence of incidents. They evaluate results and revise practices to improve policing. In addition to that, departments may combine an array of data with street intelligence and crime analysis to produce better assessments about what might happen next if they take various actions.

According to these initiatives, in this thesis we study and analyze historical crime incident data for the city of Volos and extract useful conclusions by applying statistical and data analysis. Moreover, facility location and routing algorithms are studied and implemented in order to come to conclusions regarding points where police officers should do patrol and how to plan the route of a police vehicle in case of a day with high volume of criminality. Moving one step forward, we apply prediction algorithms based on Kalman filter to forecast the number of future crime incidents in different sectors of the city of Volos. The outcomes of this study presented in this dissertation aim at providing insights in the occurrence of crime incidents and how police officers can server them more efficiently.

# Contents

Introduction	1
1. CRIME ANALYSIS	3
1.1 Introduction to Crime Analysis and History Review	3
1.2 Types of Crime Analysis	9
1.3 Technology of Crime Analysis	11
2. PROPOSED METHODOLOGY	13
2.1 Data and Statistical Analysis	13
2.1.1 Definition of Data Processing Cycle	13
2.1.2 Data Interpretation	16
2.1.3 Statistical Analysis of Data	17
2.1.4 Spatial Data Analysis	19
2.1.5 Big Data Analysis	21
2.2 Prediction Techniques	24
2.2.1 Predictive Analytics and its Process	24
2.2.2 Predictive Analytic Models	26
2.2.3 Predictive Analytic Models Applications and Techniques	27
2.3 Cluster Analysis	32
2.3.1 Clustering definition and stages	32
2.3.2 Clustering Techniques	33
2.3.3 Types of Cluster	35
2.4 Routing Problems	37
2.4.1 Introduction to Routing Problems	37
2.4.2 Routing Problem Type	39
2.4.3 The Dial-A-Ride Problem	42
2.4.4 Vehicle Routing Problem	45
3. THE CASE STUDY OF VOLOS	51
3.1 The Police Department of Volos	51
3.2 Police Data	54
3.2.1 Type and Format of Data	54
3.2.2 Validation and Cleaning of Data	56
3.2.3 Transformation and Processing of Data	57
3.2.4 Mapping and Clustering of Data	59
3.3 Statistical experiments and comments on the results	62
3.4 Prediction of Feature Incidents	73
3.4.1 Splunk Machine Learning Toolkit	74
3.4.2 Preparation of data inputs	76
3.4.3 Prediction algorithms	77
3.4.4 Implementation of Prediction algorithms	79
3.4.4.1 All incident data for sector 1	82
3.4.4.2 All incident data for sector 2	84
3.4.4.3 All incident data for sector 3	87
3.4.4.4 Burglaries for sector 1	89
3.4.4.5 Burglaries for sector 2	92
3.4.4.6 Burglaries for sector 3	94
3.4.4.7 Car Robberies for sector 1	97
3.4.4.8 Car Robberies for sector 2	99

3.4.4.9 Car Robberies for sector 3	102
3.4.4.10 Motorcycle Robberies for sector 1	104
3.4.4.11 Motorcycle Robberies for sector 2	104
3.4.4.12 Motorcycle Robberies for sector 3	107
3.5 Estimation of the Center Point of the Busiest Day Graph Network	110
3.5.1 Busiest Day	110
3.5.2 Minimum Spanning Tree of the Busiest Day Graph Network	112
3.5.3 Estimation of the Absolute Center Point for Busiest Day Graph Network	113
3.5.3.1 Absolute Center Algorithm	114
3.5.3.2 Center Point of the Graph Network of Busiest Day	115
3.6 The Dial -A- Ride Problem (DARP)	118
3.7 Estimation of Medians of City Sectors	125
3.7.1 City Sectors and Incidents Estimations	125
3.7.2 Median Problem	126
3.7.2.1 Problem Description	127
3.7.2.2 Solution Algorithm	128
3.7.3 Estimation of the Medians of the City Sector	128
3.7.3.1 Median Point of Sector p.25	129
3.7.3.2 Median Point of Sector p.38	132
4. CONCLUSIONS	134
5. FUTURE RESEARCH (ARTIFICIAL INTELLIGENCE)	146
5.1 Definition of Artificial Intelligence	146
5.2 Types of Artificial Intelligence	146
5.3 History Review of Artificial Intelligence	148
5.4 The Connection between Artificial Intelligence Machine Learning and Data Science	149
5.5 Contribution of Artificial Intelligence in the field of Predictive Policing	153
6. BIBLIOGRAPHY	
6.1 Greek Bibliography	158
6.2 Foreign Bibliography	158
6.3 Electronic Bibliography	161
7. APPENDIX	162
7.1 Calculation of the Rates of Reduction / Increase of offenses for a specific year compared to the previous year	162
7.2 Charts for the type of the offenses for every month over the period 2010-2017	163
7.3 Command used for prediction in Splunk Machine Learning Toolkit	169
7.4 Number of incidents in city sectors	170
7.5. Minimum distances matrices for sectors p25 and p38	172

## List of Tables

Table 3.1: The responsibilities of the police services (P.D. 7/2017 Articles 95 & 97)	52
Table 3.3.1: Statistical experiments related to theft offenses in Volos for the time period 2010-2017	63
Table 3.3.2: Increase/Reduction of offenses compared to the previous year	65
Table 3.3.2: Order of hours based on the frequency of offenses	69
Table 3.4.4.1-1: Results of LL algorithm for input data “All incidents data sector 1” for various future timespan and holdback values	79
Table 3.4.4.1-2: Results of LLT algorithm for input data “All incidents data sector 1” for various future timespan and holdback values	80
Table 3.4.4.1-3: Results of LLP algorithm for input data “All incidents data sector 1” for various future timespan and holdback values	80
Table 3.4.4.1-4: Results of LLP algorithm for input data “All incidents data sector 1” for various future timespan and holdback values	80
Table 3.4.4.1-5a: Forecasted values regarding all crime incidents for 2018 with confidence interval 95%	81
Table 3.4.4.2-1: Results of LL algorithm for input data “All incidents data sector 2” for various future timespan and holdback values	82
Table 3.4.4.2-2: Results of LLT algorithm for input data “All incidents data sector 2” for various future timespan and holdback values	82
Table 3.4.4.2-3: Results of LLP algorithm for input data “All incidents data sector 2” for various future timespan and holdback values	83
Table 3.4.4.2-4: Results of LLP5 algorithm for input data “All incidents data sector 2” for various future timespan and holdback values	84
Table 3.4.4.2-5: Forecasted values regarding all crime incidents for 2018 for sector 2 with confidence interval 95%	84
Table 3.4.4.3-1: Results of LL algorithm for input data “All incidents data sector 3” for various future timespan and holdback values	84
Table 3.4.4.3-2: Results of LLT algorithm for input data “All incidents data sector 3” for various future timespan and holdback values	85
Table 3.4.4.3-3: Results of LLP algorithm for input data “All incidents data sector 3” for various future timespan and holdback values	85
Table 3.4.4.3-4: Results of LLP5 algorithm for input data “All incidents data sector 3” for various future timespan and holdback values	85
Table 3.4.4.3-5: Forecasted values regarding all crime incidents for 2018 for sector 3 with confidence interval 95%	86
Table 3.4.4.4-1: Results of LL algorithm for input data “Burglaries for sector 1” for various future timespan and holdback values	87
Table 3.4.4.4-2: Results of LLT algorithm for input data “Burglaries for sector 1” for various future timespan and holdback values	87
Table 3.4.4.4-3: Results of LLP algorithm for input data “Burglaries for sector 1” for various future timespan and holdback values	87
Table 3.4.4.4-4: Results of LLP5 algorithm for input data “Burglaries for sector 1” for various future timespan and holdback values	88
Table 3.4.4.4-5: Forecasted values regarding burglaries for 2018 for sector 1 with confidence interval 95%	89



Table 3.4.4.5-1: Results of LL algorithm for input data “Burglaries for sector 2” for various future timespan and holdback values	89
Table 3.4.4.5-2: Results of LLT algorithm for input data “Burglaries for sector 2” for various future timespan and holdback values	90
Table 3.4.4.5-3: Results of LLP algorithm for input data “Burglaries for sector 2” for various future timespan and holdback values	90
Table 3.4.4.5-4: Results of LLP5 algorithm for input data “Burglaries for sector 2” for various future timespan and holdback values	90
Table 3.4.4.5-5: Forecasted values regarding burglaries for 2018 for sector 2 with confidence interval 95%	91
Table 3.4.4.6-1: Results of LL algorithm for input data “Burglaries for sector 3” for various future timespan and holdback values	92
Table 3.4.4.6-2: Results of LLT algorithm for input data “Burglaries for sector 3” for various future timespan and holdback values	92
Table 3.4.4.6-3: Results of LLP algorithm for input data “Burglaries for sector 3” for various future timespan and holdback values	92
Table 3.4.4.6-4: Results of LLP5 algorithm for input data “Burglaries for sector 3” for various future timespan and holdback values	93
Table 3.4.4.6-5: Forecasted values regarding burglaries for 2018 for sector 3 with confidence interval 95%	94
Table 3.4.4.7-1: Results of LL algorithm for input data “Car robberies for sector 1” for various future timespan and holdback values	94
Table 3.4.4.7-2: Results of LLT algorithm for input data “Car robberies for sector 1” for various future timespan and holdback values	95
Table 3.4.4.7-3: Results of LLP algorithm for input data “Car robberies for sector 1” for various future timespan and holdback values	95
Table 3.4.4.7-4: Results of LLP5 algorithm for input data “Car robberies for sector 1” for various future timespan and holdback values	95
Table 3.4.4.7-5: Forecasted values regarding car robberies for 2018 for sector 1 with confidence interval 95%	96
Table 3.4.4.8-1: Results of LL algorithm for input data “Car robberies for sector 2” for various future timespan and holdback values	97
Table 3.4.4.8-2: Results of LLT algorithm for input data “Car robberies for sector 2” for various future timespan and holdback values	97
Table 3.4.4.8-3: Results of LLP algorithm for input data “Car robberies for sector 2” for various future timespan and holdback values	97
Table 3.4.4.8-4: Results of LLP5 algorithm for input data “Car robberies for sector 2” for various future timespan and holdback values	98
Table 3.4.4.8-5: Forecasted values regarding car robberies for 2018 for sector 2 with confidence interval 95%	99
Table 3.4.4.9-1: Results of LL algorithm for input data “Car robberies for sector 3” for various future timespan and holdback values	99
Table 3.4.4.9-2: Results of LLT algorithm for input data “Car robberies for sector 3” for various future timespan and holdback values	100
Table 3.4.4.9-3: Results of LLP algorithm for input data “Car robberies for sector 3” for various future timespan and holdback values	100
Table 3.4.4.9-4: Results of LLP5 algorithm for input data “Car robberies for sector 3” for various future timespan and holdback values	100

Table 3.4.4.9-5: Forecasted values regarding car robberies for 2018 for sector 3 with confidence interval 95	101
Table 3.4.4.10-1: Results of LL algorithm for input data “Motorcycle robberies for sector 1” for various future timespan and holdback values	102
Table 3.4.4.10-2: Results of LLT algorithm for input data “Motorcycle robberies for sector 1” for various future timespan and holdback values	102
Table 3.4.4.10-3: Results of LLP algorithm for input data “Motorcycle robberies for sector 1” for various future timespan and holdback values	102
Table 3.4.4.10-4: Results of LLP5 algorithm for input data “Motorcycle robberies for sector 1” for various future timespan and holdback values	103
Table 3.4.4.10-5: Forecasted values regarding motorcycle robberies for 2018 for sector 1 with confidence interval 95%	104
Table 3.4.4.11-1: Results of LL algorithm for input data “Motorcycle robberies for sector 2” for various future timespan and holdback values	104
Table 3.4.4.11-2: Results of LLT algorithm for input data “Motorcycle robberies for sector 2” for various future timespan and holdback values	105
Table 3.4.4.11-3: Results of LLP algorithm for input data “Motorcycle robberies for sector 2” for various future timespan and holdback values	105
Table 3.4.4.11-4: Results of LLP5 algorithm for input data “Motorcycle robberies for sector 2” for various future timespan and holdback values	105
Table 3.4.4.11-5: Forecasted values regarding motorcycle robberies for 2018 for sector 2 with confidence interval 95	106
Table 3.4.4.12-1: Results of LL algorithm for input data “Motorcycle robberies for sector 3” for various future timespan and holdback values	107
Table 3.4.4.12-2: Results of LLT algorithm for input data “Motorcycle robberies for sector 3” for various future timespan and holdback values	107
Table 3.4.4.12-3: Results of LLP algorithm for input data “Motorcycle robberies for sector 3” for various future timespan and holdback values	107
Table 3.4.4.12-4: Results of LLP5 algorithm for input data “Motorcycle robberies for sector 3” for various future timespan and holdback values	108
Table 3.4.4.12-5: Forecasted values regarding motorcycle robberies for 2018 for sector 3 with confidence interval 95%	109
Table 3.5.1-1: The 8×8 matrix containing the distances between all points of busiest day	111
Table 3.5.2-1: MST edges with distance cost and total distance cost of MST	113
Table 3.5.3.2-2: Coordinates of incident points and absolute center for busiest day	117
Table 3.6.1: Details of the busiest day	118
Table 3.6.2: Aggregate table of the DARP results	124
Table 3.7.1-1: Number of incidents occurred in city sectors p30 to p40	126
Table 3.7.3.1-1: Crime incidents fall within sector p25	130
Table 3.7.3.1-2: Sum of each row of minimum matrix after the multiplication of each row with weight $h_j$	131
Table 7.1.1: Aggregate table of offenses over the period 2010-2017	162
Table 7.4.1: Number of incidents in each city sector	171
Table 7.5.1: Minimum distance matrix of sector p25	172
Table 7.5.2: Minimum distance matrix of sector p38	173



## List of Figures

Figure 2.2.1: Predictive Analytics Process	26
Figure 2.2.2: Two-dimensional view of the result of clustering a set of input data into two clusters	31
Figure 2.4.1. Seven-Bridge problem at Königsberg	40
Figure 2.4.2. Example of Chinese Postman Problem graph	40
Figure 3.2.1-1: Format of data	55
Figure 3.2.3-1: Format of data after the addition of coordinates	58
Figure 3.2.3-2: Format of data after the addition of coordinates and extraction of time data values	59
Figure 3.2.4-1: Visualization of the individual incident points and the clusters for all years	61
Figure 3.3.1: Type of offense for all the years (2010-2017)	64
Figure 3.3.2: Type of offense for all the years, translated into percentages	64
Figure 3.3.3: Total of offenses for every year (2010-2017)	65
Figure 3.3.3: Total of offenses for every year (2010-2017)	65
Figure 3.3.4: Comparative chart for the months over all the years	64
Figure 3.3.5: Total of offenses per month over the period 2010-2017	64
Figure 3.3.6: Comparative chart for the weekdays over the period 2010-2017	67
Figure 3.3.7: Total of offenses for every weekday over the period 2010-2017	68
Figure 3.3.8: Total of offenses for every hour over the period 2010-2017	68
Figure 3.3.9: Total of offenses for every hour over the period 2010-2017, translated into percentage	69
Figure 3.3.10: Total of offenses for specific weeks of the year over the period 2010-2017	70
Figure 3.3.11: Total of offenses for specific weeks of the year over the period 2010-2017	70
Figure 3.3.12 (A,B,C): Visualization of the individual incident points and the clusters for 2011	72
Figure 3.4.1-1: The search head of Splunk Machine Learning Toolkit	74
Figure 3.4.2-1: The 3 sectors of the city of Volos	75
Figure 3.4.4.1-1: Forecast chart for all incidents data in sector 1 with confidence interval 95%	81
Figure 3.4.4.2-1: Forecast chart for all incidents data in sector 2 with confidence interval 95%	83
Figure 3.4.4.3-1: Forecast chart for all incidents data in sector 3 with confidence interval 95%	86
Figure 3.4.4.4-1: Forecast chart for burglaries in sector 1 with confidence interval 95%	93
Figure 3.4.4.5-1: Forecast chart for burglaries in sector 2 with confidence interval 95%	91
Figure 3.4.4.6-1: Forecast chart for burglaries in sector 3 with confidence interval 95%	92
Figure 3.4.4.7-1: Forecast chart car robberies in sector 1 with confidence interval 95%	96
Figure 3.4.4.8-1: Forecast chart car robberies in sector 2 with confidence interval 95%	98
Figure 3.4.4.9-1: Forecast chart car robberies in sector 3 with confidence interval 95%	101
Figure 3.4.4.10-1: Forecast chart motorcycle robberies in sector 1 with confidence interval 95%	103
Figure 3.4.4.11-1: Forecast chart motorcycle robberies in sector 2 with confidence interval 95%	106

Figure 3.4.4.12-1: Forecast chart motorcycle robberies in sector 3 with confidence interval 95%	108
Figure 3.5.1-1: List of the dates and the number of crime incidents for each date	111
Figure 3.5.1-2: Incident points of busiest day	111
Figure 3.5.2-1: The MST of busiest day graph network	113
Figure 3.5.3.2-1: Incidents and absolute center of the busiest day	117
Figure 3.6.1: Trips detail for every incident of the busiest day	118
Figure 3.6.2: Representation of the steps in Access	121
Figure 3.6 3: The apply of the alg100.txt	122
Figure 3.7.1-1: City sectors	126
Figure 3.7.3.1-1: Incidents points and median point of sector p25	132
Figure 3.7.3.2-1: Incidents points and median point of sector p38	133
Figure 4.1: The most common type of offense over the period 2010-2017	135
Figure 4.2: The month with most offenses over the period 2010-2017	136
Figure 4.3: The year with most offenses over the period 2010-2017	136
Figure 4.4: The weekday with most offenses over the period 2010-2017	137
Figure 4.5: The hour with most offenses over the period 2010-2017	137
Figure 4.6: The specific week of the year with most offenses over the period 2010-2017	138
Figure 4.7: The important day of the year with most offenses over the period 2010-2017	138
Figure 4.8: Prediction for 2018 with confidence interval 95% regarding all crime incidents	140
Figure 4.9: Crime prediction for 2018 with confidence interval 95%	140
Figure 4.10: Prediction for 2018 with confidence interval 95% regarding Burglaries	141
Figure 4.11: Burglary prediction for 2018 with confidence interval 95%	141
Figure 4.12: Prediction for 2018 with confidence interval 95% regarding Car Robberies	142
Figure 4.13: Car Robbery prediction for 2018 with confidence interval 95%	142
Figure 4.14: Prediction for 2018 with confidence interval 95% regarding Motorcycle Robberies	143
Figure 4.15: Motorcycle Robbery prediction for 2018 with confidence interval 95%	143
Figure 4.16: Prediction of type of offense for the year 2018	144
Figure 4.17: Crime Prediction for 2018 over the months	144
Figure 5.4-1: Connection between AI, ML and data science (Insert caption)	150
Figure 5.4-2: System principle (Insert caption)	152
Figure 7.2.1: Type of offense for the month January over the period 2010-2017	163
Figure 7.2.2: Type of offense for the month February over the period 2010-2017	163
Figure 7.2.3: Type of offense for the month March over the period 2010-2017	164
Figure 7.2.4: Type of offense for the month April over the period 2010-2017	164
Figure 7.2.5: Type of offense for the month May over the period 2010-2017	165
Figure 7.2.6: Type of offense for the month June over the period 2010-2017	165
Figure 7.2.7: Type of offense for the month July over the period 2010-2017	166
Figure 7.2.8: Type of offense for the month July over the period 2010-2017	166
Figure 7.2.9: Type of offense for the month September over the period 2010-2017	167
Figure 7.2.10: Type of offense for the month October over the period 2010-2017	167
Figure 7.2.11: Type of offense for the month November over the period 2010-2017	168
Figure 7.2.12: Type of offense for the month November over the period 2010-2017	168

Figure 7.2.13: Comparative chart for the type of offense for all the months over the period 2010-2017 169

## Acronyms

AI	Artificial Intelligence
BDA	Big Data Analytics
BI	Business Intelligence
CAD	Computer Aided Dispatch System
CPP	Chinese Postman Problem
CVRP	Capacitated Vehicle Routing Problem
DARP	Dial-a-Ride Problem
ESRI	Environmental Systems Research Institute
GIS	Geographic Information Systems
GPDP	General Pickup and Delivery Problem
IACA	International Association of Crime Analysis
ICAP	Integrated Criminal Apprehension Program
KNN	K-nearest neighbor
L.E.A.A.	Law Enforcement Assistance Administration
LAT	Latitude
LL	Local level
LLP	Seasonal local level
LLP5	Combination of LLT and LLP models
LLT	Local level trend
LON	Longitude
ML	Machine Learning
MO	Modus Operandi
MST	Minimum Spanning Tree
N.Y.P.D	New York Police Department
PDP	Pickup and Delivery Problem
PDV	Police Department of Volos
RMS	Record Management System
RMSE	Root Mean Square Error
RPP	Rural Postman Problem
SVMs	Support Vector Machines
TSP	Travelling Salesman Problem
U.S.	United States
VRP	Vehicle Routing Problem
VRPTW	Vehicle Routing Problem with Time Windows

## INTRODUCTION

Criminality as a major social problem accompanies humanity since its appearance on the planet. Despite economic and cultural progress, crime rates are constantly increasing with geometrical progress. This growth of global crime is a threat to the rule of law, without which there can be no sustainable word development. It is a fact that in our days crime is becoming more and more cruel, inhuman and unjustifiable. Due to continuing increase and violence of crime, its prevention and prediction is particularly important.

The overarching goal of this dissertation was to study data provided by the Police Department of Volos (PDV), which refer to thefts as criminal offences over the period 2010 - 2017, in the city of Volos. By studying and processing these data were produced statistical results related to the type of the offenses, the year, the month, the weekday, the hour and specific weeks and important days of the year. As prediction of crime incidents is essential in order to prevent and decrease future occurrences, we produced prediction results for 2018 by applying Splunk Machine Learning Toolkit and dividing the city of Volos into three big sectors indicating the west, the center and the east parts of the city. The prediction outcomes are related to which sector will be more in danger, who is the month that the most offenses will take place and what kind of offense will be the most common one

Subsequently was defined the Busiest day, the day with the largest number of crime incidents and was determined the absolute center, which indicates the point where the police car can be parked for stakeout, so as to minimize the maximum distance to or from the crime incidents. In order to study how the police cans serve all the incidents occurred in the busiest day was applied the DARP problem and was concluded that the police vehicle can reach all the crime incidents points on time. Furthermore the city of Volos was divided in 68 sectors



and the median points for sectors with high criminality were estimated by applying the Single Median Algorithm. The median points are considered valuable for police officers, as they indicate which minimize the average distance to or from the incidents occurred in the sectors. These points can be considered as hot spots across the city and police officers can do patrol around them.

## **1. CRIME ANALYSIS**

### **1.1 Introduction to Crime Analysis and History Review**

Crime analysis is a profession and process in which a set of quantitative and qualitative techniques are used to analyze data valuable to police agencies and their communities. It includes the analysis of crime and criminals, crime victims, disorder, quality of life issues, traffic issues, and internal police operations, and its results support criminal investigation and prosecution, patrol activities, crime prevention and reduction strategies, problem solving, and the evaluation of police efforts. Crime analysis can occur at various levels, including tactical, operational, and strategic (IACA, 2014).

Crime analysts study crime reports, arrests reports, and police calls for service to identify emerging patterns, series, and trends as quickly as possible. They analyze these phenomena for all relevant factors, sometimes predict or forecast future occurrences, and issue bulletins, reports, and alerts to their agencies. They then work with their police agencies to develop effective strategies and tactics to address crime and disorder. Other duties of crime analysts may include preparing statistics, data queries, or maps on demand; analyzing beat and shift configurations; preparing information for community or court presentations; answering questions from the public and the press; and providing data and information support for a police department's CompStat process. (short for compare statistics which was the computer file name of the original program) is a combination of management, philosophy, and organizational management tools for police departments.

To understand crime and law enforcement statistics, one must first understand the data on which the statistics are based. Knowledge of what the data do and do not describe is important in applying and interpreting statistics accurately and effectively. There are many considerations that must be taken into account with these data types as these issues can affect

the selection and interpretation of subsequent statistical analyses. Crime data is the primary type of data used in crime analysis. There are many issues about crime data that can lead to misinterpretation or misuse of statistics. Crime represented by police crime data does not represent all crime occurring in society. Thus, we typically see reports that name this type of data, “reported crime” or “crimes known to the police.” This is an important distinction for anyone interpreting or trying to understand crime problems based on police data: that is, that we may not be aware of the entire problem and the police data may portray a biased picture. For example, we know that a very low number of rapes are reported to the police. Thus, if we report rape statistics rising or falling we must be very cautious and mention the fact that even though the police are seeing an increase or decrease, the actual number may be changing in a different way since we do not know the actual number of rapes being committed. This is particularly relevant for certain types of crimes (domestic violence, drug crimes, white collar crimes) and not as much for others (motor vehicle theft, arson, murder).

Crime data captured by police agencies is dynamic, not static. In other words, information about crime incidents is constantly being updated. For example, a person may report a crime that occurred two days before to an officer who then takes a day or two to complete the report. The case is assigned to a detective who begins to investigate the crime by identifying suspects a few days, weeks, or years later. The victim may call the police department with new information about the crime or correct erroneous information from the original report. The officer may arrest a suspect and clear the case. As you can see, the data for this incident are constantly being changed and updated, and a report including this case could change based on when the crime analyst downloaded the data for statistics. That the data surrounding crime incidents are constantly changing creates an issue for analysis. One of the issues is “real-time” data. Many agencies emphasize performing statistics on the most recent data—as recent as an hour ago. Yet the likelihood of that data changing soon after the

initial report is recorded is fairly high because of additional information from investigation, identification of suspects, errors in the original report, and so on. Theoretically, however, after a certain amount of time, the likelihood of the data being changed substantially is significantly lower. Unfortunately, there is no research in this area and analysts must take an educated guess about when the likelihood for further changes to the data is low. In many cases, crime data will be downloaded around the fifteenth of the month following when it is reported and then updated in six month or one year intervals to capture arrests and clearances. In any case, the use of “real-time” data should be done with extreme caution.

All crime data recorded by the police have two sets of dates. The first is the date the report of the crime was written. (In some cases, the time the report was written is also recorded, but this information is not essential, except for certain operations studies.) The second is the date(s) and time(s) the crime actually occurred. This is included in crime data because the date a crime was reported is not always the same as when it occurred. In many cases, the exact date and time that the crime occurred are known (e.g., robbery, assault); however, in other types of crime, the exact date and time of the crime are not known because they were not witnessed by anyone (e.g., burglary, auto theft, larceny). In these crimes, victims report their best estimate of when the crime “could” have occurred. These are called “first possible” and “last possible” date and time or “from” and “to.” The difference between these two sets of dates (when the crime was reported vs. when the crime occurred) is very important as they serve different purposes depending on the nature of the analysis. All crime reported to the police is counted based on the date it is reported, because it would be impossible to count it by when it occurred, since this is not always known. As noted above, crime data is dynamic and crimes can be reported days, weeks, months, even years after they occur. Counts would need to be constantly updated based on date of occurrence and the ranges would cause further issues in counting. Date of report is constant, it does not change,

and thus counts of crime are based on this variable. However, it can affect the interpretation of crime statistics. For example, if a large number of burglaries occur over the Christmas/New Year's holiday, victims may not report them until after the new year when they return home. Thus, a large number of crimes would be counted in the following month and even the following year when they may have occurred the previous month and year. Although there is nothing an analyst or police can do to correct this problem, it should be considered during the analysis process. Finally, the date of occurrence, though not adequate for counting crime, is important for analysis and is used for identifying patterns and series when the crime occurred is much more important than when it was reported. The various ways in which crime is measured affect statistics created from them.

Police departments typically have two data systems in which crime, arrests, and calls for service, among other data, are housed. Even though these systems are slightly different from agency to agency, the basic purposes and functions are the same and thus these systems are discussed generally here. The two types of systems are referred to as a computer aided dispatch system (CAD) and a records management system (RMS). Although many small agencies may not have these systems and others may have additional specialized databases housed in specific units, the CAD and RMS are key systems for data in policing and crime analysis. A CAD or a computer-aided dispatch system is a highly specialized telecommunications and geographic display technology created to support police and public safety response operations. CADs are typically used for all emergency operations, which in addition to police include fire and ambulance services (Boba, 2013).

The introduction of the crime analysis discipline begins with what was been identified as the first modern police force. This police force was created in 1829 by a man who is commonly identified as the father of modern community Policing; Robert Peel. Robert Peel maintained that proper statistics were an important aspect of sound policing focused mostly

on basic crime statistics for the agency and attempts to keep sound records. August Vollmer, chief of the Berkeley, California Police Department in the early 1900s, introduced the English technique of systematic classification of known offender Modus Operandi (MO) to the United States. Vollmer developed the technique of examining recorded calls for service to perform beat analyses and was instrumental in promoting the use of "pin" or "spot" maps for visually identifying areas where crime and calls were concentrated. More important he used crime information to create patrol districts. Orlando Winfield (O.W) Wilson expanded on Vollmer's beat analysis system to make crime analysis more scientific and advanced in both its calculations and use. He included hazard formulas, or the assignment of weighting factors to various categories of crimes and services calls, in an effort to provide a systematic approach to the allocation of patrol resources. It was however not until the second edition of Wilson's Police Administration (1963) that first mention was made of the term "Crime analysis". (Boba,2013)

By the late 1960s, crime analysis units began to be established in the nation's larger police organizations. These units were primarily responsible for the detection of criminal modus operandi the discovery of crime patterns within geographical areas, and the determination of relationships between known offenders and crimes. The expansion of crime analysis was further influenced through development of the Integrated Criminal Apprehension Program (ICAP) by the Law Enforcement Assistance Administration (L.E.A.A.) in the 1970s. L.E.A.A. published a series of manuals on crime analysis between 1973 - 1977. When L.E.A.A. lost its funding in 1982, crime analysis entered a sort of dark years.

In the 1990s crime analysis took a new direction based on the book "Problem Oriented Policing" by University of Wisconsin professor Herman Goldstein. This book describes a set of procedures that seek to make police operations more effective by focusing on the crime

problem rather than the crime incident, and by funding ways to eliminate root causes before the problems themselves develop. Another important event that took place in the same decade was the establishment of the International Association of Crime Analysis (IACA), founded by a group of analysts seeking to share information and ideas, advocate for professional standards and provide educational opportunities. In 1992 started the certificate program in crime analysis by the California Department of Justice. Also in the 1990s arrived serious advances in the form of powerful, affordable technology, GIS software and relational database software. The institution of crime mapping, more specifically Geographic Information Systems (GIS) provided capabilities such as raster mapping; utilizing topographic survey software to map out changes in crime densities in a given area. Buffer analysis allowed analysts to generate a buffer around targeted areas and then identifying or select features based on whether they fall inside or outside the boundary of the buffer. (A History of Crime Analysis, 2011). During the first half of the decade, two companies - ESRI (Environmental Systems Research Institute) and Mapinfo introduced the first crime mapping applications accessible to the average of crime analysts. Many new advanced GIS features - from raster mapping to buffer analysis to three dimensional imaging - were suddenly available to analysts everywhere.

In 1994 the New York Police Department Incorporated a GIS system named the "compstat" or comparative statistics. This system allowed the N.Y.P.D to conduct extensive crime analysis as well as strategy development and management accountability utilizing crime mapping data (A History of Crime Analysis, 2011).

In 1997 the National Institute of Justice Crime Mapping Research Center started the promotion, research, evaluation, development and dissemination of GIS technology and spatial analysis of crime data to provide the U.S. Law enforcement community with tools necessary to reduce crime.

The decade of the 2000s has been very difficult for crime analysis. The September 11th led to a shift in funding and focus away from analysis and towards terrorism/homeland security. Recession has hit agencies very hard, reducing budgets farther and making training and advancement for existing analysts difficult.

However, it was during the period of 2005-2015 that the biggest leaps were made in the crime analysis profession. Increases in the number and type of databases available to crime analysts were substantial. Computerized analytical capabilities exploded. Training for analysts became available and was embraced in both the collegiate and professional training industries. Top decision-makers embraced the notion that traditional law enforcement had been replaced largely with pro-active policing models (much of which were leveraging successful business models). Today's analyst is college-educated (often at the graduate level), enters the job market with one or two law enforcement internships on their resume, has substantial experience with data sources, sets and analysis and understands the end game when it comes to providing meaningful input to key decision-makers based on their analytical discoveries. Analysts have become involved at every level in a policing organization, from presenting courses to new recruits to training officers in their annual training to community presentations on crime prevention techniques, crime analysts have embraced the role crime scientists – scientists who can provide much needed information and suggestions for police responses based on what is known to be impactful and effective. Analysts are now sitting on Executive Teams who impact the entire agency and hence, the community.

## **1.2 Types of Crime Analysis**

The types of crime analysis are organized around several factors, including the nature and the source of the data, the techniques applied, the results of the analysis, the regularity and frequency of the analysis and the indented audience and purpose. No typology will ever result



in a set of definitions that are completely exclusive or exhaustive; we must always be prepared for some overlap in definitions depending on the circumstances, as well as new ideas and techniques to emerge. However, according to IACA (International Association of crime Analysis) there are four major categories of crime analysis, ordered from specific to general:

- ***Intelligence crime analysis:*** The International Association of crime analysis defines intelligence analysis as the collection and dissemination of information about criminals particularly organizations and conspiracies.
- ***Tactical crime analysis:*** Tactical crime analysis is the study of recent criminal incidents and potential criminal activity through the examination of characteristics such as how, when and where the activity has occurred to assist in pattern development, investigative lead and suspect identification, and case clearance. Its subjects areas include the analysis of space, time, offender, victim, and modus operandi for individual high-profile crimes, repeat incidents, and crime patterns with a specific focus on crime series. (Boba,2013)
- ***Strategic crime analysis:*** Is the analysis of data directed towards development and evaluation of long-term strategies, policies, and prevention techniques. Its subjects include long-term statistical trends, hot spots, and problems. Although it often starts with data from police records systems, strategic analysis usually includes the collection of primary data from a variety of other sources through both quantitative and qualitative methods.
- ***Administrative crime analysis:*** Is the analysis directed towards the administrative needs of the police agency, its government and its community. As a broad category, it includes a variety of techniques and products, performed both

regularly and on request, including statistics, data printouts, maps and charts.  
(IACA,2014)

### **1.3 Technology of Crime Analysis**

Historically, data sets were stored locally on individual computers or network drives. However, the expansion of computer technology has changed the way data is stored. More recently data sets are being stored on internet accessible files. These files are served within organizations, across organizations and even to the general public.

Analyzing crimes and understanding where crime is likely to happen in the future requires a variety of computerized tools and programs to determine likely repeat suspects and other crucial details. Data from computer aided dispatch (CAD) and records management systems (RMS) is compiled using crime analysis software which can make it easy to analyze trends, generate graphs and heat maps. Analyzing geographic information is essential for analyzing and predicting crime. This data includes maps of the locations of crimes. Many agencies handle this through a program called Geographic Information Systems, or GIS.

GIS (Geographical Information Systems) which is one of the very effective quantitative data analysis methods which have become one of the best technologies used nowadays by these security institutions to improve the crime investigation quality, because maps have the power of offering crime analysts the crime related issues in a notion of graphical mechanism. GIS facilitates the modeling of the workflow of a crime and captures its best practices. The location where crimes or activities occur and the relationship of those places to one another and to other information is an important factor in crime analysis (Chamikara et al.,2012). It is not only important where a crime takes place but also the characteristics of those places and the environment in which the crime occurs. Thus, examination of spatial data such as streets networks, parcel information, orthophotographs, school locations, business and residential zoning, among others, is imperative for effective crime analysis. Simple maps that display the locations where crimes or concentrations of crimes have occurred can be used to help direct patrols to place where they are most needed. Policy makers in police departments might use

more complex maps to observe trends in criminal activities, and maps may prove invaluable in solving criminal cases. Digital maps are the quickest means of visualizing the entire crime scenario. The locations of crime events, arrests, etc. can be routinely displayed on maps. This provides an easy method of viewing activities in an area rather than searching through a list of events. Maps can also be used to convey more than one type of information at a time. Crime locations can be symbolized according to the day of week, type of crime, modus operandi (a particular suspect's method of operation when committing a crime) or frequency (Agrawal et al., 1993)

In addition, mapping and GIS can support community and problem oriented policing. Mapping and GIS can show detailed relationships between the crime, victim, and the offender. Some other very important aspects of mapping and GIS are, showing of demographic and population changes, assisting in resource allocation, integrating data from community and government sources, providing effective communication tools. GIS aids crime analysis also by identifying and highlighting suspicious incidents and events that may require further investigation. Supporting patterns and trend analysis across multiple jurisdictions, enhancing the implementation of various policing methodologies to reduce overall crime and disorder, and integrating traditional and non-traditional law enforcement data to improve overall analysis are few other very important applications of GIS which aids the crime analysis process in a drastic manner. One other very important facet of GIS is educating the public with visual information to clarify crime concerns and enlist community action, providing tools and techniques to capture crime series and forecast future crime occurrences (Koperski and Han, 1995). Primary goal of law enforcement is anyhow to prevent crimes through the methods other than apprehension. Therefore, GIS lends itself particularly well to assist for crime analysts towards the need of crime prevention of many security institutions.

## 2. PROPOSED METHODOLOGY

### 2.1 Data and Statistical Analysis

#### 2.1.1 Definition of Data and Data Processing Cycle

According to Merriam Webster Online Dictionary data is defined as the following

- factual information (as measurements or statistics) used as a basis for reasoning, discussion or calculation
- information output by a sensing device or organ that includes both useful and irrelevant or redundant information and must be processed to be meaningful
- information in numerical form that can be digitally transmitted or processed

Taking from the above definitions, a practical approach to define data is that data is numbers, characters, images, or other method of recording in a form which can be assessed to make a determination or decision about a specific action.

There are two types of data:

- ✓ **Qualitative data:** data that is represented either in a verbal or narrative format. These types of data are collected through focus groups, interviews opened ended questionnaire items, and other less structured situations.
- ✓ **Quantitative data:** is data that is expressed in numerical terms, in which the numeric values could be large or small. Numerical values may correspond to a specific category or label. (Introduction to data analysis Handbook,2006)

Data on its own has no meaning, only when interpreted by some kind of data processing system does it take on meaning and become information. By closing examining data we can find patterns to perceive information and then information can be used to enhance knowledge.

In any research, the step of analysis of the data is one of the most crucial tasks requiring proficient knowledge to handle the data collected as per the predecided research design of the

project. Data analysis is the process of evaluating data using analytical and logical reasoning to examine each component of the data provided. This form of analysis is just one of many steps that must be completed when conducting a research experiment. Data from various sources is gathered, reviewed, and then analyzed to form some sort of finding or conclusion. There are a variety of specific data analysis methods, some of which include data mining, text analytics, business intelligence, and data visualizations.

Once data is collected, following steps are taken to process the data into more measurable and concise manner. Data processing is simply the conversion of raw data to meaningful information through a process. Data is manipulated to produce results that lead to a resolution of a problem or improvement of an existing situation. Similar to a production process, it follows a cycle where inputs (raw data) are fed to a process (computer systems, software, etc.) to produce output (information and insights). The general understanding is that data analysis and processing are one and the same. However a number of researchers and authors are of the opinion that both of them are two very distinct steps in the research process where data processing leads to data analysis.

#### **Stages of the Data Processing Cycle:**

- ✓ **Collection:** is the first stage of the cycle, and is very crucial, since the quality of data collected will impact heavily on the output. The collection process needs to ensure that the data gathered are both defined and accurate, so that subsequent decisions based on the findings are valid. This stage provides both the baseline from which to measure, and a target on what to improve.
- ✓ **Preparation:** is the manipulation of data into a form suitable for further analysis and processing. Raw data cannot be processed and must be checked for accuracy. Preparation is about constructing a data set from one or more data sources to be used for further exploration and processing. Analyzing data that has not been carefully

screened for problems can produce highly misleading results that are heavily dependent on the quality of data prepared.

- ✓ **Input:** is the task where verified data is coded or converted into machine readable form so that it can be processed through an application. Data entry is done through the use of a keyboard, scanner, or data entry from an existing source. This time-consuming process requires speed and accuracy. Most data need to follow a formal and strict syntax since a great deal of processing power is required to breakdown the complex data at this stage. Due to the costs, many businesses are resorting to outsource this stage.
- ✓ **Processing:** is when the data is subjected to various means and methods of powerful technical manipulations using *Machine Learning* and *Artificial Intelligence algorithms* to generate an output or interpretation about the data. The process may be made up of multiple threads of execution that simultaneously execute instructions, depending on the type of data. There are applications available for processing large volumes of heterogeneous data within very short periods.
- ✓ **Output and interpretation:** is the stage where processed information is now transmitted and displayed to the user. Output is presented to users in various report formats like graphical reports, audio, video, or document viewers. Output need to be interpreted so that it can provide meaningful information that will guide future decisions.
- ✓ **Storage:** is the last stage in the data processing cycle, where data, and metadata (information about data) are held for future use. The importance of this cycle is that it allows quick access and retrieval of the processed information, allowing it to be passed on to the next stage directly, when needed (Hassania & Gahnouchia, 2015)

The **Data Processing Cycle** is a series of steps carried out to extract useful information from raw data. Although each step must be taken in order, the order is cyclic. The output and storage stage can lead to the repeat of the data collection stage, resulting in another cycle of data processing. The cycle provides a view on how the data travels and transforms from collection to interpretation, and ultimately, used in effective business decisions.

There are number of methods and techniques which can be adopted for processing of data depending upon the requirements, time availability, software and hardware capability of the technology being used for data processing. There are number of types of data processing methods.

**Types of data processing:**

- ✓ **Manual data processing:** In this method data is processed manually without use of machine or electronic device. This methods might be accompanied with automatic method for completion of the data processing.
- ✓ **Mechanical data processing:** Data processing is done by use of mechanical device or very simple electronic devices like calculator and type writers. When the need for processing is simple this method can be adopted.
- ✓ **Electronic data processing:** This is the fastest and best available method with highest reliability and accuracy. Technology used is latest as this method uses computers and employed in most of the agencies. The use of softwares forms the part of this type of data processing (Mishra & Gupta, 2017)

**2.1.2. Data Interpretation**

Once the data has been processed and analyzed, the final step required in the research process is the interpretation of data. The line between analysis and interpretation is very thin. Through interpretation one understands what the given research findings really mean and

what is the underline generalization which is manifested through the data collected. This can be descriptive, or analytical, or theoretical. The data is interpreted from the point of the research questions and hypothesis is tested. While interpretation is done generalizations are drawn. Thus, interpretation consists of conclusion that the researcher has reached after the data has been processed and analyzed.

Data presentation follows a structure like:

- Describe: Pen down the "facts" observed/ heard after filtering the non relevant data
- Classify: Group the material based similarities, categorize, and make headings.
- Interpret: Identify important features and patterns in the light of the research questions or hypothesis and then represent them.

Various forms of diagrams are used so as data to be represented. They in a very meaningful way highlight the salient features of the data which makes them easy to understand (Vanlalhiriati & Singh, 2015).

### **2.1.3 Statistical Analysis of Data**

Statistics is concerned with the scientific method by which information is collected, organized, analyzed and interpreted for the purpose of description and decision making. Croxton and Cowden, two well known statisticians have introduced a simple definition of statistics. In their words, « Statistics may be defined as the science of collection, presenting and analysis and interpretation of numerical data». It is a sound techniques or method for handling the collected data, analyzing the data and used for drawing valid inferences from them. There are different statistical approaches available to a researcher. Choices of appropriate statistical techniques are determined to a great extent by the research design, hypothesis and the kind of data that will be collected. When the data are collected, edited, classified and tabulated, they are analyzed and interpreted with the help of various statistical



tools based on the nature of investigation. Thus, the researcher is expected to have basic knowledge of statistics for carrying out the systematic analysis as well to provide accurate and precise interpretation of data (Vanlalhiriati & Singh, 2015). One of the greatest advantages of the use of statistics is that in a research with large data, it helps in reducing such data into a more manageable size for the purpose of analysis and interpretation. It also helps in comparing two or more series as well as draw inferences and conclusion of the research.

Though statistical methods are of great value to a researcher, they carry with themselves certain limitations which must be kept in mind while deciding a tool of data analysis. These limitations are:

- ✓ Qualitative values like subjective perceptions qualities and attributes are not considered under statistics. It only considers quantities. This by far is the greatest limitation of statistics.
- ✓ Statistics studies and analysis group attributes rather than individual characteristics and values.
- ✓ Statistical analysis is mostly based on average; hence the inferences drawn through them are only approximate and not exact like that of mathematics.
- ✓ Statistics only help discover, analyze certain characteristics, it only forms a part of the inference and interpretation.

There are various statistical tools which are available for the researcher's assistance.

Data analysis tools:

- ✓ *Measure central tendency:* is a single value that describes the way in which a group of data cluster around a central value. There are three measures of central tendency: the mean, the median, and the mode

- ✓ *Measure of dispersion:* is the extent to which a distribution is stretched or squeezed. Common examples of measures of statistical dispersion are the variance, standard deviation, and interquartile range.
- ✓ *Measure of relationship:* are statistical measures which show a relationship between two or more variables or two or more sets of data
- ✓ *Measure of asymmetry*
- ✓ *Other measures*

There are various statistic software available for computerized statistical data analysis, which are of great help when analyzing large quantities of data. The most commonly used are SAS (Statistical Analysis System) and SPSS (Statistical Package for Social Sciences).

#### **2.1.4 Spatial Data Analysis**

In this thesis we deal with Crime Data that we study as Spatial Data. We refer to thefts as criminal offences which are recorded according to the geographic point, the day and the time they were committed.

By Spatial Data we mean data that contain locational as well as attribute information. In fact Spatial Data refers to all types of data objects or elements that are present in a geographical space or horizon. It enables the global finding and locating of individuals or devices anywhere in the world. Spatial Data is also known as geospatial data, spatial information or geographic information. This type of data is used in geographical information systems (GIS) and other geolocation or positioning services. It consists of points, lines, polygons and other geographic and geometric data primitives which can be mapped by location, stored with an object as metadata or used by a communication system to locate end user devices (Reid, 2017).

In describing the nature of spatial data it is important to distinguish between the discreteness or continuity of the space on which the variables are measured, and the discreteness and the continuity of the variable values (measurements) themselves. If the space is continuous, variable values must be continuous valued since continuity of the field could not be preserved under discrete valued variables. If the space is discrete, or if a continuous space has been made discrete, variable values may be continuously valued or discrete valued (nominal or ordinal valued).

### **Types of spatial data:**

- ✓ ***Point pattern data:*** Data set consisting of a series of point location in some study region at which events of interest have occurred, such as cases of a disease or incidence of a type of crime.
- ✓ ***Field data (geostatistical data):*** Data that relate to variables which are conceptually continuous and whose observations have been saved at a predefined and fixed set of point locations.
- ✓ ***Area data:*** Where area values are observations associated with a fixed number of areal units that may form a regular lattice, as with remotely sensed images, or be a set of irregular areas or zones, such as counties, districts, census zones, and even countries.
- ✓ ***Spatial interaction data:*** Consisting of measurements each of which is associated with a pair of point locations, or pair of areas (Fisher, 2011).

Spatial data analysis focuses on detecting patterns and exploring and modeling relationships between such patterns in order to understand processes responsible for observed patterns. In this way, spatial data analysis emphasizes the role of space as a potentially

important explicator of socioeconomic systems, and attempts to enhance understanding of the working and representation of space, spatial patterns, and processes.

As mentioned before in this thesis we study *Crime Data*, so it's important to make a short reference to the mapping and spatial analysis of crime. In its most basic form, crime mapping is the use of Geographic Information System (**GIS**) to visualize and organize spatial data for more formal statistical analysis. Spatial analysis can be employed in both an exploratory and well as a more confirmatory manner with the primary purpose of identifying how certain community or ecological factors (such as population characteristics or the built environment) influence the spatial patterns of crime. Two topics of particular interest include examining for evidence of the diffusion of crime and in evaluating the effectiveness of geographically targeted crime reduction strategies. Crime mapping can also be used to visualize and analyze the movement or target selection patterns of criminals. Mapping software allows for the creation of electronic pin-maps and by spatially organizing the data, GIS increases the analytical value of these maps. Crime mapping allows researchers and practitioners to explore crime patterns, offender mobility, and serial offenses over time and space. Within the context of local policing, crime mapping provides the visualization of crime clusters by types of crimes, thereby validating the street knowledge of patrol officers. Crime mapping can be used for allocating resources (patrol, specialized enforcement) and also to inform how the concerns of local citizens are being addressed (Chainey et al, 2005).

### **2.1.5 Big Data Analysis**

Big Data is a term that describes voluminous amount of data that is structural, semi-structural and sub-structural data that has potential to be mined for information. In fact Big Data is data that are so voluminous and complex that traditional data-processing application software are inadequate to deal with them. Big data is defined through the 3Vs which are

presented by Laney in 2001 as “high-volume, high velocity and high-variety information assets that demand cost-effective, innovative forms of information handling for improved insight and decision making” (Laney, 2001). In 2012, Gartner updated the definition as follows: “Big Data is high volume, high velocity, and/or high variety of information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization”. Other authors, researchers and engineers have extended the 3Vs to 4Vs and 5Vs by adding Value and Veracity to the definition of Big Data.

In the concept of "Big Data" it is not really the amount of data that makes the novelty, but rather the combination of 3 Vs. These fundamental characteristics are described as follows:

- ✓ **Volume:** refers to large amounts of any kind of data from any different sources,
- ✓ **Variety:** refers to different types of data collected via sensors, smart phones or social networks, such as videos, images, text, audio, and so on. Moreover, these data can be in structured or unstructured formats.
- ✓ **Velocity:** refers to the speed of data transfers. The data content is constantly changing (Hassania & Gahnouchi, 2017)

Big Data challenges include capturing data, data storage, data analysis, search, sharing, transfer, visualization, querying, updating, information privacy and data source. Lately, the term "Big Data" tends to refer to the use of predictive analytics, user behavior analytics, or certain other advanced data analytics methods that extract value from data, and seldom to a particular size of data set. Analysis of data sets can find new correlations to spot business trends, prevent diseases, combat crime and so on. Big Data uses inductive statistics and concepts from nonlinear system identification to infer laws (regressions, nonlinear relationships, and causal effects) from large sets of data with low information density to reveal relationships and dependencies, or to perform predictions of outcomes and behaviors.

Big Data Analytics (BDA), is the process of examining large data sets that containing a variety of data types like big data to uncover all hidden patterns, unknown correlations, market trends, customer preferences and other useful business information. Then analytical findings can lead to more effective marketing, new revenue opportunities, better customer service, improved operational efficiency, competitive advantages over rival organizations and other business benefits. The primary goal of Big Data Analytics is to help companies make more informative business decisions by enabling data scientists, predictive modelers and other analytics professionals to analyze large volumes of transactional data, as well as other forms of data that may be untapped by more conventional Business Intelligence (BI) programs. Big data burst upon the scene in the first decade of the 21st century, and the first organization to embrace it were online and start-up firms. Arguably, firms like Google, LinkedIn, and eBay and Face book were built around big data from the beginning (Kuchipudi Sravanthi et al, 2015).

The required type of analysis has to deal with data which are not only big in term of quantity, but are also generated with various formats and in high speed. So, big data analytics is when advanced analytic techniques operate on big data. Big data analytics research can be classified into five areas: text analytics, multimedia analytics, web analytics, network analytics, and mobile analytics (Vashisht & Gupta, 2015). We present bellow a brief review of big data analytics techniques:

- ✓ *Text analytics or text mining:* refers to techniques that extract information from textual data. It involves statistical analysis, computational linguistics, and machine learning. This kind of analytics allows organizations to convert large amount of human-generated texts into meaningful summaries that support decision making. Text analytics systems are based on text representation and natural language processing (NLP).

- ✓ *Audio analytics or speech analysis:* used to analyze and extract information from unstructured audio data.
- ✓ *Video analytics:* Involves several techniques to monitor, analyze and extract meaningful information from video streams.
- ✓ *Social media analytics:* is the analysis of structured and unstructured data collected from various social media channels.
- ✓ *Predictive analytics:* comprises several techniques for predicting future outcomes based on the past or historical and current data (Gandomi et al,2015).

Big Data Analytics Applications (BDA Apps) are a new category of software applications that leverage largescale data, which is typically too large to fit in memory or even on one hard drive, to uncover actionable knowledge using large scale parallel-processing infrastructures (D. Fisher, 2012). The big data can come from sources such as runtime information about traffic, stock market updates, usage information of an online game, or the data from any other rapidly growing data-intensive software system.

## **2.2 Prediction Techniques**

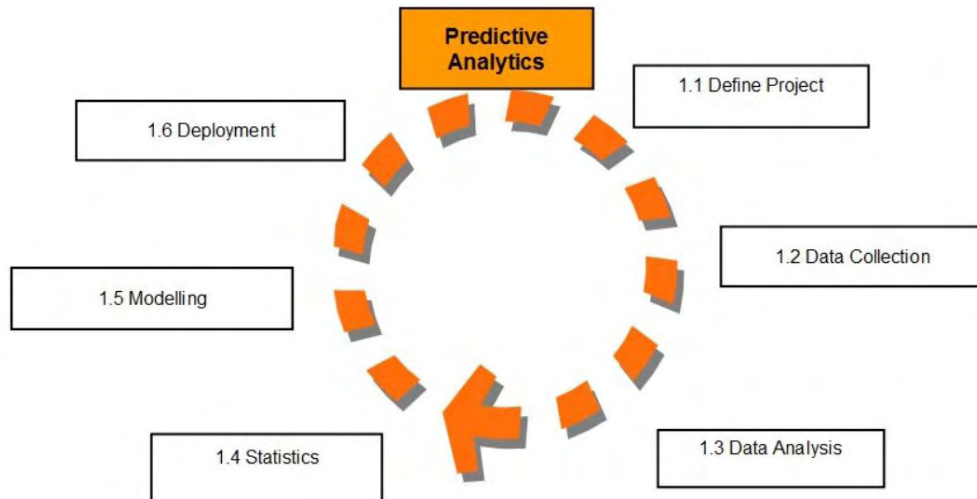
### **2.2.1 Predictive Analytics and its Process**

Predictive analytics is a form of advanced analytics that uses both new and historical data to forecast activity, behavior and trends. It involves applying statistical analysis techniques, analytical queries and automated machine learning algorithms to data sets to create predictive models that place a numerical value on the likelihood of a particular event happening (Techtarget, 2018).

The process involved in predictive analytics follows 7 steps:

- 1. Define Project:** Define what shall be the outcome of the project, the deliverables, business objectives, and based on that gather those data sets that are going to be used.
- 2. Collect Data:** This is more of the big basket where all data from various sources are binned for usage. This gives a complete view about the customer interactions.
- 3. Analysis of Data:** Here the data are inspected, cleansed, transformed and modelled in order to discover useful information and arrive at conclusion.
- 4. Statistics:** This enables to validate the findings, assumptions and hypothesis and test them using statistical models.
- 5. Modelling:** Provide accurate predictive models for the future. Through multi model evaluation the best option could be chosen as the required solution from the options available.
- 6. Deployment:** Through the predicative model deployment an option is created to deploy the analytics results into everyday effective decision. This way the results, reports and other outputs can be taken based on modelling.
- 7. Monitoring:** Models are monitored to control and check the performance to ensure that the desired results are obtained as expected (Fhyzics Buisness Consultants).





**Figure 2.2.1:** *Predictive Analytics Process*

### 2.2.2 Predictive Analytic Models

Predictive analytics is often used to mean predictive models. However, we are increasingly using the term to describe related analytic disciplines used to improve customer decisions. Since different forms of predictive analytics tackle slightly different customer decisions, they are commonly used together. So we have 3 main types in the field of predictive analytics, which are discussed below.

#### ***Predictive Models:***

Predictive models analyze past performance to predict the possibility that a special behavior is exhibited by a customer in the future. This category also encompasses models that detect subtle data patterns to answer questions about customer behavior, such as fraud detection models. Operational processes often include predictive models which are activated during live transactions. The models analyze historical and transactional data to isolate patterns such as fraudulent transaction, a risky customer or a customer likely to switch

providers. These analyses weigh the relationship between hundreds of data elements to isolate each customer's risk or potential, which guides the action on that customer.

### ***Descriptive Models:***

Unlike predictive models that predict a single customer behavior, descriptive models identify many different connections between customers and products. Descriptive models encode relationships into data in a way that is often used to classify customers or prospects into groups. For example, a descriptive model may categorize customers into various groups with different buying patterns. Descriptive modeling tools can be used for the development of further models that can simulate huge numbers of individualized agents and make predictions

### ***Decision Models:***

Decision models predict the outcomes of complex decisions in much the same way predictive models predict customer behavior. Decision models make predictions of what is going to happen in case a given action is taken, through the mapping of the relationships between all the elements of a decision. These models can be utilized in optimization, maximizing certain results while minimizing others. Decision models are generally used to develop decision logic or a group of business regulations that will produce the desired action for every customer or condition (FICO).

## **2.2.3 Predictive Analytic Models Applications and Techniques**

Predictive analytics can be of use in many applications and had a great impact in some of them in recent years. Some of these applications are CRM, Health care, Collection Analytics, Cross Sell, Fraud Detection, Risk Management, Direct Marketing and Underwriting. As a result, predictive analytics find a wide range of usage in telecom, insurance, banking,

marketing, financial services, retail, travel, health care, pharmaceuticals, oil and gas and a host of other industries where organizations are getting to take decisions based on data.

The idea behind any predictive model is to create a mapping function between a set of input data fields and a target variable. How you can make this feasible depends on the size and complexity of your data set, but there is a number of tested predictive modelling techniques you can use across a variety of applications.

### ***Support Vector Machines (SVMs):***

SVMs are linear models with the difference that they have different margin-based loss function. It is a supervised machine learning technique that analyzes smaller datasets and recognizes patterns which can be used for classification and regression analysis. This technique uses a hyperplane to divide datasets into two distinct classes. Often, SVMs require three-dimensional mapping to ensure the widest possible margin between the hyperplane and the two data classes. Although results in small datasets are accurate, training time for larger datasets is usually a limiting factor. SVMs are great for specific classification tasks such as facial and image recognition.

### ***Decision Trees:***

Decision trees work by recursively partition data into smaller subsets. At each new branch of the decision tree, data are further split until a classification or decision is made. Decision trees are computationally cheap and easy to understand. However, they are prone to **overfitting** and need to be pruned regularly. While a model may fit training data well, it can often do a poor job of classification in the real-world. Boosted decision trees use additive boosting techniques to combine the outcomes of weaker decision trees and provide a more weighted measure of accuracy. Decision trees are often used because they are easy to

understand and interpret. The most common decision trees algorithms used these days are Random Forests and Boosting Trees.

### ***Naïve Bayes:***

Naïve Bayes is a group of algorithms that use Bayes Theorem for calculating probability. This predictive modelling technique classifies items assuming that each variable is independent to each other. A **Bayesian recommendations system** creates a probabilistic model of personalized item recommendations, drawing on previous user behavior. Naïve Bayes models are originally easy to build and interpret and produce very fast results. However, its naivety comes from the assumption that every variable is independent, which is often not the case. Simple text classification tasks such as spam detection and sentiment analysis use Naïve Bayes to predict probabilities.

### ***Neural Networks:***

Neural networks mimic the human brain by imitating the way that neurons relay information. They have the ability to model extremely complex relationships and they are very powerful and flexible. In predictive modelling, a neural network uses three layers to detect patterns in data: an input layer, a hidden layer and an output layer. They aim to use the mathematical functions in the hidden layer to produce an output free of noise. Neural networks can be used in both classification and regression and are great at handling a huge amount of non-linear data. They are mainly used for predictions about time series data such as weather data or economic trends (Redpiexe).

### ***Linear regression:***

Linear regression is one of the most famous modeling techniques. This technique is among the first few topics which people pick while learning predictive modeling. It is suitable for numeric prediction that is often used in statistical applications. In this

technique, the dependent variable is continuous, independent variables can be continuous or discrete, and the regression line is linear. In linear regression, we predict value of one variable from the value of another variable. The variable we are predicting is called the target variable and the variable we are basing our predictions on, is called the predictor variable. The purpose of this technique is to find optimal weights for the training instances by minimizing the error between the real and the predicted values. As long as the data set contains more instances than attributes this is easily done using the least square method. Linear regression is quite intuitive and easy to understand but the negative is that it can't handle non-numerical attributes well enough and that it can't handle more complex nonlinear problems

### ***Principal Component Analysis:***

The purpose of principal component analysis is to derive a small number of independent linear combinations of a group of variables that retain as much of the information in the original variables as possible.

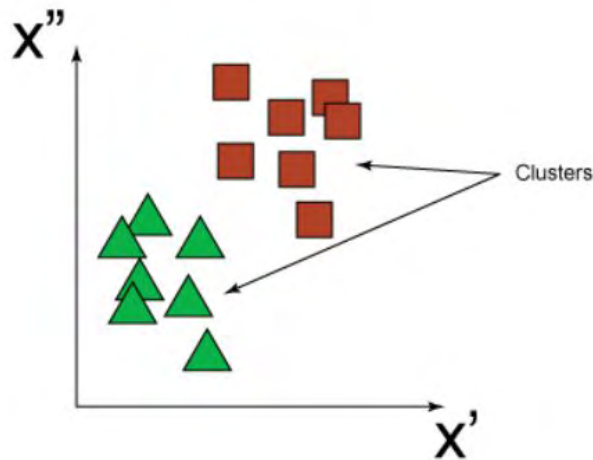
### ***K-nearest neighbor (KNN):***

The KNN algorithm belongs to the pattern recognition and is a nonparametric method for classification and regression. It finds a group of  $k$  objects in the training set that are closest to the test object, and bases the assignment of a label on the predominance of a particular class in this neighborhood

### ***K-means clustering:***

K-means clustering is a type of unsupervised learning, which is utilized in case you have unlabeled data. K-means algorithm separates a given dataset into  $k$  clusters. Given a set of data points in form of vectors, we can make clusters of points based on distances between them. The input the algorithm has taken is the number of clusters which are to be generated

and the number of iterations in which it will try to converge clusters. The representative of a cluster is the mean of all items that belong to the same cluster which is also called a cluster centroid.



**Figure 2.2.2:** *Two-dimensional view of the result of clustering a set of input data into two clusters*

### ***Time Series Data Mining:***

Time series data represents a collection of values obtained from sequential measurements over time at a particular interval. Time series data mining stems from the desire to reify our natural ability to visualize the shape of data. Humans rely on complex schemes so they are able to perform such tasks. Time series data mining combines traditional data mining and forecasting techniques. Data mining techniques such as sampling, clustering and decision trees are applied to data collected over time with the goal of improving predictions.

## 2.3 Cluster Analysis

Cluster analysis has been widely used in many sciences such as biology, software engineering, statistics, psychology and other social sciences, in order to find groups in large amounts of data. The dimensions of these datasets make analysis and validation of the results while they constantly become larger. The increase in the number of publications involving this subject has made utilization of clustering wide within computer science and especially the data base community. Cluster analysis divides data into groups those that are meaningful, useful, or both. If meaningful groups are the target, then the clusters should capture the natural structure of the data. In some cases, however, cluster analysis is only a useful starting point for other purposes, such as data summarization.

### 2.3.1 Clustering Definition and Stages

Clustering is described as the separation of data into groups of similar objects. Cluster analysis groups data objects into clusters such that objects which belong to the same cluster are similar, while those which belong to different ones are dissimilar. From the machine learning perspective, clustering can be viewed as unsupervised learning concepts. From the definition above we can understand that clustering cannot be a one-step process. So the clustering process follows the stages below:

Data Collection: In this first stage the relevant data objects from the underlying data sources are carefully extracted. In our context, data objects are distinguished by their individual values for a set of attributes.

Initial Screening: Refers to the massaging of data after its extraction from the source, or sources. This stage is closely connected to Data Cleaning, a process widely utilized in Data Warehousing.

Representation: The data are prepared in order to become suitable for the clustering algorithm. At this stage we have the examination of the characteristics and dimensionality of data and the similarity measure is chosen.

Clustering Tendency: Checks whether the data to be used have a natural tendency to cluster or not. This stage can be often ignored, especially in case of large datasets.

Clustering Strategy: Careful choice of clustering algorithm and initial parameters.

Validation: This is one of the last and, in our opinion, most under-studied stages. Validation is often based on manual examination and visual techniques. However, as the number of data and their dimension get larger, we have no means to compare the results with preconceived ideas or other clusterings.

Interpretation: At this final stage clustering results are combined with those of other studies.

### 2.3.2 Clustering Techniques

#### Basic

Partitional and Hierarchical clustering techniques are the two types of basic techniques. Their definitions are as follows:

**Partitional**: Given a database of objects, a partitional clustering algorithm constructs partitions of the data, where each cluster optimizes a clustering criterion, such as the minimization of the sum of squared distance from the mean within each cluster. One of the issues with such algorithms is their high complexity, as some of them exhaustively enumerate all possible groupings and try to find the global optimum. Even for a small amount of objects, the number of partitions is large. That's why, common solutions start with an initial, usually random, partition and proceed with its refinement. A better practice would be to run the partitional algorithm for different sets of initial points and investigate whether all solutions



lead to the same final partition. Partitional Clustering algorithms try to locally improve a certain criterion. First, they compute the values of the similarity or distance, they order the results, and pick the one that optimizes the criterion.

**Hierarchical:** Hierarchical algorithms create a hierarchical decomposition of the objects. They are either agglomerative (bottom-up) or divisive (top-down):

(a) Agglomerative algorithms begin with each object being a separate cluster itself, and successively merge groups according to a distance measure. The clustering may stop if all objects are in one group or at any other point the user wills. These methods generally follow a greedy-like bottom-up merging.

(b) Divisive algorithms follow the opposite strategy. They begin with a single group of all objects and successively divide groups into smaller ones, until each object falls in one cluster, or as desired. Divisive approaches split the data objects in disjoint groups at every step, and follow the same pattern until all objects fall into a separate cluster. This is similar to the approach followed by divide-and-conquer algorithms. Most of the times, both approaches suffer from the fact that once a merge or a split is committed, it cannot be undone or refined.

### Data Mining

Apart from the two basic categories, many other methods have emerged in cluster analysis, which are mainly focused on specific problems or specific data groups available. These methods include:

*Density-Based Clustering:* These algorithms group objects according to specific density objective functions. Density is usually defined as the number of objects in a particular neighborhood of a data objects. In these approaches a given cluster continues growing as long as the number of objects in the neighborhood exceeds some parameter. This is considered to

be different from the idea in partitional algorithms that use iterative relocation of points given a certain number of clusters.

*Grid-Based Clustering:* The main focus of these algorithms is spatial data, i.e., data that model the geometric structure of objects in space, their relationships, properties and operations. The objective of these algorithms is to quantize the dataset into a number of cells and then work with objects belonging to these cells. They do not relocate points but rather build several hierarchical levels of groups of objects. In this sense, they are closer to hierarchical algorithms but the merging of grids, and consequently clusters, does not depend on a distance measure but it is decided by a predefined parameter.

*Model-Based Clustering:* These algorithms find proper approximations of model parameters that best fit the data. They can be either partitional or hierarchical, depending on the structure or model they hypothesize about the data set and the way they refine this model to identify partitionings. They are closer to density-based algorithms, in that they grow particular clusters so that the preconceived model is improved. However, they sometimes start with a fixed number of clusters and they do not use the same concept of density.

*Categorical Data Clustering:* These algorithms are specifically developed for data where Euclidean, or other numerical-oriented, distance measures cannot be applied. In the literature, we find approaches close to both partitional and hierarchical methods (Andritsos P., 2002).

### **2.3.3 Types of Clusters**

Clustering target is to find useful clusters, where usefulness is defined by the goals of the data analysis. There are many deferent notions of a cluster that are useful in practice. In order to visually illustrate the deference among these types of clusters, we use two-dimensional points, as our data objects. We stress, however, that the types of clusters described here are equally valid for other kinds of data.

Well-Separated: A cluster is a set of objects in which each object is closer to each other object in the cluster than to any object out of the cluster. Sometimes a threshold is used to specify that all the objects in a cluster must have distance between each other. This idealistic definition of a cluster is satisfied only when the data contains natural clusters that are quite far from each other. The distance between any two points in different groups is larger than the distance between any two points within a group. Well-separated clusters do not need to be globular, but can have any shape.

Prototype-Based: A cluster is a group of objects in which each object is closer to the prototype that defines the cluster than to the prototype of any other cluster. For data with continuous attributes, the prototype of a cluster is often a centroid. When a centroid is not meaningful, such as when the data has categorical attributes, the prototype is often a medoid. For many types of data, the prototype can be considered as the most central point, and in such occasions, we commonly refer to prototype-based clusters as center-based clusters. These kind of clusters tend to be globular.

Graph-Based: If the data is represented as a graph, where the nodes are objects and the links represent connections among objects, then a cluster can be defined as a connected component. An important example of graph-based clusters are contiguity-based clusters, where connection exists between two objects only if they are within a specified distance of each other. This implies that each object in a contiguity-based cluster is closer to some other object in the cluster than to any point in a different cluster. This definition of a cluster is useful when clusters are irregular or intertwined, but can have trouble when noise is present since. Other types of graph-based clusters are also possible. In case we add connections between objects in the order of their distance from one another, a cluster is formed when a set of objects forms a clique. Like prototype-based clusters, such clusters tend to be globular.

Density-Based: A cluster is a dense region of objects that is surrounded by a region of low density. The two circular clusters are not merged, because the bridge between them fades into the noise. A density-based definition of a cluster is often employed when the clusters are irregular or intertwined, and when noise and outliers are present.

Shared-Property: In general, we can define a cluster as a group of objects that share some property. This definition encompasses all the previous definitions of a cluster. However, this approach also includes new types of clusters. A triangular cluster is adjacent to a rectangular one, and there are two intertwined clusters. In both cases, a clustering algorithm would need a very specific concept of a cluster to successfully detect these clusters. These process where we find new clusters is described as conceptual clustering. However, too sophisticated a notion of a cluster would take us into the area of pattern recognition, and thus, we only consider simpler types of clusters in this book (Tan et al, 2006).

## **2.4 Routing Problems**

### **2.4.1 Introduction to Routing Problems**

Design of routes for people or transportation systems is nowadays one of the main problems in urban service systems. Routing problems can be referred as the operational part in the management of distribution and transportation networks. On the one hand, we have cases where routes must be designed so that they traverse in an exhaustive way the streets in a neighbourhood or in a specific part of a city or, occasionally, in a whole city. On the other hand, we have objectives that must visit a number of some given geographical points in order to provide some service or deliver and collect. Decisions related to facilities number and location, fleet size, customer-depot allotment, and allotment of transportation services between locations concern the design of the network and may be viewed as strategic and tactical. The daily problem of routing vehicles to deliver goods from local depots to

customers can be classified as operational. This distinction made does not only have to do with the decisions involved but also with the time-span and frequency of these decisions.

For the first type of routing problems, we can set examples such as street cleaning, mail delivery, trash collection from houses and leaflet and flyer distribution. Some examples we can see for the point-visiting type of routing problem are urban buses routing, newspaper distribution to kiosks, the routine inspection of vending machines, and the delivery of shipments to addresses.

These problems are categorized into two types, *edge-covering* and *node-covering* problems. Specific subsets of these problems have been of great interest on the fields of mathematics and operational research. For instance, there is a large number of scientific reports and papers writing about the famous travelling salesman problem(TSP), which is the most well-known node-covering problem.

In fact, in real applications, many additional complications can be met in practice when we come to routing problems. This can make the solution of the problem a difficult task. Some typical complications include:

- ✓ the presence of time windows associated with each customer, i.e., an earliest and latest time at which the customer can be visited;
- ✓ the presence of pickup and delivery transportation requests, i.e., customers requiring that the associated load is picked up at a specified location and delivered to a corresponding delivery location;
- ✓ the presence of multiple depots
- ✓ a planning horizon of several days in which each customer requires to be visited a given number of times according to specific day combinations (Bartolini E. et al, 2009).

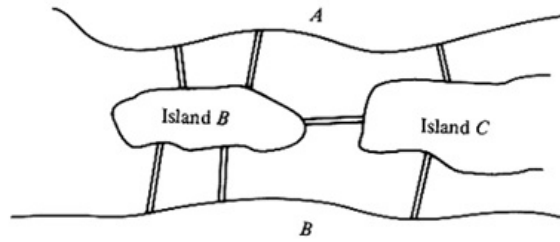
### 2.4.2 Routing Problem Types

#### **TSP:**

The description of the travelling salesman problem (TSP) is so easy but one can find great difficulties when it comes to its solution, this is the reason mathematicians and computer scientists, have shown great interest for this problem. In TSP we are trying to find the route that will cost the traveling salesman the least and can take him to visit exactly once each of a list of  $m$  spots and then return to the home spot. The TSP is the most famous of a larger class of problems known as combinatorial optimization problems. It is included in a set of such problems known as NP-complete. So efficient algorithms could be found for all other problems in the NP-complete class as long as one can be found for the TSP. Until today this kind of algorithm has not yet been found for the TSP. Nowadays we can see that many practical optimization problems of truly large scale are solved to optimality routinely. So the question of what it is that makes a problem difficult may remain unanswered, the computational record of specific instances of TSP problems taken from practical applications is optimistic.

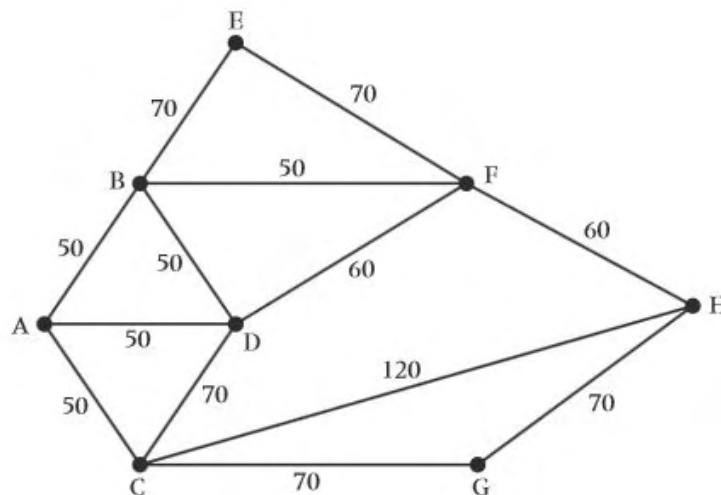
#### **CPP:**

Königsberg (now Kaliningrad) is a Russian city in which in the 18th century there were seven bridges that spanned a forked river (Pregel) that flows past an island. One day a man who wished to walk along each bridge, thought of the possibility to do this but with the restriction that a bridge must only be traversed at least once. The man thought if there exists such a continuous tour which satisfies the requirement. This problem was solved in 1735 in an ingenious way by the Swiss mathematician and physicist Leonhard Euler. His solution of the problem was the start of a very important branch of mathematics called Graph Theory.



**Figure 2.4.1.** *Seven-Bridge problem at Königsberg*

A similar problem famous in the field of Graph Theory is the Chinese Postman Problem, also known as Postman Tour or Route Inspection Problem. The name comes from the fact that an early paper discussing this problem appeared in the journal *Chinese Mathematics*. In this problem, the postman must deliver all the mail to the city using the shortest path possible. This is done when the postman passes each street once and then returns to the start. Consider the case of a mailman who is responsible for the delivery of mail in a city area. The mailman must always begin his delivery route at a start location (the post office), must traverse every single street in this area and, eventually, must return to the start location. Take each location he must deliver a point and the streets that connect these locations as edges. There is an example in the Figure 2.4.2 below



**Figure 2.4.2.** *Example of Chinese Postman Problem graph*

The question here is which will be the design of the mailman's route in order to minimize the total walking distance, while at the same time traversing every street at least once. This type of edge-covering problem is known as the *Chinese postman's problem*.

Nowadays, efficient algorithms exist for solving the CPP on *undirected* graphs and on *directed* graphs. Many researchers tried to develop a similarly efficient procedure for solving the CPP on a *mixed* graph, but as it seems this last problem belongs to a class of very hard problems for which it is unlikely that polynomial algorithms will ever be found.

### **RPP:**

The rural postman problem (RPP) is a general case of Chinese Postman Problem where a subset of the set of links of a given graph is “required” to be traversed at a minimum cost. If this subset does not form a connected graph but forms a number of disconnected components the problem is NP-complete and is also a generalization of the travelling problem.

### **GPDP:**

In the General Pickup and Delivery Problem (GPDP) a group of routes must to be developed in order to satisfy transportation requests. A number of vehicles is available to operate the routes. A given capacity, a start location and an end location corresponds to each vehicle. Each transportation request specifies the size of the load to be transported, the locations where it is to be picked up and the locations where it is to be delivered. Each load has to be transported by one vehicle from its set of origins to its set of destinations without any transshipment at any other locations.

Three well-known and extensively studied routing problems are special cases of the GPDP. The Pickup and Delivery Problem (PDP), the Dial-a-Ride Problem (DARP) and the Vehicle Routing Problem (VRP). In the PDP, each transportation request specifies a single origin and a single destination and all vehicles depart from and return to a central depot. The



DARP is a PDP in which the loads to be transported represent people. Therefore, we usually speak of clients or customers instead of transportation requests and all load sizes are equal to one. The VRP is a PDP in which either all the origins or all the destinations are located at the depot.

The GPDP is introduced in order to be able to deal with various complicating characteristics found in many practical pickup and delivery problems, such as transportation requests specifying a set of origins associated with a single destination or a single origin associated with a set of destinations, vehicles with different start and end locations, and transportation requests evolving in real time (M.W.P Savelsberh, M. Sol, 2005).

### **2.4.3 The Dial-A-Ride Problem**

The Dial-a-Ride Problem (DARP) consists of designing vehicle routes and time schedules for  $n$  users who specify pickup and delivery requests between origins and destinations. The aim is to plan a set of  $m$  minimum cost vehicle routes capable of accommodating as many users as possible, under a set of constraints. The most common example arises in door-to-door transportation for elderly or disabled people.

From a modeling point of view, the DARP generalizes a number of vehicle routing problems such as the Pickup and Delivery Vehicle Routing Problem (PDVRP) and the Vehicle Routing Problem with Time Windows (VRPTW). What makes the DARP different from most such routing problems is the human perspective. When transporting passengers, reducing user inconvenience must be balanced against minimizing operating costs. In addition, vehicle capacity is normally constraining in the DARP whereas it is often redundant in PDVRP applications, particularly those related to the collection and delivery of letters and small parcels (Molenbrunch et al, 2017).

Dial-a-ride services may operate according to a static or to a dynamic mode. In the first case, all transportation requests are known beforehand, while in the second case requests are gradually revealed throughout the day and vehicle routes are adjusted in real-time to meet demand. In practice pure dynamic DARPs rarely exist since a subset of requests is often known in advance. At this point we should mention that after more than twenty years of research, excellent heuristics exist for the static case. It is now possible to solve instances with several hundreds of users within reasonable times and it should be possible to apply decomposition techniques for larger instances involving, say, two or three thousand users. The dynamic version of the problem has to be studied more. This involves the construction of an initial solution for a limited set of requests known in advance and the design of features capable of determining whether a new request should be served or not and if so, how existing routes should be modified to accommodate it. Also it should be possible to update a partially built solution to deal with cancellations and other unforeseen events such as traffic delays and vehicle breakdowns. Finally, advanced systems should make full use of new technologies such as vehicle positioning systems now common in the area of emergency medical services (Brotcorne et al. 2003).

Most studies on the DARP assume the availability of a fleet of  $m$  homogeneous vehicles based at a single depot. There may be several depots, especially in wide geographical areas, and the fleet is sometimes heterogeneous. Some vehicles are designed to carry wheelchairs only, others may only cater to ambulatory passengers and some are capable of accommodating both types of passenger. The main consideration in some problems is to first determine a fleet size and composition capable of satisfying all demand, while in other contexts, the aim is to maximize the number of requests that can be served with a fixed size fleet. Some systems routinely turn down several requests each day. A compromise consists of

serving some of the demand with a core vehicle fleet and using extra vehicles (e.g., taxis) if necessary. We can consider two possible problems:

- Minimize costs subject to full demand satisfaction and side constraints.
- Maximize satisfied demand subject to vehicle availability and side constraints.

The most common cost elements relate to regular fleet size and operation, occasional use of extra vehicles, and drivers' wages. Quality of service criteria include route duration, route length, customer waiting time, customer ride time (i.e., total time spent in vehicles), and difference between actual and desired delivery times. Some of these criteria may be treated as constraints or as part of the objective function. A common trend in DARP models is to let users impose a time window on both their departure and arrival times, but this may be unduly constraining for the transporter, particularly if these time windows are narrow. According to Jaw et al. (1986) users should be able to specify a time window on the arrival time of their outbound trip and on the departure time of their inbound trip. The transporter then determines a planned departure time for the outbound trip and a planned arrival time for the inbound trip, while satisfying an upper bound on the ride time. In practice, since travel times are somewhat uncertain, the outbound departure time communicated to the user should be slightly earlier than the scheduled time.

Several models have been suggested for a number of variants of the DARP for example:

- *A three-index formulation for the DARP:* In this formulation, constraints ensure that each request is served once by the same vehicle and guarantee that each vehicle starts and ends its route at the depot. Also they define starts of service times, vehicle loads and user ride times, respectively, while other constraints ensure that these will be feasible.
- *A two-index formulation for the DARP:* Ropke et al. (2007) have proposed two models and a branch-and-cut algorithm for the PDP with time windows (PDPTW)

and for the DARP, where all vehicles are identical. The PDPTW is a DARP without the maximum ride time constraints (Cordeau & Laporte, 2007).

One of the simplest cases of the DARP is that where all users are served by a single vehicle, *The single-vehicle DARP*. There are algorithms for the static case, where all requests are known in advance, and for the dynamic case, where they are gradually revealed in real-time. In this study we use the single-vehicle DARP so as to find the order in which the incidents will be served by the one patrol.

#### **2.4.4 Vehicle Routing Problem**

The globalization of the economy leads to a rapidly growing trade of goods in our world. Limited merchandise and transportation resources, high planning complexity and the continuously increasing cost pressure through strong competition between logistics companies make it essential to use computer-aided systems for the planning of the transports. An important subtask in this context is the operational planning of specialized transportation vehicles. These optimization tasks are called Vehicle Routing Problems (VRPs). These problems of finding optimal routes for groups of vehicles belongs to the class of NP-hard combinatorial problems. The practical and theoretical importance of this NP-hard optimization problem and its varieties has been the subject over hundred of scientific papers. Therefore, many specific solvers for different Vehicle Routing Problems can be found in the literature. The negative here is that most of these solvers are highly specialized and inflexible and it needs a lot of effort to adapt them to modified problems. Also, most real-world problems are often much more complicated than the idealized problems out of literature and they also change over time.

The classical VRP is one of the most famous problems in combinatorial optimization, and its study has given rise to several exact and heuristic solution techniques of general

applicability. It generalizes the Traveling Salesman Problem (TSP) and is often defined under capacity and route length restrictions. For example, when vehicle capacity constraints are present the problem is denoted as Capacitated Vehicle Routing Problem (CVRP) and when we have to deal with a time interval in which each customer has to be served we are talking about a Vehicle Routing Problem with Time Windows (VRPTW).

The Vehicle Routing Problem (VRP) is a problem where one must design optimal routes for delivery or collection from one or several depots to a number of geographically scattered points (cities, customers, etc.). The VRP has an important role in the fields of physical distribution and logistics. There exists a large variety of VRPs and a broad literature on this type of problems.

The purpose of the classical VRP is to find a set of routes at a minimal cost (finding the shortest path, minimizing the number of vehicles, etc) starting and ending the route at the depot so that the demands of all nodes are reached. Each node can be visited only once, by only one vehicle, and each vehicle has a limited capacity. Some formulations also present constraints on the maximum travelling time (Barnhart & Laporte, 2007) (Laporte, 1991) (Caric & Gold, 2008).

### ***Variants of VRP***

#### **CVRP:**

The basic and probably most well-studied vehicle routing problem is the Capacitated Vehicle Routing Problem (CVRP). The CVRP is one of the most studied combinatorial optimization problems and since it was first proposed it has been studied in many variants. CVRP was first defined by Dantzig and Ramser in 1959. In that study, they used distance as a surrogate for the cost function. Since then, the cost of traveling from node  $i$  to node  $j$ , i.e.,  $c_{ij}$ , has usually been taken as the distance between those nodes. The real cost of a vehicle

traveling between two nodes depends on many variables: the price of the fuel, the vehicle load, the maintenance needed, the time for visiting the customers, the fuel consumption, the distance traveled to a given node, the depreciation of the tires and the vehicle, driver wages, total distance traveled, etc. Most of the attributes are actually distance or time based and can be approximated by the distance. In the CVRP a set of customers, each with an associated requirement of some product, must be supplied from a single depot by a homogeneous fleet of vehicles of known capacity. The problem is to design a set of simple circuits for the vehicles, called routes, starting and ending at the depot, and such that each customer is supplied by exactly one route. The total load of each route must not exceed the vehicle capacity, and the objective is to minimize the sum of route costs.

The Capacitated Vehicle Routing Problem (CVRP) is defined on a graph  $G = (V, A)$  where  $V = \{0, 1, 2, \dots, n\}$  is the set of nodes, 0 is the depot and the remaining nodes are customers. The set  $A = \{(i, j): i, j \in V, i \neq j\}$  is an arc (or edge) set. Each customer  $i \in V \setminus \{0\}$  is associated with a positive integer demand  $q_i$  and each arc  $(i, j)$  is associated with a travel cost  $c_{ij}$ . There are  $m$  vehicles with identical capacity  $Q$ . The CVRP consists of determining a set of  $m$  vehicle routes satisfying the following conditions:

- Each route starts and ends at the depot,
- Each customer is visited by exactly one route,
- The total demand of each route does not exceed the vehicle capacity  $Q$ ,
- The total cost of all routes is minimized.

The CVRP is NP-hard as it contains the well-known Traveling Salesman Problem (TSP) as a special case. In fact, the TSP corresponds to a CVRP where a single vehicle of capacity greater or equal to the sum of all customers demands is available at the depot (Bartolini E. & Marzo 2009) (Caric & Gold, 2008).

### **VRPTW:**

Time-constrained routing problems have grown as an important research area in the last twenty years as a consequence of the increasing importance of the time dimension for manufacturing and transportation companies engaged in an effort to compete on service quality. In many applications such as bank and postal deliveries, industrial refuse collection or school bus routing, customer service must occur according to strict time constraints called time windows. Time windows correspond to given time intervals, associated with each customer, imposing that a customer can only be visited at a time within its time window. The VRP with Time Windows (VRPTW) is a generalization of the CVRP that models this kind of problems. In the VRPTW a travel time is associated with each arc and with each customer are associated both a time window and a service time. Usually, vehicles are allowed to arrive at a customer location before the earliest time imposed by the corresponding time window, but in this case, the vehicle must wait until the customer earliest time before starting to service it. The objective of the VRPTW is to design a set of routes of minimum cost to serve all customers within the imposed time windows without exceeding the vehicle capacity. In some applications involving flexible time schedules it is permitted to violate time window constraints at a cost, usually a linear function of the amount of time window violation. In this case time windows are called soft. (Bartolini E. & Marzo 2009)

The VRPTW is an important generalization of the classical VRP in which service at every customer  $i$  must begin within a given time window  $[a_i, b_i]$ . A vehicle is allowed to arrive before  $a_i$  and wait until the customer becomes available, but arrivals after  $b_i$  are prohibited. The VRPTW has numerous applications in distribution management. Common examples are delivery of food, beverages and newspapers and commercial and industrial waste collection. The VRPTW is NP-hard since it generalizes the CVRP which is obtained when  $a_i = 0$  and  $b_i = \infty$  for every customer  $i$ . In the case of a fixed fleet size, even finding a feasible solution to

the VRPTW is itself an NP-complete problem. As a result, research on the VRPTW has concentrated on heuristics. Nevertheless, when the problem is sufficiently constrained, realistic size instances can be solved optimally through mathematical programming techniques. This section presents a mathematical formulation of the VRPTW followed by a description of some of the most important available exact and heuristic algorithms. It is worth pointing out that while exact methods usually minimize distance, most heuristics consider a hierarchical objective which first minimizes the number.

### **VRPSD:**

The Vehicle Routing Problem with Stochastic Demand (VRPSD) is a variation of the classical VRP, where each customer can be satisfied by more than one vehicle. So, for the VRPSD, except the delivery routes, we have to determine the amount to be delivered to each customer in each vehicle. The option of splitting a demand makes it possible to satisfy a customer whose demand exceeds the vehicle capacity. Splitting may also allow decreasing costs (Caric & Gold, 2008).

### **IRP:**

The Inventory Routing Problem (IRP) is an important variant of the VRP which integrates routing decisions with inventory control. We find this problem in environments where Vendor Managed Inventory (VMI) resupply policies are employed. These policies allow a vendor to choose the timing and size of deliveries. In exchange for this freedom, the vendor agrees to ensure that there is sufficient product for its customers. In a more traditional relationship, where customers call in their orders, large inefficiencies can occur due to the timing of customers' orders. Realizing the cost savings opportunities of vendor managed inventory policies, however, is not a simple task, particularly with a large number and variety of customers. The inventory routing problem achieves this goal by determining a distribution



strategy that brings long-term distribution costs to the minimum. This description of the inventory routing problem focuses primarily on distribution. Inventory control is restricted to ensuring that no stock outs occur at the customers. Inventory control takes a more prominent role when inventory holding costs are taken into consideration. In the inventory control literature, the resulting environment is usually referred to as a one warehouse multi retailer system. IRPs are very different from VRPs. VRPs are found when customers place orders and the vendor, on any given day, assigns the orders for that day to vehicle routes. In IRPs it is the delivery company and not the customer that decides the amount of product for delivery to which customers each day. Customer orders do not exist. Instead, the delivery company.

### **3. THE CASE STUDY OF VOLOS**

#### **3.1 The Police Department of Volos**

In this thesis we have processed and studied data which refer to thefts as criminal offences, in the city of Volos, capital of the district of Magnesia, for the period 2010 - 2017. The data was provided by the Greek Police and in particular by the Security Department of Volos. At this point we should refer to the Greek Police and the Police Department of Volos.

The Greek Police in its current form was created in 1984, with the merger of the Gendarmerie and the City Police (Law 1481/1-10-1984, Official Government Gazette A' 152). According to Law 2800/2000 Greek Police is a Security Body and its mission is:

- Securing public peace, order and the unhindered social welfare of citizens including the policing, general policing and traffic police.
- Prevention and suppression of crime and protection of the State and Democratic Republic within the framework of the constitutional order, which includes the exercise of the public and the state security police.
- Prevention of illegal entry and exit of foreigners in Greece and control of the observance of the provisions regarding the entry, exit, stay and work of the foreigners in the country, including the exercise of the foreign police and border protection.

The Greek Police is constituted by Central and Regional services. The Headquarters of the Hellenic Police is the supreme authority of these services. Its task is to ensure that the Police Force fulfills its mission as part of the policy of the Hellenic Ministry of Citizen Protection. To this end it plans, directs, monitors and controls the activities of the services and ensures the necessary conditions for the exercise of their responsibilities. The responsibilities of the police services are indicated in the following table.

**Table 3.1:** The responsibilities of the police services (P.D. 7/2017 Articles 95 & 97)

<b>Safety Department (Public-State)</b>	<b>Office tackling racist violence</b>	<b>Department of foreigners</b>	<b>Department of drugs</b>
<p>a. The prosecution of crimes against life, property and other property rights,</p> <p>b. the prosecution of financial and electronic crime,</p> <p>c. control and prosecution of illicit drug trafficking,</p> <p>d. the prosecution of smuggling and antiquities,</p> <p>e. the concern for the protection of minors and the implementation of the provisions on morals,</p> <p>f. checking compliance with the provisions concerning the memoirs and the protection of the national currency and exchange</p> <p>g. surveillance of the places where suspected criminals are frequented and the control of those persons.</p> <p>i. the search for missing persons and lost and stolen objects,</p> <p>j. search and arrest of persecuted persons.</p> <p>k. the protection of the State and of the democratic system,</p>	<p>a. intervene of its own motion or following a relevant indictment, indictment or complaint in the investigation and prosecution of crimes relating to the commission, in the preparation or in any way of public incitement, provocation or aggravation of acts, offenses or manifestations of acts or actions which may be discriminatory, hatred or violence against persons or a group of persons by reason of their race, color, religion, birth, national or ethnic origin,</p> <p>b. collect, process and use appropriate information and data relating to the commission or preparation of offenses with racist characteristics,</p> <p>c. develops collaborations with co-competent government agencies and bodies, as well as other social organizations and organizations, in the fulfillment of its mission and more effective management of incidents of racist violence,</p> <p>d. Initiates or provides assistance to initiatives by other authorities, agencies and agencies aimed at preventing and tackling racist violence,</p>	<p>a. control of the legal residence and work of foreigners,</p> <p>b. the implementation of legislation on foreigners,</p> <p>c. the handling of issues relating to the adoption of administrative acts and measures relating to foreigners,</p> <p>d. the handling of political asylum and refugee issues and the implementation of relevant legislation,</p> <p>e. controlling the legal movement of nationals and foreigners and applying the law on the protection of the national currency and foreign exchange, if the area of the Sub-Directorate is a point of entry and exit from the country.</p>	<p>a. control of drug and psychotropic drug trafficking, pre-trial action and drafting of relevant cases,</p> <p>b. the monitoring of suspected drug trafficking, both domestic and foreign, for the purpose of preventing and suppressing related crimes,</p> <p>c. control, in cooperation with the competent agencies and services, of airports, ports, railway stations and car stations, drug trafficking suspects,</p> <p>d. the supervision, in particular of public areas suitable for the transport, and trafficking of drugs, as well as of public centers suspected of trafficking or drug use,</p> <p>e. co-operation with local authorities, agencies and also Local Counterfeiting Prevention Councils to tackle drug-related crime more effectively,</p> <p>f. Surveillance of neighborhood schools and areas frequented by young people in order to effectively protect pupils and youth in general from drug-related crime,</p> <p>g. monitoring the methods and means of action of drug dealers, analyzing and evaluating the</p>

	<p>e. oversees areas and areas where there is an increased risk of racist attacks,</p> <p>f. cooperate with international organizations and agencies as well as with agencies and representatives of sensitive social groups who have been or are at risk of receiving racist attacks for the purpose of more effective incident management and fuller protection for these groups,</p> <p>g. maintains a special record of incidents of racist violence,</p> <p>h. informs victims or complainants of racist violence about their rights,</p> <p>i. ensure that the competent services are informed in cases where medical or nursing care or psychological support is required for victims of racist attacks and the presence of an interpreter if so requested or deemed necessary,</p>		<p>conclusions and recommending measures aimed at tackling drug-related crime more effectively, both in the preventive and the suppressive sectors,</p> <p>h. developing a network for collecting information on the methods, sites, modes, means and persons used or involved in any way in the handling, promotion, distribution and use of narcotic drugs and drugs;</p> <p>i. informing the head of department of the information material gathered so that it can be properly exploited.</p>
--	--	--	---

Since 2017 have been attached to the Safety Department, whose responsibilities are listed in table 3.1, the Crime Prevention and Suppression Groups named « ΟΠΙΚΕ » in greek. Their mission is to intervene in areas where serious crime (theft, robbery, blackmailing, drug trafficking, etc.) takes place and measures are needed, in addition to those developed in the context of routine policing, in order to deal effectively with delinquency, the creation of a safe environment and the consolidation of citizens' sense of security.

Volos Security Sub directorate belongs administratively to Magnesia Police Directorate. Its territorial jurisdiction includes the administrative region of municipal units of Aisionia, Volos and Nea Ionia, the municipal unit of Agria of Volos Municipality and the part of the hill named « Goritsa » of Portaria's municipal unit of Volos Municipality. that is enclosed by the stream named « Rema Karias », the contribution of the stream « Rema Karias » with the flood protection ditch until the labeled boundary line of the cement factory « AGET Iraklis » which follows and ends east of the fuel depots « Elinoil » to the sea (No. 7001/2/1478-μβ').

The above apply from May 2017, while the main difference from the previous years is that the municipal unit of Agria was not included in the territorial jurisdiction of Volos Security Sub directorate.

## **3.2 Police data**

### **3.2.1 Type and format of data**

The Police Department of Volos (PDV) provided data regarding incidents occurred in the city of Volos, capital of the district of Magnesia. The type of incidents is related to burglaries, car robberies and motorcycle robberies from the past eight years. More specifically, the data is dated from January 2010 until December 2017.

PDV records and stores all incidents electronically in a database system according to a specific format. Therefore, the incident data were handed over in a spreadsheet format. In Figure 3.2.1-1, the format of the data is illustrated.

	A	B	C	D	E	F
1	Id	Date	Time	Offense	Type of Offense	Address
2	1	1/1/2010	10:00:00 AM	ΚΛΟΠΗ	ΠΛΗΜ/ΜΑ	ΙΩΑΚΕΙΜ ΜΗΤΡΟΠΟΛΙΤΟΥ 39,Ν.Ιωνία
3	2	1/1/2010	12:00:00 AM	ΚΛΟΠΗ ΑΥΤ/ΤΩΝ	ΠΛΗΜ/ΜΑ	ΧΑΤΖΗΠΕΤΡΟΥ 5,Βόλος
4	3	1/3/2010	1:00:00 PM	ΚΛΟΠΗ ΑΥΤ/ΤΩΝ	ΠΛΗΜ/ΜΑ	ΦΙΛΙΚΗΣ ΕΤΑΙΡΕΙΑΣ 56,Βόλος
5	4	1/3/2010	4:00:00 AM	ΚΛΟΠΗ ΔΙΚΥΚΛΩΝ	ΠΛΗΜ/ΜΑ	ΒΛΑΧΑΒΑ 73,Βόλος
6	5	1/4/2010	3:00:00 AM	ΚΛΟΠΗ ΔΙΚΥΚΛΩΝ	ΠΛΗΜ/ΜΑ	ΕΘΝΙΚΩΝ ΑΓΩΝΩΝ 04,Βόλος
7	6	1/8/2010	6:00:00 AM	ΚΛΟΠΗ	ΠΛΗΜ/ΜΑ	ΠΑΠΑΔΙΑΜΑΝΤΗ 1,Βόλος
8	7	1/8/2010	12:01:00 AM	ΚΛΟΠΗ	ΠΛΗΜ/ΜΑ	ΘΡΑΚΩΝ 28,Βόλος
9	8	1/9/2010	11:00:00 AM	ΚΛΟΠΗ	ΠΛΗΜ/ΜΑ	ΚΟΥΝΤΟΥΡΙΩΤΟΥ ,Βόλος
10	9	1/10/2010	9:00:00 PM	ΚΛΟΠΗ	ΠΛΗΜ/ΜΑ	ΚΟΝΤΑΡΑΤΟΥ ΔΗΜΑΡΧΟΥ 26,Βόλος
11	10	1/11/2010	10:00:00 AM	ΚΛΟΠΗ ΑΥΤ/ΤΩΝ	ΠΛΗΜ/ΜΑ	ΜΠΟΥΚΟΥΒΑΛΑ 1,Βόλος
12	11	1/11/2010	1:00:00 PM	ΚΛΟΠΗ ΑΥΤ/ΤΩΝ	ΠΛΗΜ/ΜΑ	ΑΝΑΓΝΩΣΤΟΠΟΥΛΟΥ 30,Βόλος
13	12	1/11/2010	11:00:00 AM	ΚΛΟΠΗ ΑΥΤ/ΤΩΝ	ΠΛΗΜ/ΜΑ	ΣΠΥΡΙΔΗ ,Βόλος
14	13	1/11/2010	8:00:00 AM	ΚΛΟΠΗ ΔΙΚΥΚΛΩΝ	ΠΛΗΜ/ΜΑ	ΔΗΜΟΥ ΓΙΑΝΝΗ 48,Βόλος
15	14	1/11/2010	8:00:00 AM	ΚΛΟΠΗ ΔΙΚΥΚΛΩΝ	ΠΛΗΜ/ΜΑ	ΑΝΑΛΗΨΕΩΣ 61,Βόλος
16	15	1/12/2010	2:30:00 AM	ΚΛΟΠΗ	ΠΛΗΜ/ΜΑ	ΧΑΤΖΗΜΙΧΑΛΗ 3Γ,Βόλος
17	16	1/13/2010	10:00:00 AM	ΚΛΟΠΗ ΑΥΤ/ΤΩΝ	ΠΛΗΜ/ΜΑ	ΙΩΛΚΟΥ 250,Βόλος
18	17	1/13/2010	1:00:00 PM	ΚΛΟΠΗ ΑΥΤ/ΤΩΝ	ΠΛΗΜ/ΜΑ	ΑΛΜΥΡΟΥ ,Βόλος
19	18	1/13/2010	9:15:00 PM	ΚΛΟΠΗ ΔΙΚΥΚΛΩΝ	ΠΛΗΜ/ΜΑ	ΦΙΛΙΠΠΟΥ ΙΩΑΝΝΗ ,Βόλος
20	19	1/13/2010	3:00:00 AM	ΚΛΟΠΗ ΔΙΚΥΚΛΩΝ	ΠΛΗΜ/ΜΑ	ΦΙΛΙΚΗΣ ΕΤΑΙΡΕΙΑΣ ,Βόλος

Figure 3.2.1-1: Format of data

There is particular information that has to be recorded for every incident as it is obvious from Figure 3.2.1-1. This information is: the identification number (id), the date, the time, the type of offence, the level of offence and the address of the incident. The id is a unique numerical digit assigned to every incident in order to differentiate it from the other incidents. The date and time of every incident are recorded, so that the actual point in time when the incident occurred will be known. This information is valuable for statistical analysis and forecasting. The type of offence is related to the substantive type of the incident. In this study the available data is related to burglaries, car robberies and motorcycle robberies. The level of offence is related to the legal nature of the incident. In this study we do not focus on this feature, as it is related to law and justice studies. However, it is worth mentioning that most of the offences are misdemeanors (minor crimes) and there is also a small percentage of crimes which were recorded as felonies

(serious crimes). The address, as the last feature of every recorded incident, defines the location of the incident. This attribute is substantial, because in this study the location is translated into terms of coordinates, latitude (lat) and longitude (lon) for every incident. The coordinates can be used in order to determine the graph of the incidents and apply algorithms that can provide valuable information regarding the area with the highest rate of crime, the median and center points of the graph where the patrol car can stake out and the planning of route of the patrol car on how police forces can intervene in the incidents more efficiently.

Having data with adequate attributes and information contributes to efficient data analysis. However, in order to produce meaningful and practical outcomes, it is important to review and validate the feature values of the data. In the chapter 3.2.2, the validation and cleaning activities done for this study are presented in detail.

### **3.2.2 Validation and cleaning of data**

The data of crime incidents from 2010 until 2017 in the city of Volos was reviewed in detail and validated before proceeding with the analysis and any kind of calculations. This is a substantial preliminary activity that should be implemented in every data set in order to avoid inaccurate outcomes during the data analysis activity.

With the assistance of the police officers, the data was double checked and amendments were implemented. In some incidents, the recording of the address was not accurate. An empirical address was stored instead of the official address where the incident happened. The use of an address based on experience causes an issue regarding the estimation of the coordinates of the incident, because Google Maps could not recognize it. For this purpose, all incidents stored with an empirical address were corrected based on police officers experience, so that they have the official address.

In parallel with data validation, data cleaning was implemented also. This activity is related to the exclusion of incomplete and irrelevant data from the initial dataset. It is required in order to avoid inaccurate outcomes during the statistical and computational analysis. In this study there were two cases regarding data cleaning. Recorded incidents with incomplete information and incidents that were occurred far away from the city and the suburbs of the city of Volos needed to be removed from the initial dataset.

Regarding incidents with incomplete information, a small number of data found without the feature of time and/or address. These two characteristics are important for the analysis. Thus, the incidents with shortage of time and/or address were deleted from the dataset. On the other hand, all incidents located far away from the city and suburbs of Volos, e.g. robberies occurred on Skiathos island, were excluded from the dataset and further statistical analysis, because they might have affected the calculations and caused inaccurate results.

With the completion of these two activities, the data is ready to be used for processing and extraction of useful information. In the subchapter 3.2.3, the transformation of field values and processing of data is presented thoroughly.

### **3.2.3 Transformation and processing of data**

Having accurate data after validation and cleaning activities, it is time to proceed with the transformation of the field values into practical numeric values, which are going to be used in the calculations. In this case study, the values of the address feature are needed to be transformed in terms of coordinates (latitude, longitude). For this purpose the map tool <http://194.177.201.113/loisgmaps> was used. This tool is a web interface application created by Professor Athanasios Lois based on Google Maps. By uploading a simple text file including in every row the id and the address information of each incident on this tool, the outcome is a simple text file with the id and the coordinates of each incident. The coordinates were stored back into



the spreadsheet with the incidents, so they can be used later in data analysis. Figure 3.2.3-1 illustrates how the spreadsheet looks like after the addition of coordinates into the data.

1	Id	Date	Time	Offense	Type of Offense	Address	Latitude	Longitude
2	1	1/1/2010	10:00:00 AM	ΚΛΟΠΗ	ΠΛΗΜ/ΜΑ	ΙΩΑΚΕΙΜ ΜΗΤΡΟΠΟΛΙΤΟΥ 39,N.Ιωνία	39.3770138	22.9548655
3	2	1/1/2010	12:00:00 AM	ΚΛΟΠΗ ΑΥΤ/ΤΩΝ	ΠΛΗΜ/ΜΑ	ΧΑΤΖΗΠΕΤΡΟΥ 5,Βόλος	39.3549204	22.9228793
4	3	1/3/2010	1:00:00 PM	ΚΛΟΠΗ ΑΥΤ/ΤΩΝ	ΠΛΗΜ/ΜΑ	ΦΙΛΙΚΗΣ ΕΤΑΙΡΕΙΑΣ 56,Βόλος	39.3643874	22.9250742
5	4	1/3/2010	4:00:00 AM	ΚΛΟΠΗ ΔΙΚΥΚΛΩΝ	ΠΛΗΜ/ΜΑ	ΒΛΑΧΑΒΑ 73,Βόλος	39.3603123	22.9589756
6	5	1/4/2010	3:00:00 AM	ΚΛΟΠΗ ΔΙΚΥΚΛΩΝ	ΠΛΗΜ/ΜΑ	ΕΘΝΙΚΩΝ ΑΓΩΝΩΝ 04,Βόλος	39.3744956	22.9305301
7	6	1/8/2010	6:00:00 AM	ΚΛΟΠΗ	ΠΛΗΜ/ΜΑ	ΠΑΠΑΔΙΑΜΑΝΤΗ 1,Βόλος	39.3686572	22.9341718
8	7	1/8/2010	12:01:00 AM	ΚΛΟΠΗ	ΠΛΗΜ/ΜΑ	ΘΡΑΚΩΝ 28,Βόλος	39.3639648	22.9385873
9	8	1/9/2010	11:00:00 AM	ΚΛΟΠΗ	ΠΛΗΜ/ΜΑ	ΚΟΥΝΤΟΥΡΙΩΤΟΥ ,Βόλος	39.3707729	22.9454418
10	9	1/10/2010	9:00:00 PM	ΚΛΟΠΗ	ΠΛΗΜ/ΜΑ	ΚΟΝΤΑΡΑΤΟΥ ΔΗΜΑΡΧΟΥ 26,Βόλος	39.3589409	22.9524679
11	10	1/11/2010	10:00:00 AM	ΚΛΟΠΗ ΑΥΤ/ΤΩΝ	ΠΛΗΜ/ΜΑ	ΜΠΟΥΚΟΥΒΑΛΑ 1,Βόλος	39.3529455	22.9231415
12	11	1/11/2010	1:00:00 PM	ΚΛΟΠΗ ΑΥΤ/ΤΩΝ	ΠΛΗΜ/ΜΑ	ΑΝΑΓΝΩΣΤΟΠΟΥΛΟΥ 30,Βόλος	39.3650311	22.9236221
13	12	1/11/2010	11:00:00 AM	ΚΛΟΠΗ ΑΥΤ/ΤΩΝ	ΠΛΗΜ/ΜΑ	ΣΠΥΡΙΔΗ ,Βόλος	39.3653133	22.9522622
14	13	1/11/2010	8:00:00 AM	ΚΛΟΠΗ ΔΙΚΥΚΛΩΝ	ΠΛΗΜ/ΜΑ	ΔΗΜΟΥ ΓΙΑΝΝΗ 48,Βόλος	39.3709993	22.952842
15	14	1/11/2010	8:00:00 AM	ΚΛΟΠΗ ΔΙΚΥΚΛΩΝ	ΠΛΗΜ/ΜΑ	ΑΝΑΛΗΨΕΩΣ 61,Βόλος	39.3693638	22.9459098
16	15	1/12/2010	2:30:00 AM	ΚΛΟΠΗ	ΠΛΗΜ/ΜΑ	ΧΑΤΖΗΜΙΧΑΛΗ 3Γ,Βόλος	39.3637424	22.9263855
17	16	1/13/2010	10:00:00 AM	ΚΛΟΠΗ ΑΥΤ/ΤΩΝ	ΠΛΗΜ/ΜΑ	ΙΩΑΚΟΥ 250,Βόλος	39.3745925	22.9583204
18	17	1/13/2010	1:00:00 PM	ΚΛΟΠΗ ΑΥΤ/ΤΩΝ	ΠΛΗΜ/ΜΑ	ΑΛΜΥΡΟΥ ,Βόλος	39.3615923	22.9354121
19	18	1/13/2010	9:15:00 PM	ΚΛΟΠΗ ΔΙΚΥΚΛΩΝ	ΠΛΗΜ/ΜΑ	ΦΙΛΙΠΠΟΥ ΙΩΑΝΝΗ ,Βόλος	39.356904	22.9598133

Figure 3.2.3-1: Format of data after the addition of coordinates

Processing of data is the next step, after the transformation of address data into coordinates, in order to extract valuable information from the values of the rest of the features. Regarding date feature, many useful values can be extracted, such as the day, the name of the weekday, the month and the year when the incident took place. For this purpose, Excel functions were applied in the values of date feature. In particular, for extracting the day on which an incident happened, the DAY(<cell\_name\_including\_date\_value>) function was used. This function returns the day of the month, which is a number from 1 to 31, given a date value. In order to extract the month in which an incident happen, the MONTH (<cell\_name\_including\_date\_value>) function was used. This function return the month of the year, which is a number from 1 to 12, given a date value. For the year extraction, the YEAR(<cell\_name\_including\_date\_value>) function was applied. The YEAR function returns the year component of a date as a 4-digit number. In addition, the weekday of each incident based on the date values was estimated by using a combination of CHOOSE and WEEKDAY Excel functions. For example, in order to estimate the weekday on which the incident with id equals to 1 occurred the CHOOSE

(WEEKDAY(B2,2),"Monday","Tuesday","Wednesday","Thursday","Friday","Saturday","Sunday") was applied in the date value of the incident. The returned result is “Friday” and it is validated by using the calendar. In Figure 3.2.3-2, the format of the spreadsheet is presented after the application of the Excel commands.

The spreadsheet is enriched with 6 additional columns. The first 5 additional columns, from G to K, contain numerical values related to coordinates and time data. The last column, L, contains text values related to the weekday on which each incident took place. It is essential to have these additional columns in order to perform further calculations and statistical analysis. In subchapter 3.2.4, the mapping and the clustering of data based on the coordinate values is explained thoroughly.

	A	B	C	D	E	F	G	H	I	J	K	L
1	Id	Date	Time	Offense	Type of Offense	Address	Latitude	Longitude	Year	Month	Day	Weekday
2	1	1/1/2010	10:00:00 AM	ΚΛΟΠΗ	ΠΛΗΜ/ΜΑ	ΙΩΑΚΕΙΜ ΜΗΤΡΟΠΟΛΙΤΟΥ 39, Ν. Ιωνία	39.3770138	22.9548655	2010	1	1	Friday
3	2	1/1/2010	12:00:00 AM	ΚΛΟΠΗ ΑΥΤ/ΤΩΝ	ΠΛΗΜ/ΜΑ	ΧΑΤΖΗΠΕΤΡΟΥ 5, Βόλος	39.3549204	22.9228793	2010	1	1	Friday
4	3	1/3/2010	1:00:00 PM	ΚΛΟΠΗ ΑΥΤ/ΤΩΝ	ΠΛΗΜ/ΜΑ	ΦΙΛΙΚΗΣ ΕΤΑΙΡΕΙΑΣ 56, Βόλος	39.3643874	22.9250742	2010	1	3	Sunday
5	4	1/3/2010	4:00:00 AM	ΚΛΟΠΗ ΔΙΚΥΚΛΩΝ	ΠΛΗΜ/ΜΑ	ΒΛΑΧΑΒΑ 73, Βόλος	39.3603123	22.9589756	2010	1	3	Sunday
6	5	1/4/2010	3:00:00 AM	ΚΛΟΠΗ ΔΙΚΥΚΛΩΝ	ΠΛΗΜ/ΜΑ	ΕΘΝΙΚΩΝ ΑΓΩΝΩΝ 04, Βόλος	39.3744956	22.9305301	2010	1	4	Monday
7	6	1/8/2010	6:00:00 AM	ΚΛΟΠΗ	ΠΛΗΜ/ΜΑ	ΠΑΠΑΔΙΑΜΑΝΤΗ 1, Βόλος	39.3686572	22.9341718	2010	1	8	Friday
8	7	1/8/2010	12:01:00 AM	ΚΛΟΠΗ	ΠΛΗΜ/ΜΑ	ΘΡΑΚΩΝ 28, Βόλος	39.3639648	22.9385873	2010	1	8	Friday
9	8	1/9/2010	11:00:00 AM	ΚΛΟΠΗ	ΠΛΗΜ/ΜΑ	ΚΟΥΝΤΟΥΡΙΩΤΟΥ, Βόλος	39.3707729	22.9454418	2010	1	9	Saturday
10	9	1/10/2010	9:00:00 PM	ΚΛΟΠΗ	ΠΛΗΜ/ΜΑ	ΚΟΝΤΑΡΑΤΟΥ ΔΗΜΑΡΧΟΥ 26, Βόλος	39.3589409	22.9524679	2010	1	10	Sunday
11	10	1/11/2010	10:00:00 AM	ΚΛΟΠΗ ΑΥΤ/ΤΩΝ	ΠΛΗΜ/ΜΑ	ΜΠΟΥΚΟΥΒΑΛΑ 1, Βόλος	39.3529455	22.9231415	2010	1	11	Monday
12	11	1/11/2010	1:00:00 PM	ΚΛΟΠΗ ΑΥΤ/ΤΩΝ	ΠΛΗΜ/ΜΑ	ΑΝΑΓΝΩΣΤΟΠΟΥΛΟΥ 30, Βόλος	39.3650311	22.9236221	2010	1	11	Monday
13	12	1/11/2010	11:00:00 AM	ΚΛΟΠΗ ΑΥΤ/ΤΩΝ	ΠΛΗΜ/ΜΑ	ΣΠΥΡΙΔΗ, Βόλος	39.3653133	22.9522622	2010	1	11	Monday
14	13	1/11/2010	8:00:00 AM	ΚΛΟΠΗ ΔΙΚΥΚΛΩΝ	ΠΛΗΜ/ΜΑ	ΔΗΜΟΥ ΠΙΑΝΝΗ 48, Βόλος	39.3709993	22.952842	2010	1	11	Monday
15	14	1/11/2010	8:00:00 AM	ΚΛΟΠΗ ΔΙΚΥΚΛΩΝ	ΠΛΗΜ/ΜΑ	ΑΝΔΡΩΤΕΩΣ 61, Βόλος	39.3693638	22.9459098	2010	1	11	Monday
16	15	1/12/2010	2:30:00 AM	ΚΛΟΠΗ	ΠΛΗΜ/ΜΑ	ΧΑΤΖΗΜΙΧΑΛΗ 3Γ, Βόλος	39.3637424	22.9263855	2010	1	12	Tuesday
17	16	1/13/2010	10:00:00 AM	ΚΛΟΠΗ ΑΥΤ/ΤΩΝ	ΠΛΗΜ/ΜΑ	ΙΩΑΚΟΥ 250, Βόλος	39.3745925	22.9583204	2010	1	13	Wednesday
18	17	1/13/2010	1:00:00 PM	ΚΛΟΠΗ ΑΥΤ/ΤΩΝ	ΠΛΗΜ/ΜΑ	ΑΛΜΥΡΟΥ, Βόλος	39.3615923	22.9354121	2010	1	13	Wednesday
19	18	1/13/2010	9:15:00 PM	ΚΛΟΠΗ ΔΙΚΥΚΛΩΝ	ΠΛΗΜ/ΜΑ	ΦΙΛΙΠΠΟΥ ΙΩΑΝΝΗ, Βόλος	39.356904	22.9598133	2010	1	13	Wednesday

Figure 3.2.3-2: Format of data after the addition of coordinates and extraction of time data values

### 3.2.4 Mapping and Clustering of data

The coordinate values, Latitude and Longitude columns, are used in order to map the incidents and have a general view of the location of data. Given the location of data, meaningful conclusions can be extracted regarding the distribution of incidents in the city of Volos for the past 8 years. In addition to that and by applying also filters on the rest of the features of the data additional and more particular conclusions can be drawn. For instance, by selecting to apply a

filter on the Year column and on Offense column, mapping of part of the data regarding specific year and type of offence can be produced and particular inferences can be drawn. For this purpose, a text file containing the latitude and longitude of the incidents is uploaded on the web interface application of Professor Athanasios Lois and the result is a map having the incidents points marked on it.

Furthermore, the visualization of the data is useful for clustering them based on their attributes. In this study clustering of the data is done based on the latitude and longitude values of the incident data. The incidents are clustered by using again the web interface application of Professor Athanasios Lois. In Figure 3.2.4-1, the incidents are presented into clusters. The color of the cluster reflects the amount of incident points which belong to the cluster. If a cluster consists of less than 10 incident points, the color of the cluster is blue. If a cluster consists of more than 10 incident points, the color of the cluster is yellow. And, if a point does not belong to a cluster, it is presented with a red pin on the map as an individual point.



Figure 3.2.4-1: Visualization of the individual incident points and the clusters for all years

The clustering of incidents data is significant, as based on the outcome the city of Volos is divided into sectors. Given the sectors of the city, further calculations can be executed, such as the estimation of the median of each sector, which is described in more detail in paragraph 3.7.

### **3.3 Statistical experiments and comments on the results**

After the data transformation and processing as described in section 3.2 we conducted statistical experiments with the use of spreadsheets of Microsoft Excel. In particular we applied filters to produce searches based on the year, the month, the weekday and the hour that offenses took place, as well as and the kind of offense. Also searches related to specific days and important time period of the year were produced. Specifically were implemented searches related to the first and last week of the year and to Easter week, and also to specific days, important for the country. These weeks and specific days of the year were chosen to be taken into account in the statistical experiments, as they are the most important public holidays for the country. During these public holidays the shops and most businesses are closed and it's common for people to travel and leave their houses without security the most of the time. This situation favors the commission of offenses. All the statistical experiments carried out are listed in the following table 3.3.1.

After the statistical experiments have been carried out, we have developed comparative charts, so as to compare the searches results and to come to conclusions related to the number of offenses that occurred every year, every month, every weekday, every hour and every important time period and day during the year. Also results were obtained with regard to the type of offense so as to indicate the most common offense. Subsequently the above charts are presented and commented.

Table 3.3.1: Statistical experiments related to theft offenses in Volos for the time period  
2010-2017

Statistical experiments related to theft offenses in Volos for the time period 2010-2017

- ✓ **All the offenses for every year**
- ✓ **All the offenses for every month**
- ✓ **All the offenses for every weekday**
- ✓ **All the offenses for every hour**
- ✓ **Every type of offense for all the years**
- Important days and time - period:*
- ✓ **First week of the year**
- ✓ **Easter week of the year**
- ✓ **Last week of the year**
- ✓ **01/01 for all the years**    *New Year's Day*
- ✓ **15/08 for all the years**    *Celebration of the Virgin Mary*
- ✓ **25/12 for all the years**    *Christmas*
- ✓ **31/12 for all the years**    *New Year's Eve*

As mentioned before we studied data related to theft offenses in the city of Volos for the time period 2010 - 2017. In fact *three categories of offenses* were examined:

- A. Burglaries** ( a general class of thefts that includes: Thefts, Distinguished Cases of Thefts, Theft - Break into)
- B. Car Robberies**
- C. Motorcycle Robberies**

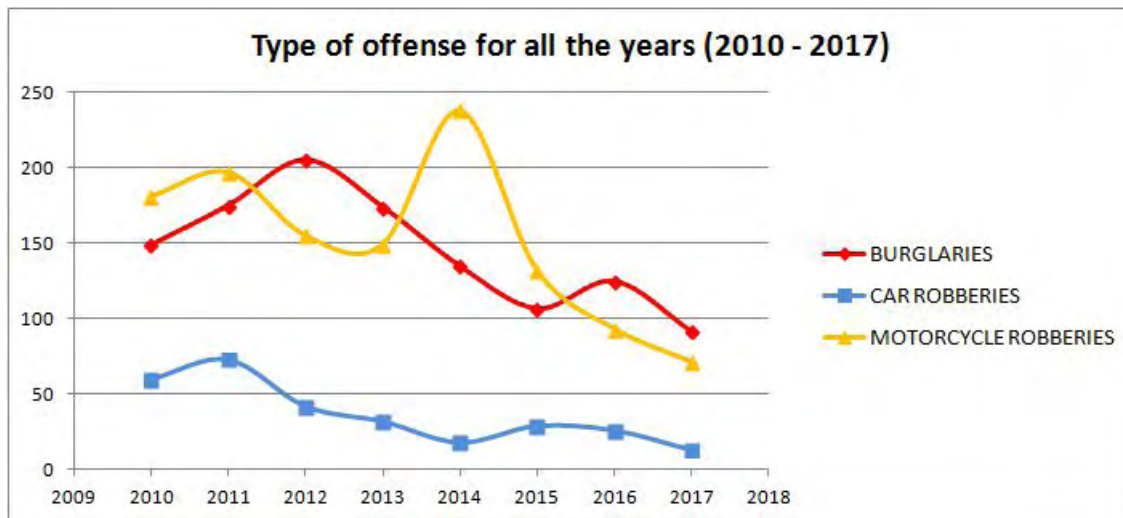


Figure 3.3.1: Type of offense for all the years (2010-2017)

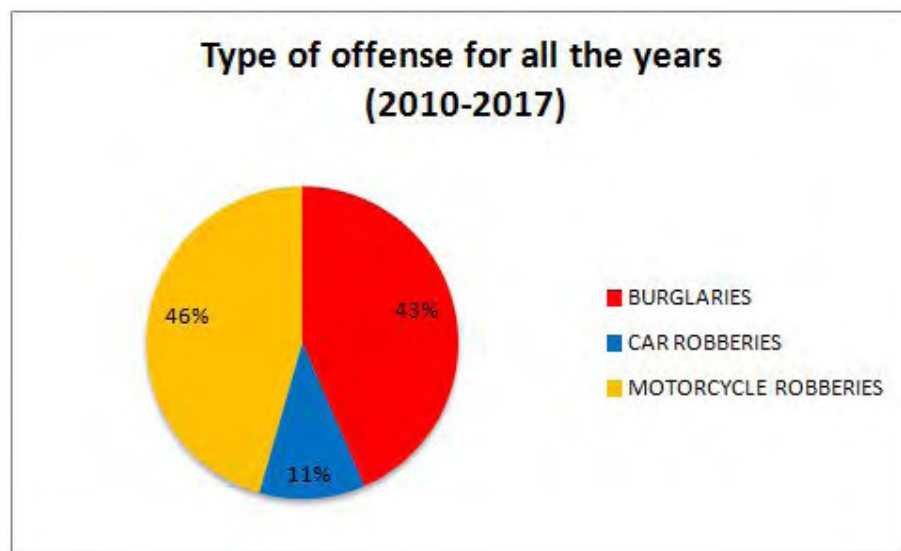


Figure 3.3.2: Type of offense for all the years, translated into percentages

The majority of offenses during the period 2010-2017 refer to Burglaries and Motorcycle Robberies, with the later to be slightly more. On the other hand Car Robberies refer to 11% of the total offenses.



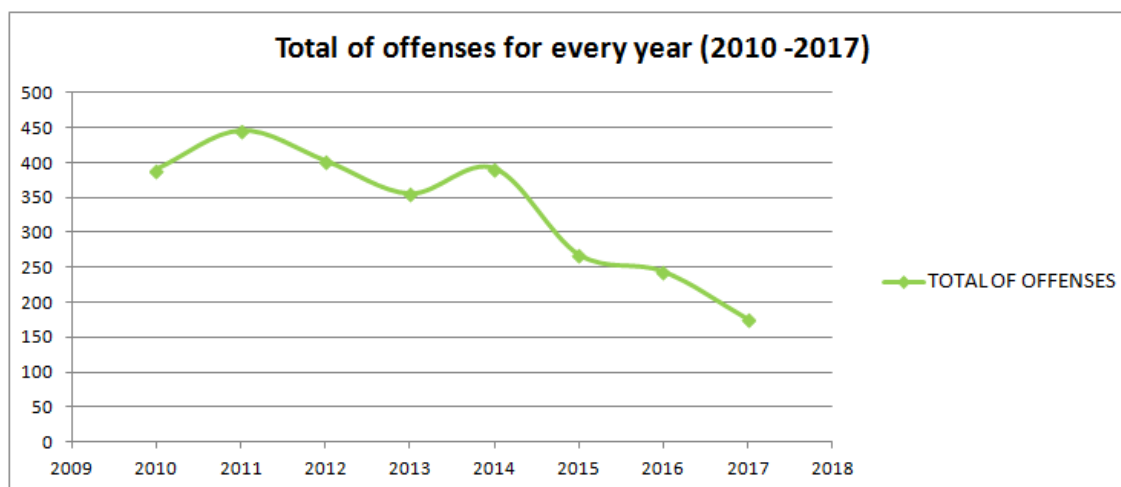


Figure 3.3.3: Total of offenses for every year (2010-2017)

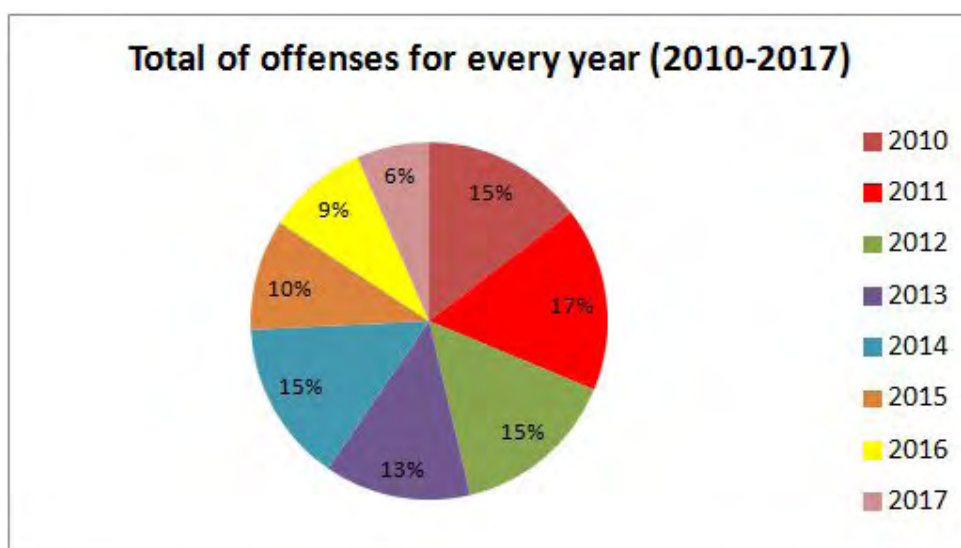


Figure 3.3.3: Total of offenses for every year (2010-2017)

Table 3.3.2: Increase/Reduction of offenses compared to the previous year

YEAR	INCREASE/REDUCTION	RATE
2011	INCREASE	14.1%
2012	REDUCTION	9.66%
2013	REDUCTION	11.69%
2014	INCREASE	10.14%
2015	REDUCTION	31.46%
2016	REDUCTION	8.95%
2017	REDUCTION	27.87%



The rates of reduction / increase of offenses in relation to the previous year shown in table 3.3.2 were derived from calculations based on the table set out in Appendix section 7.1.

For the time period studied, a general reduction of offenses was observed with only two years to show an increase over the previous year. Specifically in 2011 was an increase in offenses by 14.1% compared to 2010 and in 2014 was an increase in offenses by 10.14% compared to 2013. From 2015 onwards, we have significant reduction of offenses with 2017 showing a 54.87% decrease over 2010. The highest reduction rate between the years has been observed in 2015. In fact, in 2015, there were 31.46% more offenses than in 2014.

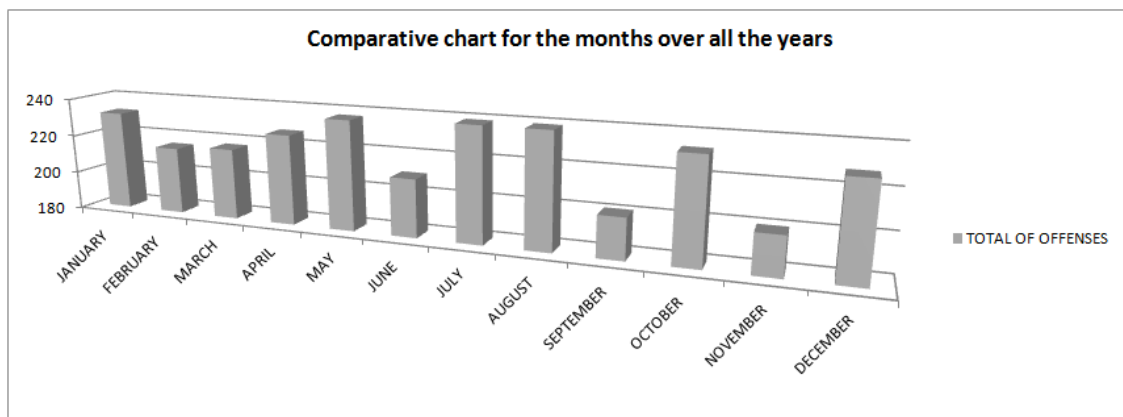


Figure 3.3.4: Comparative chart for the months over all the years

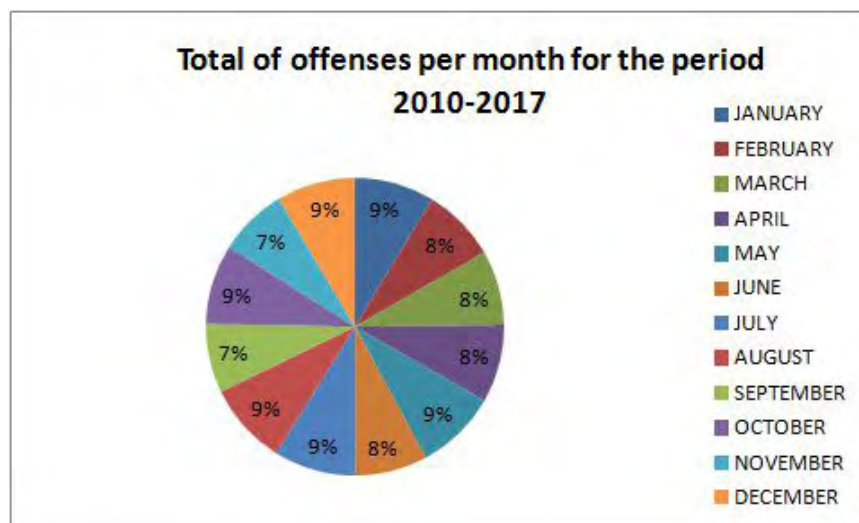


Figure 3.3.5: Total of offenses per month over the period 2010-2017

According to figure 3.3.4 and figure 3.3.5 the total of offenses does not show any particular difference between the months over the period 2010-2017. In fact the totals show a slight divergence of 1% and 2%. The months of September and November correspond to the lowest total of offenses, while the months in which the most offenses were committed are the months July, August, May, October, January and December. From the above we infer that offenses are prone to occur during the summer months, where the schools are closed and most of the people leave for summer vacations, and the months that include important days such as Christmas. Statistical experiments related to those important days and specific time periods of the year will be presented later in this section.

The charts showing information about the type of offense for every month over the period 2010-2017 are set out in the Appendix (section 7.2). Regarding these charts, the overall conclusion is that the most offenses related to Burglaries and Car Robberies, while the Motorcycle Robberies follow with a great deal of difference.

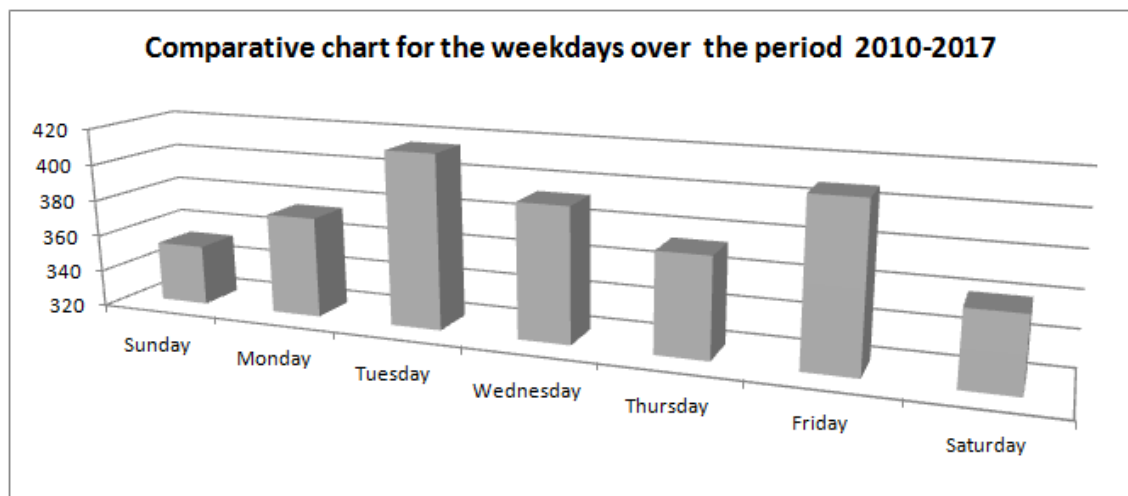


Figure 3.3.6: Comparative chart for the weekdays over the period 2010-2017

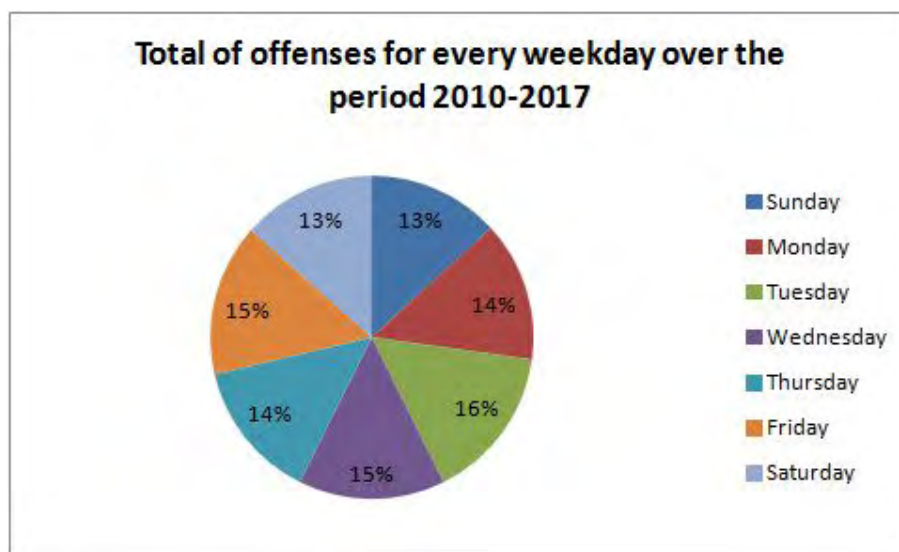


Figure 3.3.7: Total of offenses for every weekday over the period 2010-2017

As seen in the figures 3.3.6 & 3.3.7 the day with the display of most offenses is Tuesday. Then follow without much difference the day of Friday sharing the same rate with Wednesday, the day of Monday sharing the same rate with Thursday and last the day of Saturday sharing the same rate with Sunday. The percentage difference between the days is in the range of 1% to 2%.

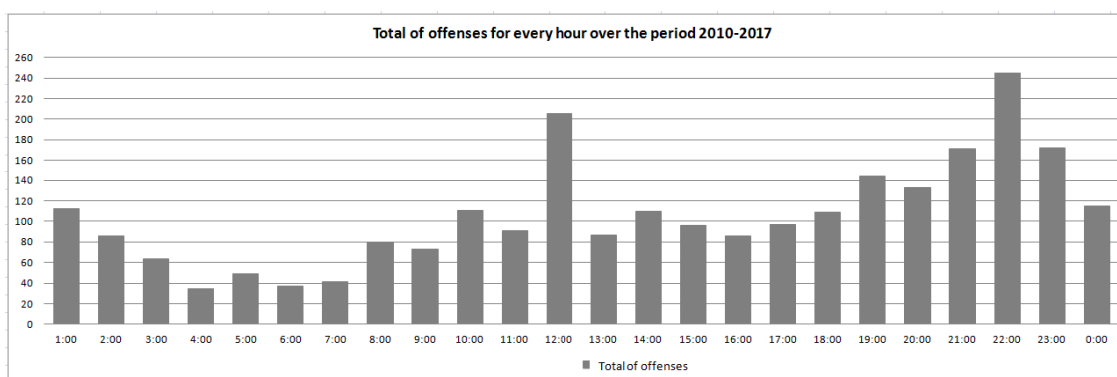


Figure 3.3.8: Total of offenses for every hour over the period 2010-2017

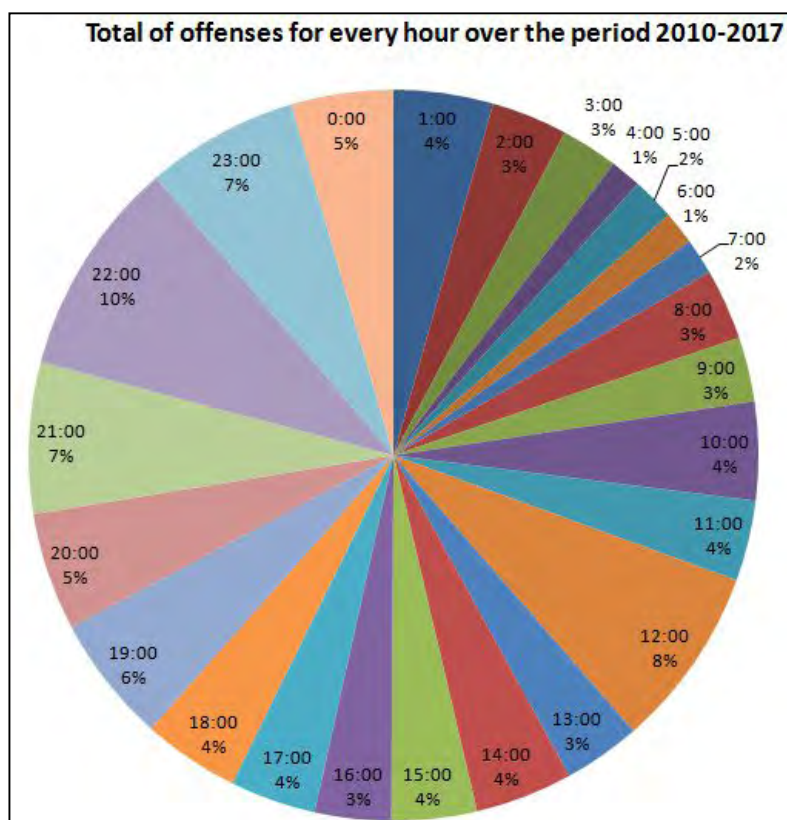


Figure 3.3.9: Total of offenses for every hour over the period 2010-2017, translated into percentage

Table 3.3.3: Order of hours based on the frequency of offenses

Group of hours	Rate occurrence of offenses
22:00	10%
12:00	8%
21:00 & 23:00	7%
19:00	6%
00:00 & 20:00	5%
01:00, 09:00, 11:00, 14:00, 15:00, 17:00, 18:00	4%
02:00, 03:00, 08:00, 09:00, 13:00, 16:00	3%
05:00, 07:00	2%
04:00	1%

Taken into consideration the figures 3.3.8 & 3.3.9 it appears that the most offenses occurred mainly in the afternoon and evening hours. In particular, 22:00 hours seems to have the highest percentage of offenses occurrence, while 04:00 seems to have the lowest one. Also, the 12:00 turns out to be in the highest risk after the 22:00. The above table 3.3.3 shows the hours in relation to the occurrence of offenses expressed in groups.

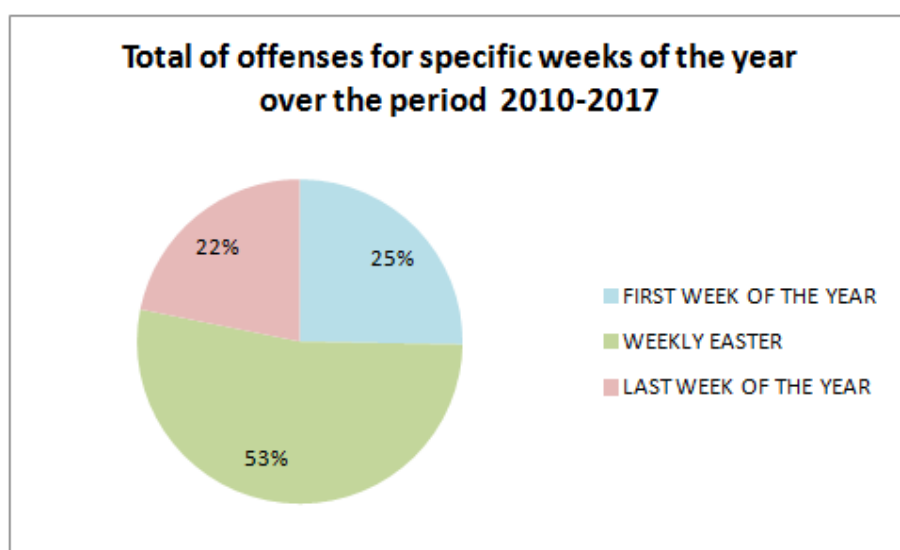


Figure 3.3.10: Total of offenses for specific weeks of the year over the period 2010-2017

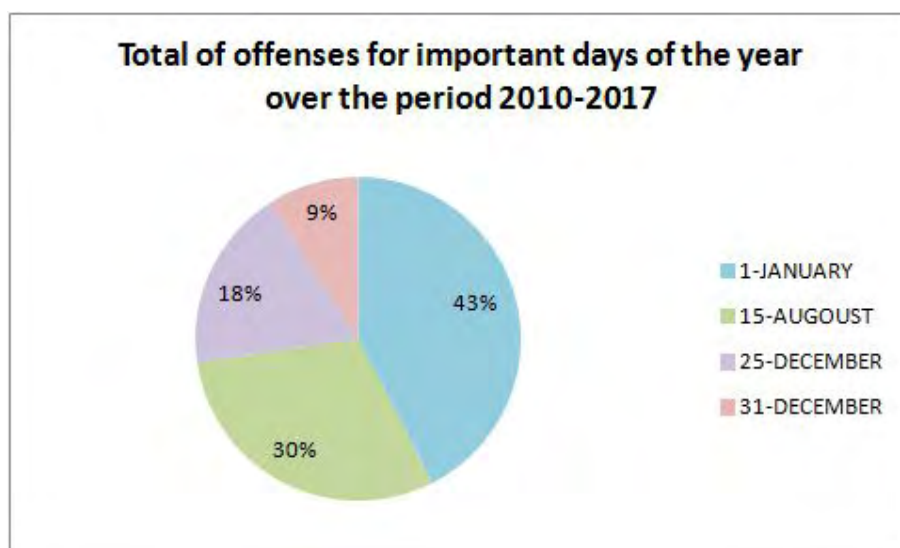
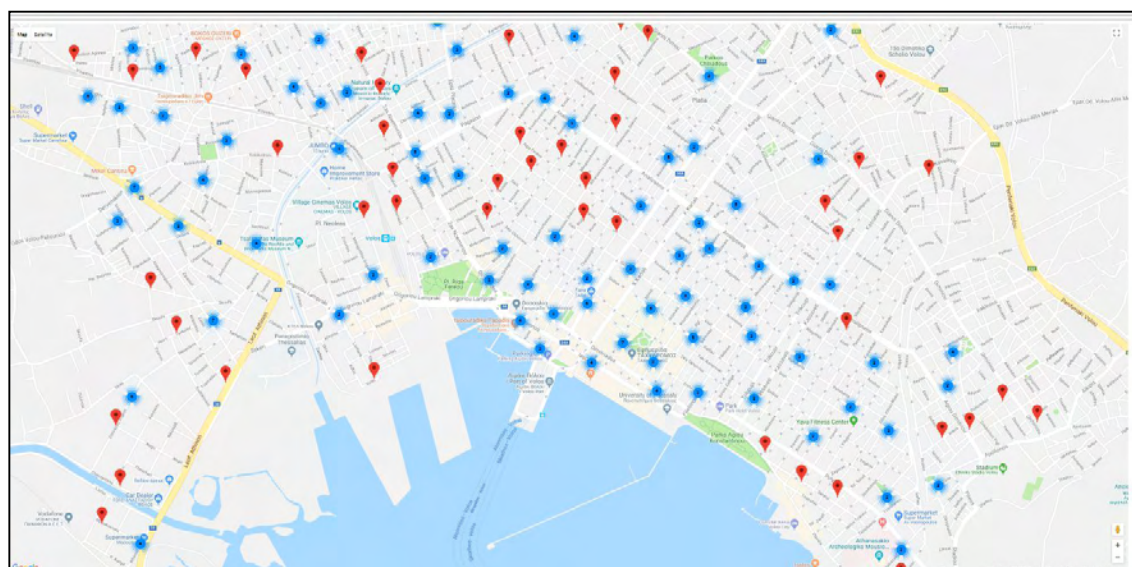


Figure 3.3.11: Total of offenses for specific weeks of the year over the period 2010-2017

As for the specific weeks of the year, it appears from figure 3.3.10 that the most offenses occurred during Easter Week. Easter is usually celebrated in Greece in May or April, according to the Orthodox calendar. Furthermore by examining and comparing the important days of the year, figure 3.3.11, it is concluded that the majority of offenses took place in the 1<sup>st</sup> of January, New Year's Day, and then in the 15<sup>th</sup> of August, Celebration of Virgin Mary (a great religious feast for the country).

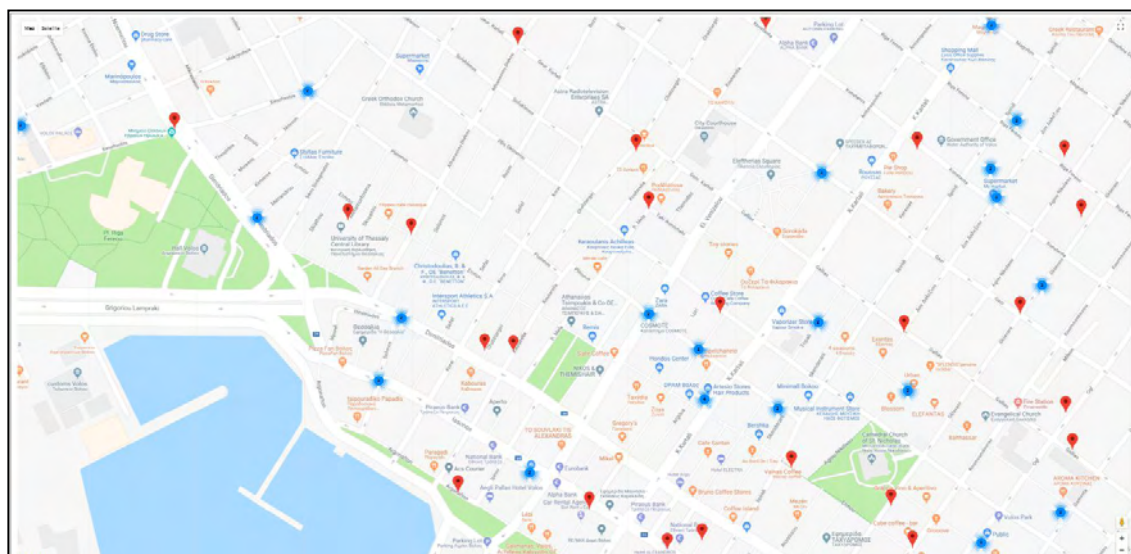
In regard to the location of incidents, the coordinate values are used in order to map the incidents and have a general view of the location of data. In particular, text files are produced with the results of the above statistical experiments, expressed in terms of Latitude and Longitude and then these text files containing the latitude and longitude of the incidents, are uploaded on the web interface application of Professor Athanasios Lois and the result is a map having the incidents points marked on it. It turned out that most of the incidents occurred nearby the main streets while they spread throughout the city. The figures below show all the incidents that took place in 2011, the year with the biggest number of offenses. By zoom in, it appears that most of the incidents occurred nearby main streets.

A)





B)



C)

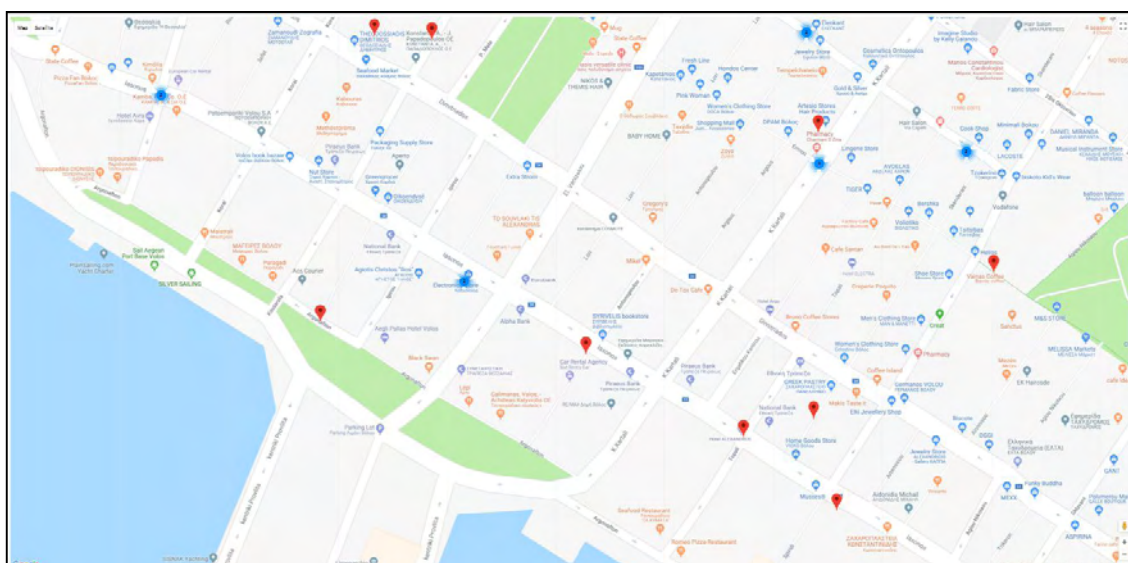


Figure 3.3.12 (A,B,C): Visualization of the individual incident points and the clusters for 2011

### **3.4 Prediction of future incidents**

Being proactive about predicting crime incidents through the use of data analytics is the best way to be prepared and act in a more effective way. Based on forecasting outcomes, important decisions can be made and appropriate solutions can be designed in order to prevent crime incidents to occur, such as the increment of policing in particular time periods and city zones. For this purpose, the Machine Learning Toolkit of Splunk software platform is used.

#### **3.4.1 Splunk Machine Learning Toolkit**

Splunk is a software platform to search, analyze and visualize the data gathered from databases, websites, applications, sensors and various devices. The most significant operation of Splunk is the real time processing of data. The data are stored and processed based on their time attribute, called timestamp. There are various techniques to import data into Splunk, such as uploading simple csv files or forwarding data from remote systems and devices by using Splunk forwarders.

Moreover, Splunk platform provides additional solutions based on the add-on toolkits and applications. One of these add-ons is the Splunk Machine Learning Toolkit which assists in applying machine learning techniques and methods against the data. Methods to analyze data including algorithms such as regression, anomaly and outlier detection are provided with this toolkit. These algorithms are essential for understanding, modeling and detecting trends in the data not easily identifiable by observation. In Figure 3.4.1-1, the search head, which is the main component for analyzing the incoming data in Splunk Machine Learning Toolkit of Splunk platform, is presented.



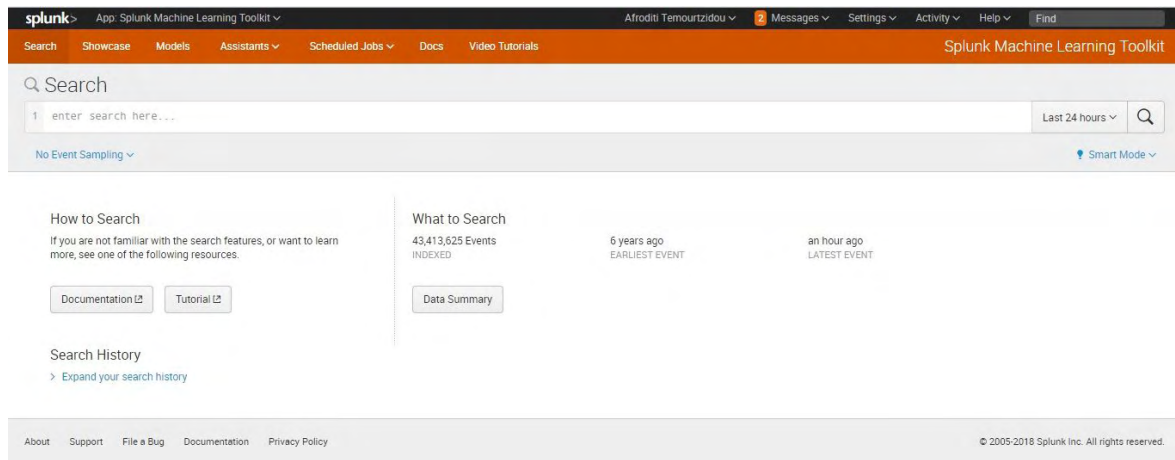


Figure 3.4.1-1: The search head of Splunk Machine Learning Toolkit

As it is stated above, prediction of crime incidents is essential in order to prevent and decrease future occurrences. Therefore, Splunk Machine Learning Toolkit is used for prediction purposes in this dissertation.

### 3.4.2 Preparation of data inputs

In order to apply the machine learning algorithms of Splunk Machine Learning Toolkit and forecast future crime incidents, it is necessary to prepare the data inputs first. The data inputs should be precisely structured. In order to achieve this, the data inputs should be created according to the available data and the requirements of the algorithm applied.

In this study, the data inputs are created based on 3 data features. These features are: the timestamp, the coordinates and the type of offense. Timestamp is the most efficient feature, because Splunk indexes and processes the data based on the time that they occurred. Thus, it is essential to take into consideration the time attribute of the incidents. In this study and concerning the time attribute of the available incident data, the data inputs will be produced based on the month and the year, i.e. Jan-2010, Feb-2010,..., Dec-2017. Regarding the coordinates, the city of Volos is divided into 3 big sectors indicating the west, the center and the east parts of the city. This division is applied based on how the city is built and because it is more practical to analyze the data and apply prediction algorithms in the sectors instead of the whole city. In Figure 3.4.2-1,

the 3 sectors of the city of Volos are illustrated. An additional feature which can be taken into consideration in the creation of the data inputs is the type of offense. Future crime incidents can be predicted as a whole. However, it is more precise to predict them based on the type of the offense. There are 3 different types of offense studied in this thesis, burglaries, car robberies and motorcycle robberies. Thus, data inputs based on the different types of offence are created.

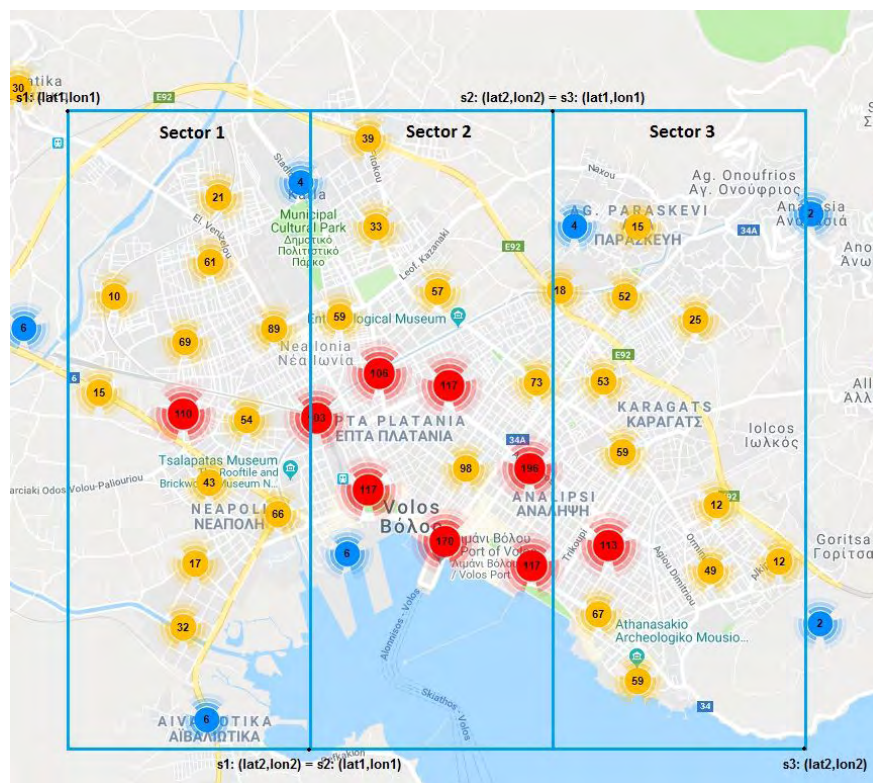


Figure 3.4.2-1: The 3 sectors of the city of Volos

For the purpose of this study 12 data inputs are created and tested on the dispensable Splunk's prediction algorithms. These data inputs are:

1. All incident data for sector 1
2. All incident data for sector 2
3. All incident data for sector 3
4. Burglaries for sector 1
5. Burglaries for sector 2
6. Burglaries for sector 3
7. Car robberies for sector 1
8. Car robberies for sector 2
9. Car robberies for sector 3
10. Motorcycle robberies for sector 1
11. Motorcycle robberies for sector 2
12. Motorcycle robberies for sector 3

### **3.4.3 Prediction algorithms**

Splunk Machine Learning Toolkit provides univariate and bivariate algorithms based on Kalman filter. Kalman filtering, also known as linear quadratic estimation (LQE), is an algorithm that uses a series of measurements observed over time, containing statistical noise and other inaccuracies, and produces estimates of unknown variables that tend to be more accurate than those based on a single measurement alone, by estimating a joint probability distribution over the variables for each timeframe (Zarchan & Musoff, 2000). The univariate algorithms of Splunk Machine Learning Toolkit are used for the purpose of this study. These are:

- Local level (LL): A model with no trends and no seasonality. Requires a minimum of 2 data points. The LL algorithm is the simplest algorithm and computes the

levels of the time series. For instance, each new state equals the previous state, plus the Gaussian noise.

- Local level trend (LLT): A model with trend, but no seasonality. Requires a minimum of 3 data points.
- Seasonal local level (LLP): A model with seasonality. The number of data points must be at least twice the number of periods, using the period attribute. The LLP algorithm takes into account the cyclical regularity of the data, if it exists. If the number of periods is known, the period argument should be specified. If the period is not set, the algorithm tries to calculate it. LLP returns an error message if the data is not periodic.
- Combination of LLT and LLP models (LLP5): If the time series is periodic, LLP5 computes two predictions, one using LLT and the other using LLP. The algorithm then takes a weighted average of the two values and outputs that as the prediction. The confidence interval is also based on a weighted average of the variances of LLT and LLP.

In subchapter 3.4.4 the application of algorithms by using the custom data inputs is implemented and the prediction results are presented thoroughly.

### **3.4.4 Implementation of prediction algorithms**

In this paragraph the results from the implementation of univariate algorithms are presented in detail. The 5 available algorithms are applied and tested for different input parameters for every data input. These implementations will produce various results regarding the number of future crime incidents that are going to be assessed based on R-squared and Root Mean Square Error (RMSE).

R-squared is a statistical measure of how close the data are to the fitted regression line. It is also known as the coefficient of determination, or the coefficient of multiple determinations for

multiple regressions. It is the percentage of the response variable variation that is explained by a linear model and it is calculated by the formula:

$$R - squared = \frac{Explained\ variation}{Total\ variation} \quad (3.4.4-1)$$

R-squared always ranges between 0 and 100%. The 0 percentage indicates that the model explains none of the variability of the response data around its mean. On the contrary, the 100 percentage indicates that the model explains all the variability of the response data around its mean. In general, the higher the R-squared, the better the model fits the data.

Root Mean Square Error (RMSE) is the standard deviation of the residuals (prediction errors). Residuals are a measure of how far from the regression line data points are. RMSE is a measure of how spread out these residuals are. In other words, RMSE indicates how concentrated the data is around the line of best fit. In general, lower values of RMSE indicate better fit. RMSE is commonly used in climatology, forecasting, and regression analysis to verify experimental results.

The formula of RMSE is:

$$RMSE = \sqrt{\overline{(f - o)^2}} \quad (3.4.4-2)$$

Where:

f = forecasts (expected values or unknown results),

o = observed values (known results).

The bar above the squared differences is the mean (similar to  $\bar{x}$ ). (Barnston, 1992)

In the following sub-paragraphs the prediction values for the next year (2018), the R-squared and RMSE are calculated for all the data input and for various combinations of future timespan and holdback values. Future timespan value specifies how many future predictions the predict command will compute and it is a non-negative number. Holdback value specifies the number of data points from the end that are not to be used by the prediction algorithm. It is used in

conjunction with the future timespan (future\_timespan) argument. For instance, 'holdback=10 future\_timespan=10' computes the predicted values for the last 10 values in the data set. By setting the holdback and future timespan values to be equal, the assessment of how accurate the predictions values are can be achieved by checking whether the actual data point values fall into the predicted confidence intervals.

Furthermore, it is essential to set the confidence intervals. The lower and upper confidence interval parameters default to lower95 and upper95. These values specify a confidence interval where 95% of the predictions are expected fall. It is typical for some of the predictions to fall outside the confidence interval. The confidence interval does not cover 100% of the predictions. The confidence interval is about a probabilistic expectation and results do not match the expectation exactly.

The assessment of prediction outcomes is done by R-squared and RMSE. From all algorithm applications, only the more suitable is presented analytically in this chapter. All prediction results and diagrams are presented in more detail in the Appendix.

#### 3.4.4.1 All incident data for sector 1

##### ➤ LL algorithm

Test id	Future timespan	Holdback	R-squared	RMSE
LL1	24	12	0.4152	2.97
LL2	36	24	0.4040	3.00
LL3	48	36	0.1149	3.66
LL4	60	48	0.0250	3.84

Table 3.4.4.1-1: Results of LL algorithm for input data “All incidents data sector 1” for various future timespan and holdback values

##### ➤ LLT algorithm

Test id	Future timespan	Holdback	R-squared	RMSE
LLT1	24	12	0.5540	2.60
LLT2	36	24	0.5077	2.73
LLT3	48	36	0.4586	2.86
LLT4	60	48	0.4689	2.83

Table 3.4.4.1-2: Results of LLT algorithm for input data “All incidents data sector 1” for various future timespan and holdback values

➤ LLP

Test id	Future timespan	Holdback	R-squared	RMSE
LLP1	24	12	0.6104	2.43
LLP2	36	24	0.5533	2.60
LLP3	48	36	0.2620	3.34

Table 3.4.4.1-3: Results of LLP algorithm for input data “All incidents data sector 1” for various future timespan and holdback values

➤ LLP5

Test id	Future timespan	Holdback	R-squared	RMSE
LLP5_1	24	12	0.6559	2.28
LLP5_2	36	24	0.6265	2.38
LLP5_3	48	36	<b>0.6892</b>	<b>2.17</b>
LLP5_4	60	48	0.5558	2.59

Table 3.4.4.1-3: Results of LLP algorithm for input data “All incidents data sector 1” for various future timespan and holdback values

From all replications of algorithm with test id LLP5\_3, which is the application of LLP5 algorithm with future timespan 48 periods and holdback 36 periods, produces the more reliable

outcome regarding the values of R-squared and RMSE. In Figure 3.4.4.1-1 the actual and forecast charts are presented together in one graph for confidence interval 95%. Also, the prediction results are presented in Table 3.4.4.1-5.



Figure 3.4.4.1-1: Forecast chart for all incidents data in sector 1 with confidence interval 95%

Time period	Lower95(prediction)	Prediction	Upper95(prediction)
2018-01-01	-10.06653568 → 0	3.371515103 → 3	16.80956588 → 17
2018-02-01	-10.93145373 → 0	3.32727736 → 3	17.58600845 → 18
2018-03-01	-14.51982473 → 0	-0.187031137 → 0	14.14576245 → 14
2018-04-01	-14.0688215 → 0	0.337195559 → 0	14.74321262 → 15
2018-05-01	-14.25274793 → 0	0.225670566 → 0	14.70408907 → 15
2018-06-01	-14.63054764 → 0	-0.080533291 → 0	14.46948106 → 14
2018-07-01	-14.21030066 → 0	0.410519891 → 0	15.03134045 → 15
2018-08-01	-15.0390757 → 0	-0.348223129 → 0	14.34262944 → 14
2018-09-01	-15.13637778 → 0	-0.376252405 → 0	14.38387297 → 14
2018-10-01	-15.3973642 → 0	-0.568710676 → 0	14.25994285 → 14
2018-11-01	-15.34525598 → 0	-0.44880486 → 0	14.44764626 → 14
2018-12-01	-15.66661568 → 0	-0.703083797 → 0	14.26044808 → 14

Table 3.4.4.1-5a: Forecasted values regarding all crime incidents for 2018 with confidence interval 95%

### 3.4.4.2 All incident data for sector 2

➤ LL algorithm



Test id	Future timespan	Holdback	R-squared	RMSE
LL1	24	12	0.4651	3.92
LL2	36	24	0.4641	3.92
LL3	48	36	0.2055	4.78
LL4	60	48	0.1603	4.91

Table 3.4.4.2-1: Results of LL algorithm for input data “All incidents data sector 2” for various future timespan and holdback values

➤ LLT algorithm

Test id	Future timespan	Holdback	R-squared	RMSE
LLT1	24	12	0.4476	3.98
LLT2	36	24	0.4521	3.97
LLT3	48	36	0.3807	4.22
LLT4	60	48	0.4320	4.04

Table 3.4.4.2-2: Results of LLT algorithm for input data “All incidents data sector 2” for various future timespan and holdback values

➤ LLP

Test id	Future timespan	Holdback	R-squared	RMSE
LLP1	24	12	0.4249	4.07
LLP2	36	24	0.3746	4.24
LLP3	48	36	0.0472	5.23
LLP4	60	48	0.1461	4.95

Table 3.4.4.2-3: Results of LLP algorithm for input data “All incidents data sector 2” for various future timespan and holdback values

➤ LLP5

Test id	Future timespan	Holdback	R-squared	RMSE
LLP5_1	24	12	0.4814	3.86
LLP5_2	36	24	<b>0.5162</b>	<b>3.73</b>
LLP5_3	48	36	0.4771	3.88
LLP5_4	60	48	0.5118	3.75

Table 3.4.4.2-4: Results of LLP5 algorithm for input data “All incidents data sector 2” for various future timespan and holdback values

From all replications of algorithm with test id LLP5\_2, which is the application of LLP5 algorithm with future timespan 36 periods and holdback 24 periods, produces the more reliable outcome regarding the values of R-squared and RMSE. In Figure 3.4.4.2-1 the actual and forecast charts are presented together in one graph for confidence interval 95%. Also, the prediction results are presented in Table 3.4.4.2-5.

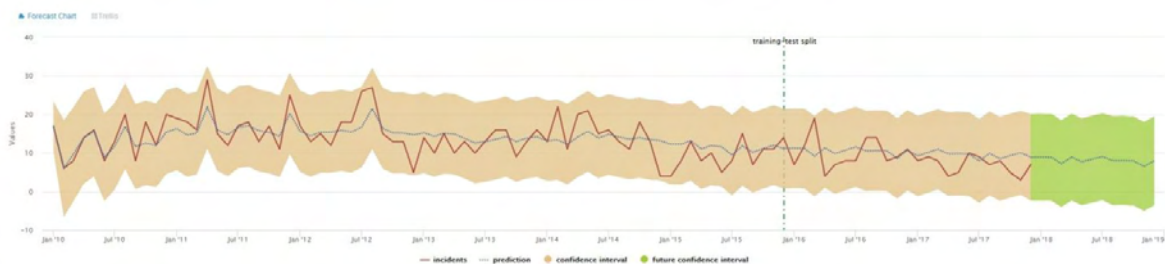


Figure 3.4.4.2-1: Forecast chart for all incidents data in sector 2 with confidence interval 95%

Time period	Lower95(prediction)	Prediction	Upper95(prediction)
2018-01-01	-2.252380539 → 0	8.955491844 → 9	20.16336423 → 20
2018-02-01	-2.344798553 → 0	8.865611943 → 9	20.07602244 → 20

2018-03-01	-3.93858941 → 0	7.27435758 → 7	18.48730457 → 18
2018-04-01	-2.280131279 → 0	8.935350586 → 9	20.15083245 → 20
2018-05-01	-3.48558817 → 0	7.732426955 → 8	18.95044208 → 19
2018-06-01	-2.796254465 → 0	8.424292306 → 8	19.64483908 → 20
2018-07-01	-2.186506539 → 0	9.036570265 → 9	20.25964707 → 20
2018-08-01	-3.469593165 → 0	8.024969028 → 8	19.51953122 → 20
2018-09-01	-3.44504112 → 0	8.052233938 → 8	19.549509 → 20
2018-10-01	-3.540968953 → 0	7.959017334 → 8	19.45900362 → 19
2018-11-01	-4.983812144 → 0	6.518883737 → 7	18.02157962 → 18
2018-12-01	-3.508981389 → 0	7.996422453 → 8	19.5018263 → 20

Table 3.4.4.2-5: Forecasted values regarding all crime incidents for 2018 for sector 2 with confidence interval 95%

### 3.4.4.3 All incident data for sector 3

#### ➤ LL algorithm

Test id	Future timespan	Holdback	R-squared	RMSE
LL1	24	12	0.0185	3.08

Table 3.4.4.3-1: Results of LL algorithm for input data “All incidents data sector 3” for various future timespan and holdback values

#### ➤ LLT algorithm

Test id	Future timespan	Holdback	R-squared	RMSE
LLT1	24	12	0.2133	2.76

Table 3.4.4.3-2: Results of LLT algorithm for input data “All incidents data sector 3” for various future timespan and holdback values

➤ LLP

Test id	Future timespan	Holdback	R-squared	RMSE
LLP1	24	12	<b>0.5565</b>	<b>2.07</b>

Table 3.4.4.3-3: Results of LLP algorithm for input data “All incidents data sector 3” for various future timespan and holdback values

➤ LLP5

Test id	Future timespan	Holdback	R-squared	RMSE
LLP5_1	24	12	0.4414	2.33

Table 3.4.4.3-4: Results of LLP5 algorithm for input data “All incidents data sector 3” for various future timespan and holdback values

Only the combination of future timespan equals to 24 and holdback equals to 12 returns reasonable values for R-squared and RMSE. Thus, no other tests are performed for the data input of “All incidents data sector 3”. From all replications of algorithm with test id LLP1, which is the application of LLP algorithm with future timespan 24 periods and holdback 12 periods, produces the more reliable outcome regarding the values of R-squared and RMSE. In Figure 3.4.4.3-1 the actual and forecast charts are presented together in one graph for confidence interval 95%. Also, the prediction results are presented in Table 3.4.4.3-5.

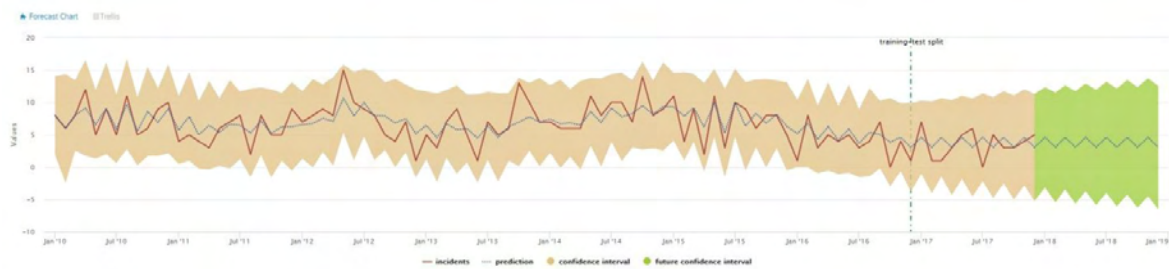


Figure 3.4.4.3-1: Forecast chart for all incidents data in sector 3 with confidence interval 95%

Time period	Lower95(prediction)	Prediction	Upper95(prediction)
2018-01-01	-3.090321038 → 0	4.620862341 → 5	12.33204572 → 12
2018-02-01	-5.399056079 → 0	3.063107982 → 3	11.52527204 → 12
2018-03-01	-3.384814929 → 0	4.620862341 → 5	12.62653961 → 13
2018-04-01	-5.611515844 → 0	3.063107982 → 3	11.73773181 → 12
2018-05-01	-3.668853457 → 0	4.620862341 → 5	12.91057814 → 13
2018-06-01	-5.818894969 → 0	3.063107982 → 3	11.94511093 → 12
2018-07-01	-3.943476949 → 0	4.620862341 → 5	13.18520163 → 13
2018-08-01	-6.021541397 → 0	3.063107982 → 3	12.14775736 → 12
2018-09-01	-4.209563862 → 0	4.620862341 → 5	13.45128854 → 13
2018-10-01	-6.219765079 → 0	3.063107982 → 3	12.34598104 → 12
2018-11-01	-4.467863993 → 0	4.620862341 → 5	13.70958868 → 14
2018-12-01	-6.413843541 → 0	3.063107982 → 3	12.5400595 → 13

Table 3.4.4.3-5: Forecasted values regarding all crime incidents for 2018 for sector 3 with confidence interval 95%

#### 3.4.4.4 Burglaries for sector 1

➤ LL algorithm

Test id	Future timespan	Holdback	R-squared	RMSE
LL1	24	12	0.1261	1.88
LL2	36	24	0.1137	1.89
LL3	48	36	0.0528	1.96

Table 3.4.4.4-1: Results of LL algorithm for input data “Burglaries for sector 1” for various future timespan and holdback values

➤ LLT algorithm

Test id	Future timespan	Holdback	R-squared	RMSE
LLT1	24	12	0.2181	1.78
LLT2	36	24	0.3148	1.66

Table 3.4.4.4-2: Results of LLT algorithm for input data “Burglaries for sector 1” for various future timespan and holdback values

➤ LLP

Test id	Future timespan	Holdback	R-squared	RMSE
LLP1	24	12	<b>0.4468</b>	<b>1.49</b>
LLP2	36	24	0.1826	1.82

Table 3.4.4.4-3: Results of LLP algorithm for input data “Burglaries for sector 1” for various future timespan and holdback values

➤ LLP5

Test id	Future timespan	Holdback	R-squared	RMSE
LLP5_1	24	12	0.3956	1.56

LLP5_2	36	24	0.3620	1.61
LLP5_3	48	36	0.2556	1.73
LLP5_4	60	48	0.3482	1.62

Table 3.4.4.4-4: Results of LLP5 algorithm for input data “Burglaries for sector 1” for various future timespan and holdback values

From all replications of algorithm with test id LLP1, which is the application of LLP algorithm with future timespan 24 periods and holdback 12 periods, produces the more reliable outcome regarding the values of R-squared and RMSE. In Figure 3.4.4.4-1 the actual and forecast charts are presented together in one graph for confidence interval 95%. Also, the prediction results are presented in Table 3.4.4.4-5



Figure 3.4.4.4-1: Forecast chart for burglaries in sector 1 with confidence interval 95%

Time period	Lower95(prediction)	Prediction	Upper95(prediction)
2018-01-01	-0.76184441 → 0	2.587856288 → 3	5.937556986 → 6
2018-02-01	-1.180968002 → 0	2.296444748 → 2	5.773857498 → 6
2018-03-01	-3.774392783 → 0	0.804058455 → 1	5.382509693 → 5
2018-04-01	-7.094491742 → 0	1.089803423 → 1	9.274098589 → 9

2018-05-01	-0.58499321 → 0	3.114448107 → 3	6.813889424 → 7
2018-06-01	-0.761945523 → 0	2.587856288 → 3	5.937658099 → 6
2018-07-01	-1.181185706 → 0	2.296444748 → 2	5.774075203 → 6
2018-08-01	-4.045873793 → 0	0.804058455 → 1	5.653990703 → 6
2018-09-01	-7.520039965 → 0	1.089803423 → 1	9.699646812 → 10
2018-10-01	-0.58540213 → 0	3.114448107 → 3	6.814298343 → 7
2018-11-01	-0.762046633 → 0	2.587856288 → 3	5.937759209 → 6
2018-12-01	-1.181403397 → 0	2.296444748 → 2	5.774292894 → 6

Table 3.4.4.4-5: Forecasted values regarding burglaries for 2018 for sector 1 with confidence interval 95%

### 3.4.4.5 Burglaries for sector 2

#### ➤ LL algorithm

Test id	Future timespan	Holdback	R-squared	RMSE
LL1	24	12	0.4598	2.14
LL2	36	24	0.5107	2.03
LL3	48	36	0.4953	2.07

Table 3.4.4.5-1: Results of LL algorithm for input data “Burglaries for sector 2” for various future timespan and holdback values

#### ➤ LLT algorithm

Test id	Future timespan	Holdback	R-squared	RMSE
LLT1	24	12	0.5732	1.90
LLT2	36	24	0.5102	2.03



Table 3.4.4.5-2: Results of LLT algorithm for input data “Burglaries for sector 2” for various future timespan and holdback values

➤ LLP

Test id	Future timespan	Holdback	R-squared	RMSE
LLP1	24	12	0.4950	2.07
LLP2	36	24	<b>0.6108</b>	<b>1.81</b>
LLP3	48	36	0.5954	1.85

Table 3.4.4.5-3: Results of LLP algorithm for input data “Burglaries for sector 2” for various future timespan and holdback values

➤ LLP5

Test id	Future timespan	Holdback	R-squared	RMSE
LLP5_1	24	12	0.5637	1.92
LLP5_2	36	24	0.5971	1.85
LLP5_3	48	36	0.3219	2.39

Table 3.4.4.5-4: Results of LLP5 algorithm for input data “Burglaries for sector 2” for various future timespan and holdback values

From all replications of algorithm with test id LLP2, which is the application of LLP algorithm with future timespan 36 periods and holdback 24 periods, produces the more reliable outcome regarding the values of R-squared and RMSE. In Figure 3.4.4.5-1 the actual and forecast charts are presented together in one graph for confidence interval 95%. Also, the prediction results are presented in Table 3.4.4.5-5.



Figure 3.4.4.5-1: Forecast chart for burglaries in sector 2 with confidence interval 95%

Time period	Lower95(prediction)	Prediction	Upper95(prediction)
2018-01-01	-13.34313765 → 0	3.06163907 → 3	19.46641579 → 19
2018-02-01	-4.328672531 → 0	3.93246052 → 4	12.19359357 → 12
2018-03-01	-13.89044855 → 0	3.06163907 → 3	20.01372669 → 20
2018-04-01	-4.485985391 → 0	3.93246052 → 4	12.35090643 → 12
2018-05-01	-14.42063339 → 0	3.06163907 → 3	20.54391153 → 21
2018-06-01	-4.640412033 → 0	3.93246052 → 4	12.50533307 → 12
2018-07-01	-14.93520582 → 0	3.06163907 → 3	21.05848395 → 21
2018-08-01	-4.792105719 → 0	3.93246052 → 4	12.65702676 → 13
2018-09-01	-15.43546885 → 0	3.06163907 → 3	21.55874699 → 22
2018-10-01	-4.941206609 → 0	3.93246052 → 4	12.80612765 → 13
2018-11-01	-15.92255374 → 0	3.06163907 → 3	22.04583188 → 22
2018-12-01	-5.087843277 → 0	3.93246052 → 4	12.95276432 → 13

Table 3.4.4.5-5: Forecasted values regarding burglaries for 2018 for sector 2 with confidence interval 95%

### 3.4.4.6 Burglaries for sector 3

#### ➤ LL algorithm

Test id	Future timespan	Holdback	R-squared	RMSE
LL1	24	12	0.0359	1.84
LL2	36	24	0.0221	1.85

LL3	48	36	0.3404	1.52
LL4	60	48	0.1508	1.73

Table 3.4.4.6-1: Results of LL algorithm for input data “Burglaries for sector 3” for various future timespan and holdback values

➤ LLT algorithm

Test id	Future timespan	Holdback	R-squared	RMSE
LLT1	24	12	0.1553	1.72
LLT2	36	24	0.0966	1.78
LLT3	48	36	0.4471	1.39

Table 3.4.4.6-2: Results of LLT algorithm for input data “Burglaries for sector 3” for various future timespan and holdback values

➤ LLP

Test id	Future timespan	Holdback	R-squared	RMSE
LLP1	24	12	0.0687	1.81
LLP2	36	24	0.1029	1.77
LLP3	48	36	0.3072	1.56
LLP4	60	48	0.2150	1.66

Table 3.4.4.6-3: Results of LLP algorithm for input data “Burglaries for sector 3” for various future timespan and holdback values

➤ LLP5

Test id	Future timespan	Holdback	R-squared	RMSE
LLP5_1	24	12	0.1329	1.74
LLP5_2	36	24	0.1266	1.75

LLP5_3	48	36	<b>0.4795</b>	<b>1.35</b>
LLP5_4	60	48	0.2743	1.60

Table 3.4.4.6-4: Results of LLP5 algorithm for input data “Burglaries for sector 3” for various future timespan and holdback values

From all replications of algorithm with test id LLP5\_3, which is the application of LLP5 algorithm with future timespan 48 periods and holdback 36 periods, produces the more reliable outcome regarding the values of R-squared and RMSE. In Figure 3.4.4.6-1 the actual and forecast charts are presented together in one graph for confidence interval 95%. Also, the prediction results are presented in Table 3.4.4.6-5.

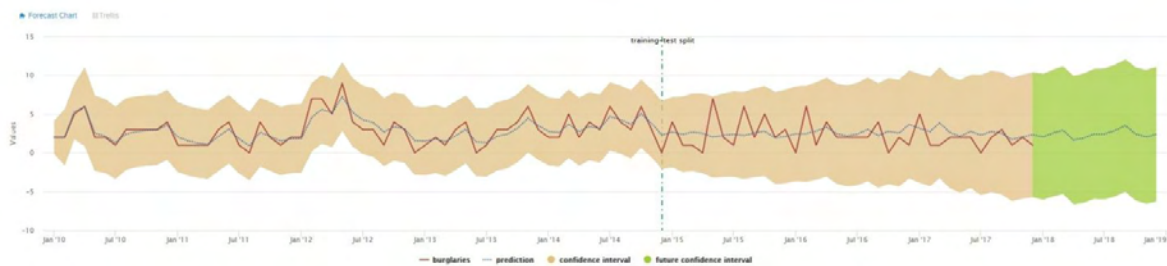


Figure 3.4.4.6-1: Forecast chart for burglaries in sector 3 with confidence interval 95%

Time period	Lower95(prediction)	Prediction	Upper95(prediction)
2018-01-01	-6.042605385 → 0	2.075260489 → 2	10.19312636 → 10
2018-02-01	-5.583559128 → 0	2.589516253 → 3	10.76259163 → 11
2018-03-01	-5.302208269 → 0	2.924990473 → 3	11.15218922 → 11
2018-04-01	-6.633281871 → 0	1.646989261 → 2	9.927260393 → 10
2018-05-01	-6.425908409 → 0	1.906417657 → 2	10.23874372 → 10
2018-06-01	-5.937087408 → 0	2.446308097 → 2	10.8297036 → 11

2018-07-01	-5.987999876 → 0	2.445510081 → 2	10.87902004 → 11
2018-08-01	-5.607068733 → 0	2.875629821 → 3	11.35832838 → 11
2018-09-01	-5.008211296 → 0	3.522777855 → 4	12.05376701 → 12
2018-10-01	-6.135232408 → 0	2.443175978 → 2	11.02158436 → 11
2018-11-01	-6.550143563 → 0	2.074838197 → 2	10.69981996 → 11
2018-12-01	-6.285618886 → 0	2.385114818 → 2	11.05584852 → 11

Table 3.4.4.6-5: Forecasted values regarding burglaries for 2018 for sector 3 with confidence interval 95%

### 3.4.4.7 Car robberies for sector 1

#### ➤ LL algorithm

Test id	Future timespan	Holdback	R-squared	RMSE
LL1	24	12	0.3293	0.88
LL2	36	24	0.3230	0.89
LL3	48	36	0.3331	0.88
LL4	60	48	0.2838	0.91

Table 3.4.4.7-1: Results of LL algorithm for input data “Car robberies for sector 1” for various future timespan and holdback values

#### ➤ LLT algorithm

Test id	Future timespan	Holdback	R-squared	RMSE
LLT1	24	12	0.6267	0.66
LLT2	36	24	0.4201	0.82

Table 3.4.4.7-2: Results of LLT algorithm for input data “Car robberies for sector 1” for various future timespan and holdback values

➤ LLP

Test id	Future timespan	Holdback	R-squared	RMSE
LLP1	24	12	0.7740	0.51
LLP2	36	24	0.7221	0.57
LLP3	48	36	0.5113	0.75
LLP4	60	48	0.3702	0.86

Table 3.4.4.7-3: Results of LLP algorithm for input data “Car robberies for sector 1” for various future timespan and holdback values

➤ LLP5

Test id	Future timespan	Holdback	R-squared	RMSE
LLP5_1	24	12	<b>0.8048</b>	<b>0.48</b>
LLP5_2	36	24	0.7723	0.52
LLP5_3	48	36	0.2933	0.91

Table 3.4.4.7-4: Results of LLP5 algorithm for input data “Car robberies for sector 1” for various future timespan and holdback values

From all replications of algorithm with test id LLP5\_1, which is the application of LLP5 algorithm with future timespan 24 periods and holdback 12 periods, produces the more reliable outcome regarding the values of R-squared and RMSE. In Figure 3.4.4.7-1 the actual and forecast charts are presented together in one graph for confidence interval 95%. Also, the prediction results are presented in Table 3.4.4.7-5.

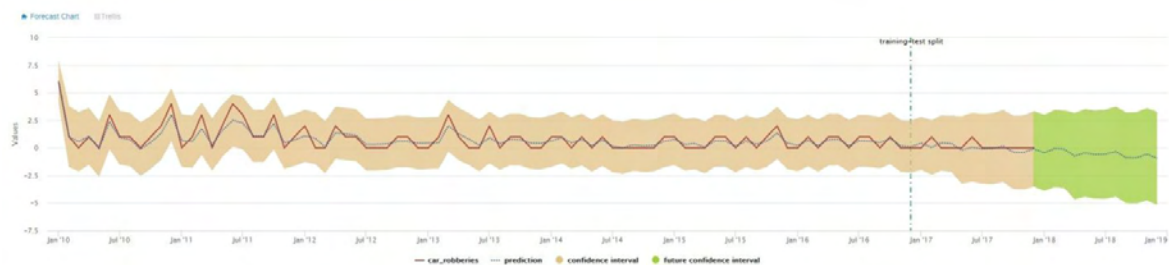


Figure 3.4.4.7-1: Forecast chart car robberies in sector 1 with confidence interval 95%

Time period	Lower95(prediction)	Prediction	Upper95(prediction)
2018-01-01	-3.887347082 → 0	-0.45068402 → 0	2.985979041 → 3
2018-02-01	-3.525953792 → 0	-0.051873623 → 0	3.422206547 → 3
2018-03-01	-3.636147762 → 0	-0.126563305 → 0	3.383021153 → 3
2018-04-01	-4.63911804 → 0	-0.742023401 → 0	3.155071238 → 3
2018-05-01	-4.398126727 → 0	-0.458192859 → 0	3.481741009 → 3
2018-06-01	-4.578663137 → 0	-0.597677681 → 0	3.383307775 → 3
2018-07-01	-4.593441178 → 0	-0.57307594 → 0	3.447289297 → 3
2018-08-01	-4.394608635 → 0	-0.336429944 → 0	3.721748748 → 4
2018-09-01	-4.990052091 → 0	-0.895529953 → 0	3.198992184 → 3
2018-10-01	-5.038469318 → 0	-0.908985563 → 0	3.220498192 → 3
2018-11-01	-4.746983912 → 0	-0.583839444 → 0	3.579305025 → 4
2018-12-01	-5.142115832 → 0	-0.946537119 → 0	3.249041593 → 3

Table 3.4.4.7-5: Forecasted values regarding car robberies for 2018 for sector 1 with confidence interval 95%

### 3.4.4.8 Car robberies for sector 2

#### ➤ LL algorithm

Test id	Future timespan	Holdback	R-squared	RMSE
LL1	24	12	0.3750	0.98

LL2	36	24	0.3322	1.02
LL3	48	36	0.3576	1.00
LL4	60	48	0.3518	1.00

Table 3.4.4.8-1: Results of LL algorithm for input data “Car robberies for sector 2” for various future timespan and holdback values

➤ LLT algorithm

Test id	Future timespan	Holdback	R-squared	RMSE
LLT1	24	12	0.4476	0.93
LLT2	36	24	0.4618	0.91

Table 3.4.4.8-2: Results of LLT algorithm for input data “Car robberies for sector 2” for various future timespan and holdback values

➤ LLP

Test id	Future timespan	Holdback	R-squared	RMSE
LLP1	24	12	0.4573	0.92
LLP2	36	24	0.3950	0.97
LLP3	48	36	0.3532	1.00
LLP4	60	48	0.3516	1.00

Table 3.4.4.8-3: Results of LLP algorithm for input data “Car robberies for sector 2” for various future timespan and holdback values

➤ LLP5

Test id	Future timespan	Holdback	R-squared	RMSE
---------	-----------------	----------	-----------	------



LLP5_1	24	12	<b>0.5216</b>	<b>0.86</b>
LLP5_2	36	24	0.4541	0.92
LLP5_3	48	36	0.1508	1.15
LLP5_4	60	48	0.0256	1.23

Table 3.4.4.8-4: Results of LLP5 algorithm for input data “Car robberies for sector 2” for various future timespan and holdback values

From all replications of algorithm with test id LLP5\_1, which is the application of LLP5 algorithm with future timespan 24 periods and holdback 12 periods, produces the more reliable outcome regarding the values of R-squared and RMSE. In Figure 3.4.4.8-1 the actual and forecast charts are presented together in one graph for confidence interval 95%. Also, the prediction results are presented in Table 3.4.4.8-5.

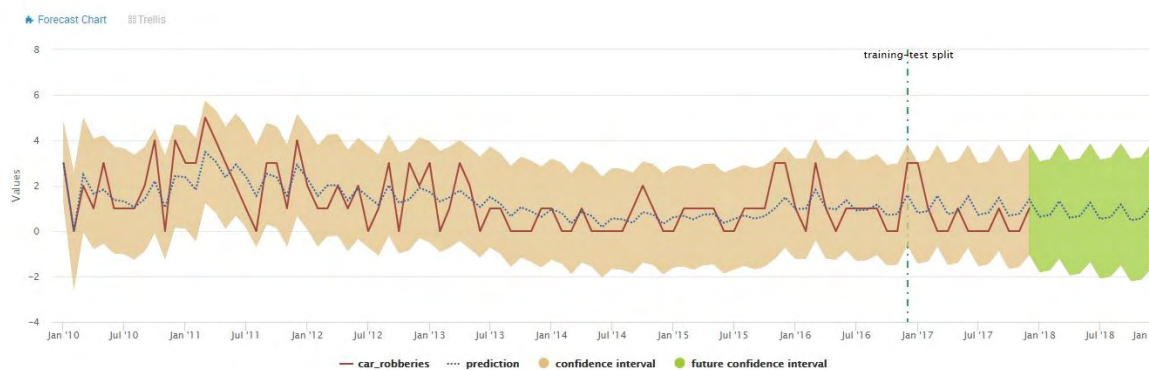


Figure 3.4.4.8-1: Forecast chart car robberies in sector 2 with confidence interval 95%

Time period	Lower95(prediction)	Prediction	Upper95(prediction)
2018-01-01	-1.805183243 → 0	0.632864214 → 1	3.070911672 → 3
2018-02-01	-1.725579387 → 0	0.718723008 → 1	3.163025402 → 3
2018-03-01	-1.199228288 → 0	1.320688705 → 1	3.840605697 → 4
2018-04-01	-1.941236738 → 0	0.585293228 → 1	3.111823194 → 3
2018-05-01	-1.865681005 → 0	0.667372566 → 1	3.200426137 → 3

2018-06-01	-1.354151093 → 0	1.24811739 → 1	3.850385872 → 4
2018-07-01	-2.073994205 → 0	0.535108672 → 1	3.144211549 → 3
2018-08-01	-2.001964604 → 0	0.613884858 → 1	3.229734319 → 3
2018-09-01	-1.503800696 → 0	1.175985369 → 1	3.855771433 → 4
2018-10-01	-2.204038569 → 0	0.482765173 → 0	3.169568915 → 3
2018-11-01	-2.135105593 → 0	0.558629858 → 1	3.252365308 → 3
2018-12-01	-1.649029045 → 0	1.10422008 → 1	3.857469206 → 4

Table 3.4.4.8-5: Forecasted values regarding car robberies for 2018 for sector 2 with confidence interval 95%

### 3.4.4.9 Car robberies for sector 3

#### ➤ LL algorithm

Test id	Future timespan	Holdback	R-squared	RMSE
LL1	24	12	0.1615	0.76
LL2	36	24	0.1445	0.77
LL3	48	36	0.1302	0.78
LL4	60	48	0.0490	0.81

Table 3.4.4.9-1: Results of LL algorithm for input data “Car robberies for sector 3” for various future timespan and holdback values

#### ➤ LLT algorithm

Test id	Future timespan	Holdback	R-squared	RMSE
LLT1	24	12	0.2138	0.74
LLT2	36	24	0.2672	0.71

LLT3	48	36	0.1643	0.76
LLT4	60	48	0.0113	0.83

Table 3.4.4.9-2: Results of LLT algorithm for input data “Car robberies for sector 3” for various future timespan and holdback values

➤ LLP

Test id	Future timespan	Holdback	R-squared	RMSE
LLP1	24	12	<b>0.4786</b>	<b>0.60</b>
LLP2	36	24	0.3028	0.70

Table 3.4.4.9-3: Results of LLP algorithm for input data “Car robberies for sector 3” for various future timespan and holdback values

➤ LLP5

Test id	Future timespan	Holdback	R-squared	RMSE
LLP5_1	24	12	0.4174	0.64
LLP5_2	36	24	0.3717	0.66
LLP5_3	48	36	0.3030	0.70
LLP5_4	60	48	0.2147	0.74

Table 3.4.4.9-4: Results of LLP5 algorithm for input data “Car robberies for sector 3” for various future timespan and holdback values

From all replications of algorithm with test id LLP1, which is the application of LLP algorithm with future timespan 24 periods and holdback 12 periods, produces the more reliable outcome regarding the values of R-squared and RMSE. In Figure 3.4.4.9-1 the actual and forecast

charts are presented together in one graph for confidence interval 95%. Also, the prediction results are presented in Table 3.4.4.9-5.

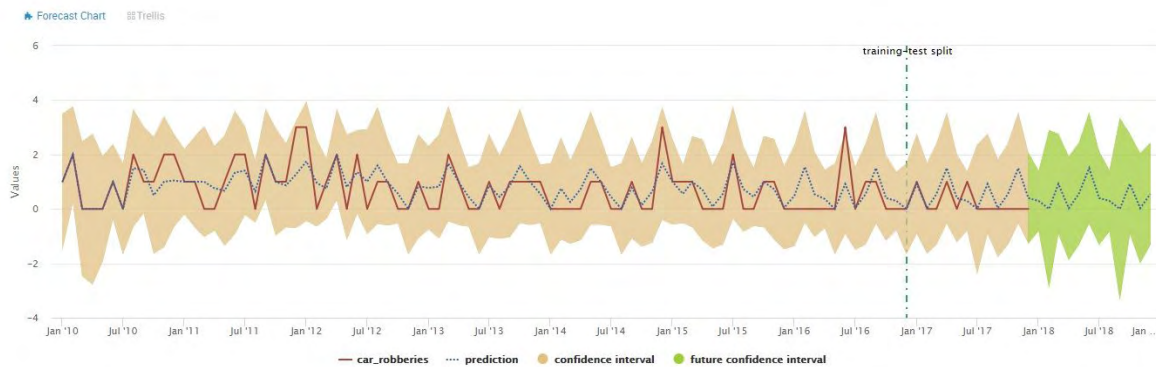


Figure 3.4.4.9-1: Forecast chart car robberies in sector 3 with confidence interval 95%

Time period	Lower95(prediction)	Prediction	Upper95(prediction)
2018-01-01	-0.812121927 → 0	0.292804512 → 0	1.39773095 → 1
2018-02-01	-2.893291556 → 0	1.27E-06 → 0	2.893294093 → 3
2018-03-01	-0.919560493 → 0	0.919234428 → 1	2.758029348 → 3
2018-04-01	-1.877981413 → 0	0.027018657 → 0	1.932018727 → 2
2018-05-01	-1.325867253 → 0	0.546551487 → 1	2.418970227 → 2
2018-06-01	-0.53999565 → 0	1.499999811 → 1	3.539995273 → 4
2018-07-01	-1.317031026 → 0	0.397702055 → 0	2.112435136 → 2
2018-08-01	-0.829756461 → 0	0.292804512 → 0	1.415365484 → 1
2018-09-01	-3.340571646 → 0	1.27E-06 → 0	3.340574183 → 3
2018-10-01	-0.9199298 → 0	0.919234428 → 1	2.758398656 → 3
2018-11-01	-1.987476777 → 0	0.027018657 → 0	2.04151409 → 2
2018-12-01	-1.333067054 → 0	0.546551487 → 1	2.426170028 → 2

Table 3.4.4.9-5: Forecasted values regarding car robberies for 2018 for sector 3 with confidence interval 95

#### 3.4.4.10 Motorcycle robberies for sector 1

➤ LL algorithm

Test id	Future timespan	Holdback	R-squared	RMSE
LL1	24	12	0.5495	1.55
LL2	36	24	0.5313	1.58

Table 3.4.4.10-1: Results of LL algorithm for input data “Motorcycle robberies for sector 1” for various future timespan and holdback values

➤ LLT algorithm

Test id	Future timespan	Holdback	R-squared	RMSE
LLT1	24	12	0.3540	1.85
LLT2	36	24	<b>0.5980</b>	<b>1.46</b>
LLT3	48	36	0.0325	2.27
LLT4	60	48	0.2438	2.01

Table 3.4.4.10-2: Results of LLT algorithm for input data “Motorcycle robberies for sector 1” for various future timespan and holdback values

➤ LLP

Test id	Future timespan	Holdback	R-squared	RMSE
LLP1	24	12	0.4597	1.70
LLP2	36	24	0.5324	1.58
LLP3	48	36	0.0870	2.20

Table 3.4.4.10-3: Results of LLP algorithm for input data “Motorcycle robberies for sector 1” for various future timespan and holdback values

Test id	Future timespan	Holdback	R-squared	RMSE
LLP5_1	24	12	0.4537	1.70

LLP5_2	36	24	0.5909	1.48
LLP5_3	48	36	0.1236	2.16
LLP5_4	60	48	0.2851	1.95

➤ LLP5

Table 3.4.4.10-4: Results of LLP5 algorithm for input data “Motorcycle robberies for sector 1” for various future timespan and holdback values

From all replications of algorithm with test id LLT2, which is the application of LLT algorithm with future timespan 36 periods and holdback 24 periods, produces the more reliable outcome regarding the values of R-squared and RMSE. In Figure 3.4.4.10-1 the actual and forecast charts are presented together in one graph for confidence interval 95%. Also, the prediction results are presented in Table 3.4.4.10-5.

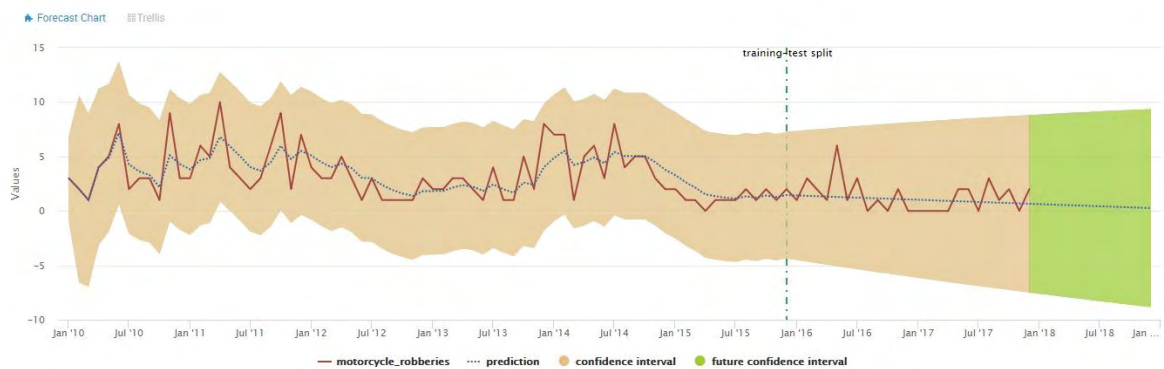


Figure 3.4.4.10-1: Forecast chart motorcycle robberies in sector 1 with confidence interval 95%

Time period	Lower95(prediction)	Prediction	Upper95(prediction)
2018-01-01	-7.609210394 → 0	0.623859244 → 1	8.856928883 → 9
2018-02-01	-7.724741255 → 0	0.590755656 → 1	8.906252568 → 9
2018-03-01	-7.839463038 → 0	0.557652068 → 1	8.954767175 → 9
2018-04-01	-7.953399109 → 0	0.52454848 → 1	9.00249607 → 9
2018-05-01	-8.066571733 → 0	0.491444892 → 0	9.049461517 → 9
2018-06-01	-8.17900214 → 0	0.458341304 → 0	9.095684748 → 9
2018-07-01	-8.290710597 → 0	0.425237716 → 0	9.141186029 → 9
2018-08-01	-8.401716464 → 0	0.392134127 → 0	9.185984719 → 9
2018-09-01	-8.512038249 → 0	0.359030539 → 0	9.230099327 → 9
2018-10-01	-8.621693663 → 0	0.325926951 → 0	9.273547566 → 9
2018-11-01	-8.730699668 → 0	0.292823363 → 0	9.316346394 → 9
2018-12-01	-8.839072513 → 0	0.259719775 → 0	9.358512063 → 9

Table 3.4.4.10-5: Forecasted values regarding motorcycle robberies for 2018 for sector 1 with confidence interval 95%

### 3.4.4.11 Motorcycle robberies for sector 2

#### ➤ LL algorithm

Test id	Future timespan	Holdback	R-squared	RMSE
LL1	24	12	0.4291	2.66
LL2	36	24	0.1786	3.19

Table 3.4.4.11-1: Results of LL algorithm for input data “Motorcycle robberies for sector 2” for various future timespan and holdback values

➤ LLT algorithm

Test id	Future timespan	Holdback	R-squared	RMSE
LLT1	24	12	0.5008	2.49
LLT2	36	24	0.4211	2.68

Table 3.4.4.11-2: Results of LLT algorithm for input data “Motorcycle robberies for sector 2” for various future timespan and holdback values

➤ LLP

Test id	Future timespan	Holdback	R-squared	RMSE
LLP1	24	12	0.5835	2.27
LLP2	36	24	0.3678	2.80
LLP3	48	36	0.2413	3.07
LLP4	60	48	0.2569	3.04

Table 3.4.4.11-3: Results of LLP algorithm for input data “Motorcycle robberies for sector 2” for various future timespan and holdback values

➤ LLP5

Test id	Future timespan	Holdback	R-squared	RMSE
LLP5_1	24	12	<b>0.6487</b>	<b>2.09</b>
LLP5_2	36	24	0.5012	2.49
LLP5_3	48	36	0.1274	3.29
LLP5_4	60	48	0.3231	2.90

Table 3.4.4.11-4: Results of LLP5 algorithm for input data “Motorcycle robberies for sector 2” for various future timespan and holdback values



From all replications of algorithm with test id LLP5\_1, which is the application of LLP5 algorithm with future timespan 24 periods and holdback 12 periods, produces the more reliable outcome regarding the values of R-squared and RMSE. In Figure 3.4.4.11-1 the actual and forecast charts are presented together in one graph for confidence interval 95%. Also, the prediction results are presented in Table 3.4.4.11-5.

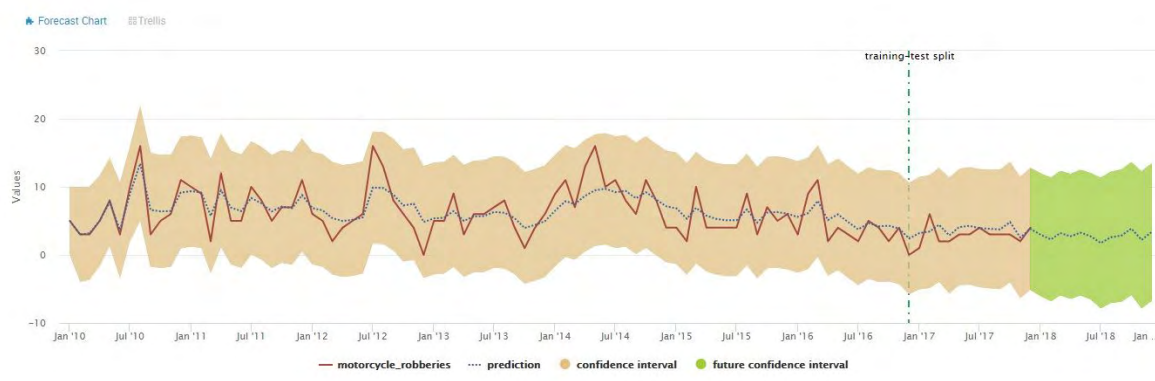


Figure 3.4.4.11-1: Forecast chart motorcycle robberies in sector 2 with confidence interval 95%

Time period	Lower95(prediction)	Prediction	Upper95(prediction)
2018-01-01	-6.049106474 → 0	2.967366258 → 3	11.98383899 → 12
2018-02-01	-6.805927424 → 0	2.264467317 → 2	11.33486206 → 11
2018-03-01	-5.920858996 → 0	3.202645253 → 3	12.3261495 → 12
2018-04-01	-6.447996314 → 0	2.727825823 → 3	11.90364796 → 12
2018-05-01	-5.950573125 → 0	3.27679536 → 3	12.50416384 → 13
2018-06-01	-6.55283859 → 0	2.725324017 → 3	12.00348662 → 12
2018-07-01	-7.8566398 → 0	1.747120559 → 2	11.35088092 → 11
2018-08-01	-7.063115469 → 0	2.594518402 → 3	12.25215227 → 12
2018-09-01	-6.88130177 → 0	2.829493534 → 3	12.54028884 → 13
2018-10-01	-5.897602228 → 0	3.86565881 → 4	13.62891985 → 14
2018-11-01	-7.844018692 → 0	2.212219799 → 2	12.26845829 → 12
2018-12-01	-6.760757255 → 0	3.350485949 → 3	13.46172915 → 13

Table 3.4.4.11-5: Forecasted values regarding motorcycle robberies for 2018 for sector 2 with confidence interval 95

### 3.4.4.12 Motorcycle robberies for sector 3

#### ➤ LL algorithm

Test id	Future timespan	Holdback	R-squared	RMSE
LL1	24	12	0.0247	1.95

Table 3.4.4.12-1: Results of LL algorithm for input data “Motorcycle robberies for sector 3” for various future timespan and holdback values

#### ➤ LLT algorithm

Test id	Future timespan	Holdback	R-squared	RMSE
LLT1	24	12	0.1360	1.83
LLT2	36	24	0.2618	1.69

Table 3.4.4.12-2: Results of LLT algorithm for input data “Motorcycle robberies for sector 3” for various future timespan and holdback values

#### ➤ LLP

Test id	Future timespan	Holdback	R-squared	RMSE
LLP1	24	12	<b>0.5239</b>	<b>1.36</b>
LLP2	36	24	0.1198	1.85

Table 3.4.4.12-3: Results of LLP algorithm for input data “Motorcycle robberies for sector 3” for various future timespan and holdback values

➤ LLP5

Test id	Future timespan	Holdback	R-squared	RMSE
LLP5_1	24	12	0.4562	1.45
LLP5_2	36	24	0.2381	1.72

Table 3.4.4.12-4: Results of LLP5 algorithm for input data “Motorcycle robberies for sector 3” for various future timespan and holdback values

From all replications of algorithm with test id LLP1, which is the application of LLP algorithm with future timespan 24 periods and holdback 12 periods, produces the more reliable outcome regarding the values of R-squared and RMSE. In Figure 3.4.4.12-1 the actual and forecast charts are presented together in one graph for confidence interval 95%. Also, the prediction results are presented in Table 3.4.4.12-5.

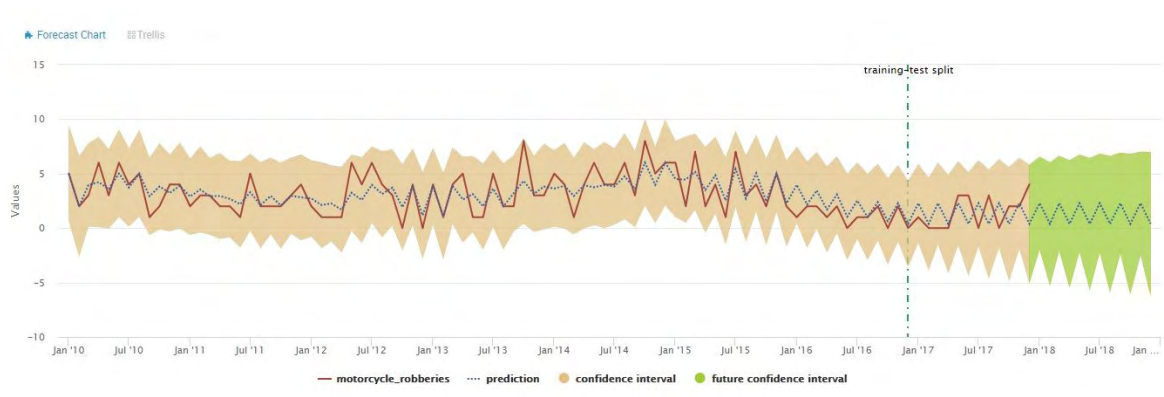


Figure 3.4.4.12-1: Forecast chart motorcycle robberies in sector 3 with confidence interval 95%

Time period	Lower95(prediction)	Prediction	Upper95(prediction)
2018-01-01	-1.967659053 → 0	2.284752598 → 2	6.53716425 → 7
2018-02-01	-5.254178209 → 0	0.384832801 → 0	6.023843812 → 6
2018-03-01	-2.065793626 → 0	2.284752598 → 2	6.635298823 → 7
2018-04-01	-5.457360243 → 0	0.384832801 → 0	6.227025846 → 6
2018-05-01	-2.161762897 → 0	2.284752598 → 2	6.731268094 → 7
2018-06-01	-5.653709569 → 0	0.384832801 → 0	6.423375172 → 6
2018-07-01	-2.255704169 → 0	2.284752598 → 2	6.825209365 → 7
2018-08-01	-5.843872389 → 0	0.384832801 → 0	6.613537992 → 7
2018-09-01	-2.347740818 → 0	2.284752598 → 2	6.917246015 → 7
2018-10-01	-6.028399047 → 0	0.384832801 → 0	6.79806465 → 7
2018-11-01	-2.437984199 → 0	2.284752598 → 2	7.007489395 → 7
2018-12-01	-6.207762828 → 0	0.384832801 → 0	6.977428431 → 7

Table 3.4.4.12-5: Forecasted values regarding motorcycle robberies for 2018 for sector 3  
with confidence interval 95%

### 3.5 Estimation of the center point of busiest day graph network

In this chapter the absolute center point of the network graph for the busiest day is estimated. By the term “busiest day”, the day with the largest number of crime incidents in the data set is defined. In order to calculate the absolute center of the graph network for the busiest day, the following steps needed to be followed:

1. The busiest day with its crime incidents is defined
2. The matrix of distances between the crime incidents of the busiest day is calculated based on incidents' coordinates
3. The minimum spanning tree of the graph network for busiest day is defined by applying Kruskal or Prim algorithm
4. The absolute center is calculated by applying the Absolute Center Algorithm.

#### 3.5.1 Busiest day

In order to estimate the busiest day the number of crime incidents needed to be counted for each unique available date in the spreadsheet. For this purpose, the Excel function COUNTIF(range, criteria) is applied. COUNTIF is a function used to count cells that meet a single criterion. It can be used to count cells with dates, numbers, and text that match specific criteria.

As a result of COUNTIF function application, a list of numbers indicating the sum of incidents for every unique date value is produced. Moreover, by sorting the numbers from largest to smallest, the dates are also sorted respectively. The busiest day is the 1<sup>st</sup> of February of 2014 with 8 incidents occurred. In Figure 3.5.1.1 a screenshot of the list of the dates and the number of crime incidents for each date are presented sorted in a descending order, so on the top of the list the busiest day is ranked. The highlighted value represents the busiest day with the number of incidents occurred on that day.

	A	B
1	Date	For all areas
2	2/1/2014	8
3	1/1/2012	7
4	3/10/2011	6
5	10/12/2011	6
6	8/15/2012	6
7	10/18/2013	6
8	11/8/2013	6
9	2/3/2014	6
10	8/11/2015	6
11	1/11/2010	5
12	4/27/2010	5

Figure 3.5.1-1: List of the dates and the number of crime incidents for each date

As a next step, the coordinates of the incidents on busiest day are used in order to calculate the distances between all points of incidents and display the incidents on the map. In Figure 3.5.1-2 the incident points of the busiest day mapped on the map is illustrated.

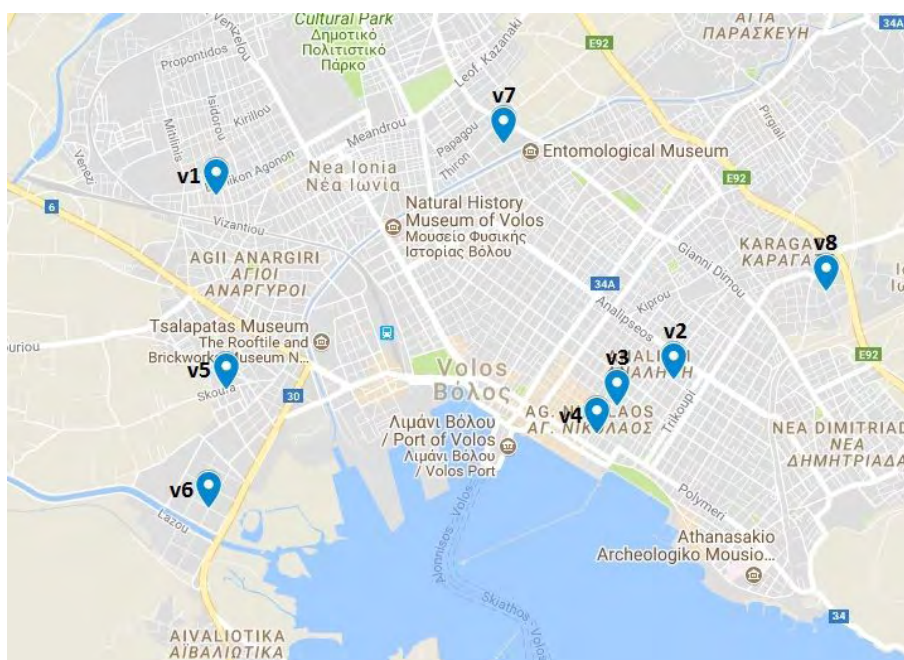


Figure 3.5.1-2: Incident points of busiest day

The distances between all incidents points are calculated with a java code scripted by Professor Athanasios Lois. The script takes as an input a text file with the coordinates of its point and uses a Google Maps Key in order to perform the calculation of distances. The outcome of the

code execution is a  $n \times n$  matrix containing the distances between all points, where  $n$  is the number of the points included in the input text file. In Table 3.5.1-1 the matrix with the distances between the 8 incident points of the busiest day is displayed. The  $8 \times 8$  matrix of the distances is used in order to estimate the minimum spanning tree (MST) of the graph network for busiest day.

Distances between nodes								
	1	2	3	4	5	6	7	8
1	0	4145	3708	3819	2150	3242	2350	5376
2	4191	0	1271	863	3596	3915	2803	1377
3	3728	495	0	263	2996	3315	2920	1857
4	3741	737	240	0	3298	3617	2750	2100
5	2085	3518	2732	2843	0	1168	2595	9895
6	3296	3912	3126	3237	943	0	3192	5322
7	2663	2364	3093	3215	2828	4068	0	2806
8	5550	1625	2490	2082	10562	5700	3020	0

Table 3.5.1-1: The  $8 \times 8$  matrix containing the distances between all points of busiest day

### 3.5.2 Minimum spanning tree of busiest day graph network

Given the matrix of distances between all points, the MST of the network graph of busiest day is estimated. For this purpose, a Java code scripted by Professor Athanasios Lois, which implements Prim algorithm, is used. The outcome of this code is the edges of the MST for the network graph of busiest day. In Table 3.5.2-1 the edges of MST with their distance cost and also the total cost distance of MST are presented. Moreover, in Figure 3.5.2-1, the MST of the network graph of busiest day is illustrated based on the outcome presented on Table 3.5.2-1.

Minimum Spanning Tree			
Edge			Distance
v1	–	v7	2350
v2	–	v8	1377
v3	–	v4	263
v4	–	v2	737
v5	–	v1	2085
v6	–	v5	943
v7	–	v2	2364
Total distance of MST			10119

Table 3.5.2-1: MST edges with distance cost and total distance cost of MST

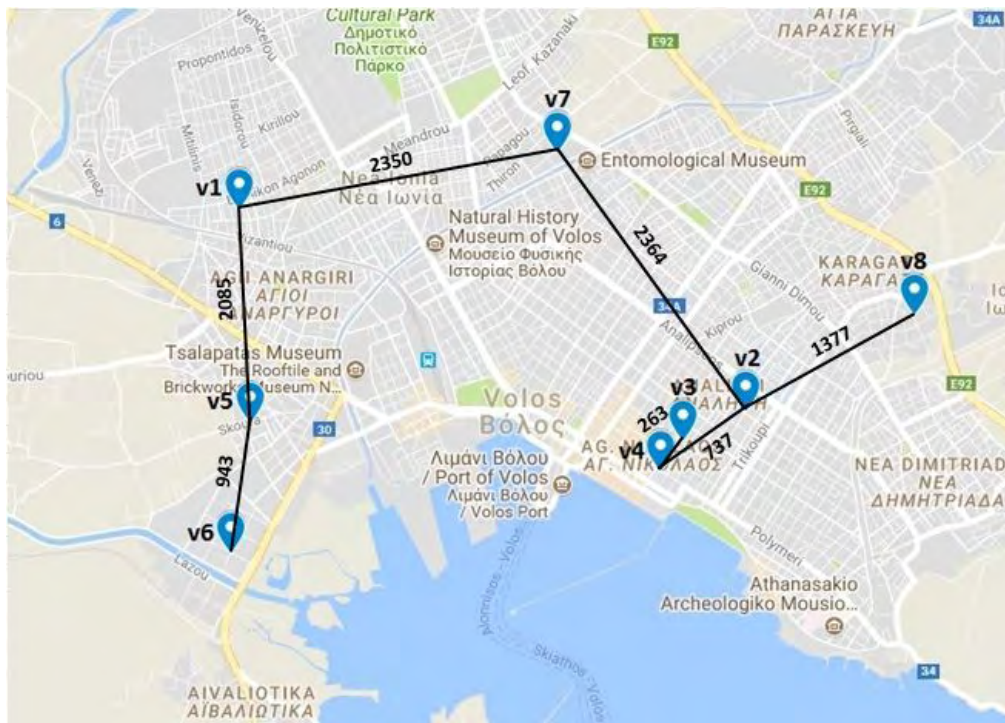


Figure 3.5.2-1: The MST of busiest day graph network

### 3.5.3 Estimation of the absolute center point for busiest day graph network

The purpose of this chapter of the thesis belongs to the category of facility location problems which is part of the extensive category of urban service system problems. These problems are concerned with determining good locations for the stationing of service vehicles or the construction of major facilities. Even though facility location problem arise in the context of both



routine and emergency services, in this chapter the emergency aspect, which refers to the center problem, is studied.

Center problems are concerned with a prespecified number of facilities which must be located so as to minimize the maximum distance (or time or cost), to or from the facilities, that any user will have to travel. Center problems, also referred to as minimax problems, are more applicable in the context of emergency urban services, such as emergency medical care, firefighting, and crime fighting.

Given the MST of the graph network for busiest day, the center point of the graph network can be estimated. In the case of police service, the absolute center indicates the point where the police car can be parked for stakeout.

There are two basic algorithms for the estimation of the center point of the graph, the node-restricted algorithm and the absolute center algorithm. The first one estimates the center point of the graph network exclusively on one of the vertices of the graph. On the other hand, absolute center algorithm can determine the center point of the graph on any point of the graph, as it takes into account not only the vertices but also the edges of the graph. Therefore, the result of the absolute center algorithm is considered more sufficient than the outcome of node-restricted algorithm which should be on one of the vertices of the graph.

For the purpose of this study the absolute center algorithm is applied. In paragraph 3.5.3.1 the absolute center algorithm is introduced and in 3.5.3.2 the outcomes after the implementation of the algorithm on the graph network of the busiest day are presented.

### 3.5.3.1 Absolute center algorithm

The absolute center problem is defined by the objective function:

$$\min_{x \in P(G)} g(x) := \max_{i=1, \dots, n} d(x, v_i), x \in P(T) \quad (3.5.3.1.1)$$

The aim is to estimate a point  $x \in P(G)$ , which minimizes the function  $g(\cdot)$ , so that  $g(x^*) \leq g(x), \forall x \in P(G)$ . This point is called optimal or absolute center. Also,  $X^*(g(G))$  denotes the set of all absolute centers of the function  $g(\cdot)$ .

The algorithm applied for resolving the absolute center problem is described by the following steps.

Step 1: Choose an arbitrary point  $x \in P(T), T = (V, E)$ .

Step 2: Select the node  $v$  in  $T$  that maximal distance to  $x$ .

$$d(v, x) := \max_{i=1, \dots, n} d(v_i, x) \quad (3.5.3.1.2)$$

Step 3: Select the node  $v'$  in  $T$  that maximal distance to  $v$ .

$$d(v', v) := \max_{i=1, \dots, n} d(v_i, v) \quad (3.5.3.1.3)$$

Step 4:  $x^* =$  center of  $P(v, v')$ :  $d(x^*, v) = d(x^*, v') = \frac{1}{2} d(v, v') = g(x^*)$ , where  $P(v, v')$  is the longest path of the minimum spanning tree of the graph.

The complexity of the algorithm is  $O(n)$ .

### ***3.5.3.2 Center point for the graph network of busiest day***

The absolute center algorithm is applied on the network graph of the busiest day in order to estimate the absolute center point, where the police car can stakeout. In Table 3.5.3.2-1 algorithm's steps and calculations are presented thoroughly.

1. Choose:	<b>x = v7</b>
2. Calculate all nodes' distances from x = v7:	$d(v1, v7) = 2350$ $d(v2, v7) = 2364$ $d(v3, v7) = 3364$ $d(v4, v7) = 3101$ $d(v5, v7) = 4435$ $d(v6, v7) = 5378$ $d(v7, v7) = 0$ $d(v8, v7) = 3741$
Select the node that has maximal distance to x:	$d(v, x) = \max\{d(v1, v7), d(v2, v7), d(v3, v7), d(v4, v7), d(v5, v7), d(v6, v7), d(v7, v7), d(v8, v7)\}$ $d(v, x) = 5378$ <b>v = v6</b>
3. Calculate all node's distances from v = v6:	$d(v1, v6) = 3028$ $d(v2, v6) = 7742$ $d(v3, v6) = 8742$ $d(v4, v6) = 8479$ $d(v5, v6) = 943$ $d(v6, v6) = 0$ $d(v7, v6) = 5378$ $d(v8, v6) = 9119$
Select the node that has maximal distance to v:	$d(v, v') = \max\{d(v1, v6), d(v2, v6), d(v3, v6), d(v4, v6), d(v5, v6), d(v6, v6), d(v7, v6), d(v8, v6)\}$ $d(v, v') = 9119$ <b>v' = v8</b>
4. Estimate absolute center x*:	$d(v6, v8) = 4559.5$ $x^* = ([v1, v7], 1531.5/2350) = ([v1, v7], 0.65)$

Table 3.5.3.2-1: Implementation of absolute center algorithm on the network graph of busiest day

The absolute center of the network graph for the busiest day is the point on edge  $[v1, v7]$  and is located on the 65% of the distance from vertex  $v1$ . The coordinates of the absolute center are: 39.37391, 22.93782. In Table 3.5.3.2-2 the coordinates of the incident points and the absolute center point of the busiest day are presented. In addition, Figure 3.5.3.2-1 illustrates how the incidents and the absolute center of the busiest day are distributed on the map.

Node	Node ID	Latitude	Longitude
v1	1749	39.37185	22.924702
v2	1750	39.36214	22.955865
v3	1751	39.36077	22.951969
v4	1752	39.35933	22.950616
v5	1753	39.36163	22.925385
v6	1754	39.35541	22.924213
v7	1755	39.37459	22.944228
v8	1756	39.36684	22.966231
Center	center	39.37391	22.93782

Table 3.5.3.2-2: Coordinates of incident points and absolute center for busiest day

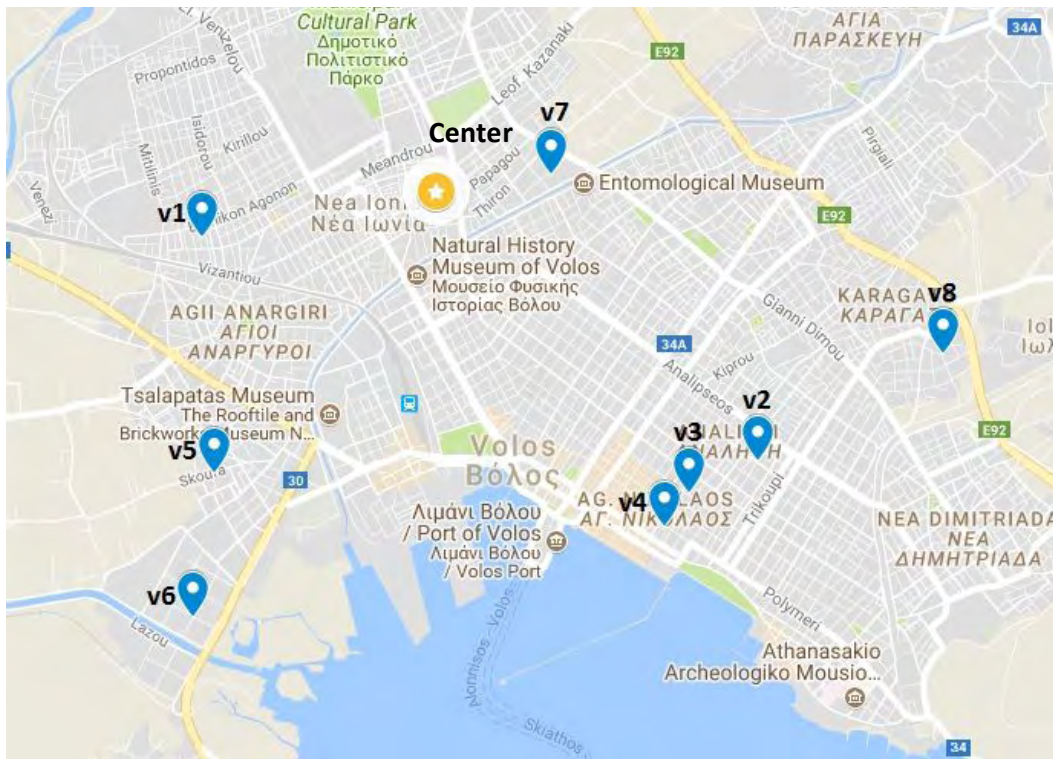


Figure 3.5.3.2-1: Incidents and absolute center of the busiest day

### 3.6 The Dial -A- Ride Problem (DARP)

As described in Chapter 2 in the Section 2.4.3, the Dial-A-Ride Problem (DARP) consists in defining a set of routes that satisfy transportation requests of users between a set of pickup points and a set of delivery points, in the presence of ride time constraints. The DARP has been examined and applied in this study in order to study how the police can serve all the incidents occurred in the busiest day.

Was selected for studying the busiest day, the day when most offenses occurred, so as to address the problem in its most demanding form. This day was defined before in Section 3.5 and is 1<sup>st</sup> February 2014. The incidents that took place that day were eight. There were no time information on half of the incidents, therefore the incident time was randomly selected for every incident lacked in time information. The following table presents the busiest day with all the details.

Table 3.6.1: Details of the busiest day

Id	Date	Time	Offense	Type of Offense	Address	Latitude	Longitude
1749	1/2/2014	15:40:00	BURGLARY	MISDEMEANOR	ENEZI ILIA 21. N. IONIA	39.37185	22.924702
1750	1/2/2014	9:00:00	BURGLARY	MISDEMEANOR	KORONIOY, VOLOS	39.36214	22.955865
1751	1/2/2014	21:00:00	BURGLARY	MISDEMEANOR	FRIXOY 1, VOLOS	39.36077	22.951969
1752	1/2/2014	19:00:00	BURGLARY	MISDEMEANOR	ODYSSEOS 6, VOLOS	39.35933	22.950616
1753	1/2/2014	22:00:00	BURGLARY	MISDEMEANOR	SAMOY 10, N. IONIA	39.36163	22.925385
1754	1/2/2014	2:00:00	MOTORCYCLE ROBBERY	MISDEMEANOR	MAGNITON 199, VOLOS	39.35541	22.924213
1755	1/2/2014	0:10:00	MOTORCYCLE ROBBERY	MISDEMEANOR	GALLIAS 65, VOLOS	39.37459	22.944228
1756	1/2/2014	1:00:00	MOTORCYCLE ROBBERY	MISDEMEANOR	SKYROY, VOLOS	39.36684	22.966231

In order to apply the DARP problem and take results we used the Microsoft Access and the windows tool, command line interpreter, CMD commands. In particular, there were created several forms in Access based on data details about the distances between the incidents, the stops

(the Police Department and the site of the offence), the vehicles used and the details of trips. Actually the Detail\_Trips\_Query\_Form were filled in with the proper time windows, as shown in the figure 3.6.1. Generally there were applied some assumptions, based on discussions with the police, so as to set the problem. The Police Services located El. Venizelou 158 - Patouxa, was determined as the base for the patrol cars. The police vehicle in service is one. The vehicle can leave for the site of the offense at the earliest, at the time of the offense and at the latest at the 40 minutes after the incident. It can stay in the scene of the incident up to 30 minutes and then can return to the police station ( the base) at the earliest at the same time that the incident occurred, this means that the police vehicle never left for the incident, and at the latest at the end of the day. And finally the time that takes to deliver was set the 10 minutes.

DESCRIPTION	inc1
FROM_NODE	1749
FROM_NODE_ET	900
FROM_NODE_LT	940
FROM_NODE_ST	30
TO_NODE	1
TO_NODE_ET	900
TO_NODE_LT	1440
TO_NODE_ST	10
MAX_RIDE_TIME	1440
AMMOUNT	1
FEATURES	NOF
DATE	

(a)

DESCRIPTION	INC2
FROM_NODE	1750
FROM_NODE_ET	540
FROM_NODE_LT	580
FROM_NODE_ST	30
TO_NODE	1
TO_NODE_ET	540
TO_NODE_LT	1440
TO_NODE_ST	10
MAX_RIDE_TIME	1440
AMMOUNT	1
FEATURES	NOF
DATE	

(b)

DESCRIPTION	INC3
FROM_NODE	1751
FROM_NODE_ET	1260
FROM_NODE_LT	1300
FROM_NODE_ST	30
TO_NODE	1
TO_NODE_ET	1260
TO_NODE_LT	1440
TO_NODE_ST	10
MAX_RIDE_TIME	1440
AMMOUNT	1
FEATURES	NOF
DATE	

(c)

DESCRIPTION	inc4
FROM_NODE	1752
FROM_NODE_FT	1140
FROM_NODE_LT	1180
FROM_NODE_ST	30
TO_NODE	1
TO_NODE_ET	1140
TO_NODE_LT	1440
TO_NODE_ST	10
MAX_RIDE_TIME	1440
AMMOUNT	1
FEATURES	nof
DATE	

(d)

DESCRIPTION	inc5
FROM_NODE	1753
FROM_NODE_ET	1320
FROM_NODE_LT	1340
FROM_NODE_ST	30
TO_NODE	1
TO_NODE_ET	1320
TO_NODE_LT	1440
TO_NODE_ST	10
MAX_RIDE_TIME	1440
AMMOUNT	1
FEATURES	nof
DATE	

(e)

DESCRIPTION	inc6
FROM_NODE	1754
FROM_NODE_FT	120
FROM_NODE_LT	160
FROM_NODE_ST	30
TO_NODE	1
TO_NODE_ET	120
TO_NODE_LT	1440
TO_NODE_ST	10
MAX_RIDE_TIME	1440
AMMOUNT	1
FEATURES	nof
DATE	

(f)

DESCRIPTION	inc7
FROM_NODE	1755
FROM_NODE_FT	10
FROM_NODE_LT	50
FROM_NODE_ST	30
TO_NODE	1
TO_NODE_ET	10
TO_NODE_LT	1440
TO_NODE_ST	10
MAX_RIDE_TIME	1440
AMMOUNT	1
FEATURES	nof
DATE	

(g)

DESCRIPTION	inc8
FROM_NODE	1756
FROM_NODE_FT	60
FROM_NODE_LT	100
FROM_NODE_ST	30
TO_NODE	1
TO_NODE_ET	60
TO_NODE_LT	1440
TO_NODE_ST	10
MAX_RIDE_TIME	1440
AMMOUNT	1
FEATURES	nof
DATE	

(h)

Figure 3.6.1: Trips detail for every incident of the busiest day

FROM\_NODE: the identification number (id) of the incident

FROM\_NODE\_ET: the earliest that the vehicle can leave for the scene of the incident

FROM\_NODE\_LT: latest the vehicle can leave for the scene of the incident

FROM\_NODE\_ST: time that the vehicle spends on the scene of the incident

TO\_NODE: the Police Department

TO\_NODE\_ET: the earliest that the vehicle returns to the Police Department

TO\_NODE\_LT: the latest that the vehicle returns to the Police Department

TO\_NODE\_ST: time that the vehicle spends at the Police Department after the return (deliver time)

MAX\_RIDE\_TIME: the maximum time that the vehicle can drive on the roads

Subsequently we proceeded to build in Access and write in CMD commands in order to run the DARP problem and take results. In fact by the use of CMD we executed the Algorithm for the DARP problem made by Professor Athanasios Lois. The results were printed in notepad format.

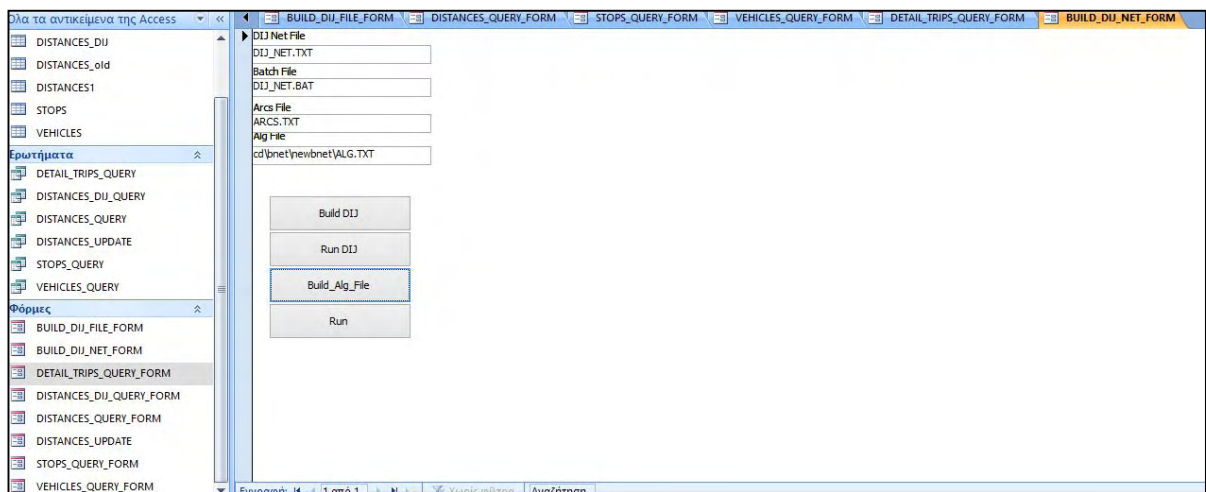
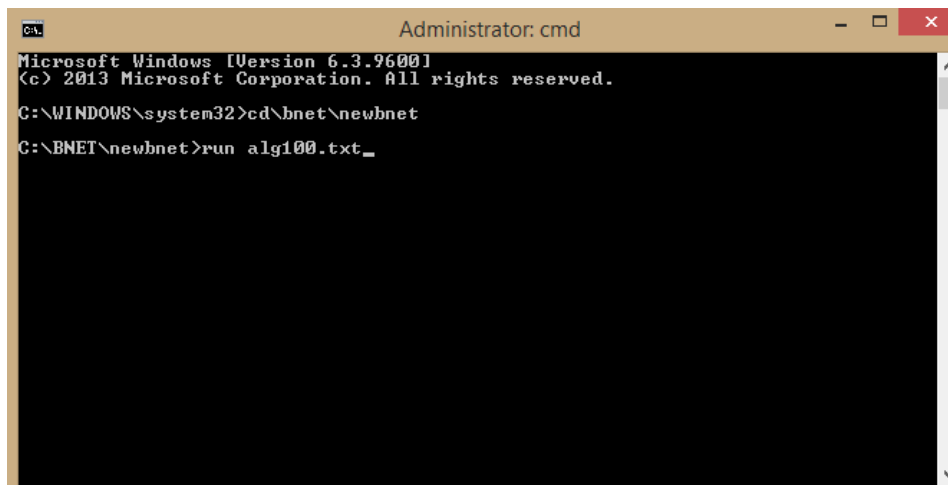


Figure 3.6.2: Representation of the steps in Access





```
Administrator: cmd
Microsoft Windows [Version 6.3.9600]
(c) 2013 Microsoft Corporation. All rights reserved.
C:\WINDOWS\system32>cd\bnet\newbnet
C:\BNET\newbnet>run alg100.txt_
```

Figure 3.6 3: The apply of the alg100.txt

**The results of the run:**

```
C:\BNET\newbnet>EODARP ALG100.TXT 10000000 3 4 5 6
DEMAND_ALLOCATED:
START_DEMANDS_SESSION
DEMAND:inc7 ID:7 ALLOCATED_ON_ROUTES:1 CHARGING_COST:1
DEMAND:inc8 ID:8 ALLOCATED_ON_ROUTES:1 CHARGING_COST:1
DEMAND:inc6 ID:6 ALLOCATED_ON_ROUTES:1 CHARGING_COST:1
DEMAND:INC2 ID:2 ALLOCATED_ON_ROUTES:1 CHARGING_COST:1
DEMAND:inc1 ID:1 ALLOCATED_ON_ROUTES:1 CHARGING_COST:1
DEMAND:inc4 ID:4 ALLOCATED_ON_ROUTES:1 CHARGING_COST:1
DEMAND:INC3 ID:3 ALLOCATED_ON_ROUTES:1 CHARGING_COST:1
DEMAND:inc5 ID:5 ALLOCATED_ON_ROUTES:1 CHARGING_COST:1
END_DEMANDS_SESSION
START_ROUTE_SESSION
ROUTE_ID:1
VEHICLE_DESCR:bom200
ROUTE_DISTANCE:91.87
ROUTE_TIME:1372
TOTAL_PASSENGERS_DISTANCE:46.818
AVERAGE_PASSENGERS_PER_DISTANCE_UNIT:0.542987
ROUTE_SLOPE:5.10389
VEHICLE_COST:16
```

VEHICLE\_CHARGING\_COST:16  
START\_SCHEDULE\_SESSION  
bom200 SDEPOT 0 -10000 0 0 0 0 1  
bom200 inc7 1 1 10 40 2 1 1755  
bom200 inc7 2 -1 49 59 0 -1 1  
bom200 inc8 3 1 67 97 0 1 1756  
bom200 inc8 4 -1 106 116 0 -1 1  
bom200 inc6 5 1 126 156 0 1 1754  
bom200 inc6 6 -1 167 177 0 -1 1  
bom200 INC2 7 1 540 570 351 1 1750  
bom200 INC2 8 -1 583 593 0 -1 1  
bom200 inc1 9 1 900 930 299 1 1749  
bom200 inc1 10 -1 937 947 0 -1 1  
bom200 inc4 11 1 1140 1170 178 1 1752  
bom200 inc4 12 -1 1185 1195 0 -1 1  
bom200 INC3 13 1 1260 1290 49 1 1751  
bom200 INC3 14 -1 1304 1314 0 -1 1  
bom200 inc5 15 1 1323 1353 0 1 1753  
bom200 inc5 16 -1 1362 1372 0 -1 1  
bom200 EDEPOT 17 -20000 1372 1372 0 0 1  
END\_SCHEDULE\_SESSION  
END\_ROUTE\_SESSION  
START\_STATISTICS\_SESSION  
ALLOCATED\_DEMANDS\_NUMBER:8  
NOT\_ALLOC\_DEM\_NUM:0  
PERCENTAGE\_OF\_NON\_ALLOCATED:0  
ALLOCATED\_ROUTES\_NUMBER:1  
NOT\_ALLOC\_ROUTE\_NUM:0  
PERCENTAGE\_OF\_ALLOCATED\_ROUTES:100  
TOTAL\_DISTANCE\_UNITS:91.87  
TOTAL\_TIME:1372  
TOTAL\_PASSENGERS\_DISTANCE:46.818  
AVGERAGE\_PASSENGERS\_DISTANCE:0.542987  
TOTAL\_DEMANDS\_SP\_DISTANCE:46.818

TOTAL\_DEMANDS\_REAL\_DISTANCE:46.818  
 TOTAL\_DEMAND\_SP\_TIME:87  
 TOTAL\_DEMAND\_REAL\_TIME:407  
 AVERAGE\_TIME\_DEVIATION:40  
 AVERAGE\_DISTANCE\_DEVIATION:0  
 AVERAGE\_DISTANCE\_DEVIATION\_PERCENTAGE:0  
 AVERAGE\_TIME\_DEVIATION\_PERCENTAGE:367  
 TOTAL\_VEHICLE\_COST:16  
 TOTAL\_TRIP\_CHARGING\_COST:8  
 PROFIT:-8  
 END\_STATISTICS\_SESSION  
 TOTAL\_EXECUTION\_TIME:0

Table 3.6.2: Aggregate table of the DARP results

Incident Id	Time that the vehicle leaves for the scene incident	Time that the vehicle leaves from the scene incident	Time that the vehicle arrives at the police station	Deliver Delay	Time that the vehicle is available again
1755	10	40	49	10	59
1756	67	97	106	10	116
1754	126	156	167	10	177
1750	540	570	583	10	593
1749	900	930	937	10	947
1752	1140	1170	1185	10	1195
1751	1260	1290	1304	10	1314
1753	1323	1353	1362	10	1372

The above table presents the DARP results showing how the busiest day can be handled.

### **3.7 Estimation of the medians of city sectors**

Volos is a medium size city covering an area of 27.68 km<sup>2</sup>. Although the size of the city is not so big, in order for the police officers to serve more efficient crime incidents occurring from time to time, the city is divided into sectors. For each sector, the crime incidents according to the historical data are defined and the median point is estimated. The median points are considered valuable information for police officers, as they indicate the points which minimize the average distance to or from the incidents occurred in the sectors. Therefore, in this chapter the median points for sectors with high criminality are estimated.

#### **3.7.1 City sectors and incidents estimation**

According to the size of the city and the clustering outcomes of the crime incidents occurred from 2010 until 2017, the city is divided into 68 sectors. In Figure 3.7.1-1, the sectors of the city are depicted.

In each sector, the number of incidents with their coordinates is determined. In Table 3.7.1-1 part of the calculations regarding the number of incidents occurred in the city sectors in and around city center is presented. (The complete table can be found in the Appendix section 7.4)

Based on incidents' coordinates, the network graph of the sector is estimated and the distances between incident points are calculated. For this purpose a Java code scripted by Professor Athanasios Lois is used. The outcome of the code's execution is a matrix containing the distances between all incident points in the sector. The distance matrix of each sector is fundamental information for estimating the median point of the sector, as it is stated in paragraph 3.7.2 in which the steps of median algorithm are defined.

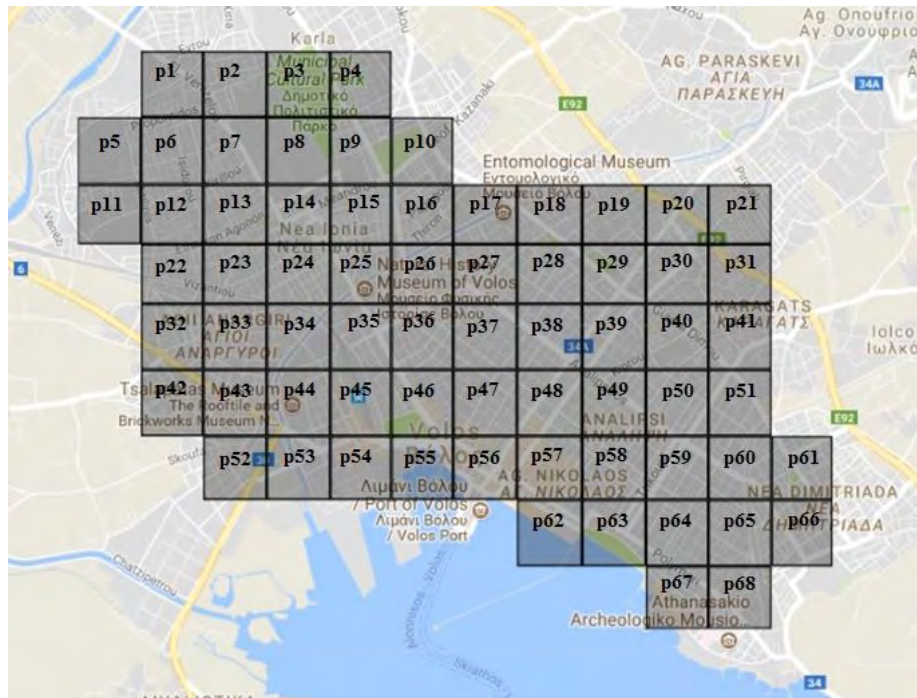


Figure 3.7.1-1: City sectors

Sector	Lat1	Lon1	Lat2	Lon2	Number of Incidents
p30	39.37256	22.955728	39.369441	22.959998	29
p31	39.37256	22.959998	39.369441	22.964236	13
p32	39.369441	22.92168	39.365999	22.92587	69
p33	39.369441	22.92587	39.365999	22.930183	53
p34	39.369441	22.930183	39.365999	22.934496	17
p35	39.369441	22.934496	39.365999	22.938637	54
p36	39.369441	22.938637	39.365999	22.942778	52
p37	39.369441	22.942778	39.365999	22.947118	42
p38	39.369441	22.947118	39.365999	22.951458	42
p39	39.369441	22.951458	39.365999	22.955728	47
p40	39.369441	22.955728	39.365999	22.959998	28

Table 3.7.1-1: Number of incidents occurred in city sectors p30 to p40

### 3.7.2 Median problem

The median problem falls into the category of facility location problems. It is related to the location of a prespecified number of facilities so as to minimize the average distance or average travel time to or from the facilities for the population of their users. Median problems arise very

often in the context of facility construction for delivery of nonemergency services, such as post offices, transportation terminals, telephone interchanges etc.

For the purposes of this study the median points of the network graph of sectors for the city of Volos are estimated. Even though police force is an emergency service, medians of city sectors estimated based on the historical data will provide valuable information regarding the points which minimizes the average distance to the incidents occurred. These points can be considered as hot spots across the city, especially for sectors with high volume of criminality and police officers can do patrol around them.

### 3.7.2.1 Problem description

Given an undirected network  $G(N, A)$  with  $n$  nodes, let  $k$  be a positive integer  $k$  ( $k = 1, 2, 3, \dots$ ) and let  $k$  district points to be chosen on the graph  $G$  to be indicated as the set  $X_k = \{x_1, x_2, \dots, x_{k-1}, x_k\}$ . Then,  $d(X_{k,j})$  indicates the minimum distance between any one of the points  $x_i \in X_k$  and the node  $j$  on  $G$ . That is,

$$d(X_k, j) = \min_{x_i \in X_k} d(x_i, j) \quad (3.7.2.1-1)$$

The  $k$ -medians of the network  $G$  are defined as follows:

Definition: A set of  $k$  points  $X_k^*$  on  $G$  is a set of  $k$ -medians of  $G$ , for every  $X_k \in G$ ,

$$J(X_k^*) \leq J(X_k) \quad (3.7.2.1-2)$$

where:

$$J(X_k) = \sum_{j=1}^n h_j d(X_k, j) \quad (3.7.2.1-3)$$

If the  $k$  points in  $X_k$  are to be the points where  $k$  facilities providing a given service will be located and if  $h_j$ , the demand weight of node  $j$ , is set equal to the fraction of all calls for the service in question that originate from  $j$  (i.e.  $\sum_{j=1}^n h_j, h_j = 1$ ), then finding the  $k$ -medians,  $X_k^*$ , of  $G$

amounts to finding the set of  $k$  locations that minimize the average travel distance to (or from) the facilities by service users. This should be clear from the definition of the function  $J(X_k)$  in (3.7.2.1-3), which is now nothing but an expression for the average travel distance. Also, it should be noted that the implicit assumption in all of the above is that demands originating at any given node  $j$  will be served exclusively by the facility that is closest to  $j$ .

### 3.7.2.2 Solution algorithm

For the purposes of this study, the Single Median Algorithm is applied in order to estimate the median point of a sector's network graph. The steps of the algorithms are:

Step 1: Obtain the minimum distance matrix for the nodes of  $G$ .

Step 2: Multiply the  $j^{\text{th}}$  column of the minimum distance matrix by the demand weight  $h_j$  ( $j = 1, 2, \dots, n$ ) to obtain the matrix  $[h_j \cdot d(i, j)]$ .

Step 3: For each row  $i$  of the  $[h_j \cdot d(i, j)]$  matrix, compute the sum of all the terms in the row. The node that corresponds to the row with the minimum sum of terms is the location for the  $l$ -median.

### 3.7.3 Estimation of the medians of city sectors

In this paragraph, the Single Median Algorithm is implemented in order to estimate the median points of the city sectors of Volos. Due to daily calculation limitations of Google Maps regarding the number of points for which the distance matrix can be estimated, the median points for the network graphs of 2 city sectors are estimated and presented.

The maximum number of points that can be used in Google Maps for the calculation of distance matrix is 46 per day. At the same time, it is considered valuable to calculate the median points of the sectors with the highest volume of crime incidents. For these reasons, it was decided to calculate the median points for city sectors p25 and p38.

Sector p25 is defined by coordinates (39.372560, 22.934496) and (39.369441, 22.938637) and sector p38 by (39.369441, 22.947118) and (39.365999, 22.951458). In sector p25, 34 crime incidents occurred in total from 2010 until 2017 and in sector p38, 42 crime incidents occurred in total from 2010 until 2017.

For the estimation of median points for p25 and p38, the coordinates of the incident points occurred between 2010 and 2017 are determined. These coordinates are used as input for the Java script in order to calculate the distances between all incident points in the network graph of each sector. The distance matrix is used in order to apply the Single Median Algorithm and estimate the median point for each sector.

### 3.7.3.1 Median point of sector p25

Sector p25 is demarcated by coordinates (39.372560, 22.934496) and (39.369441, 22.938637). In order to estimate which of the incident points lay on sector p25, a composition of EXCEL commands to construct the checking criteria is used. From the application of this command, the incident points which fall within sector p25 are determined. In Table 3.7.3.1-1 the incident points with their id and coordinates which lie in sector p25 are presented.

Id	Latitude	Longitude	Id	Latitude	Longitude
132	39.37121	22.937608	1248	39.37213	22.936152
147	39.37169	22.935294	1380	39.37209	22.937949
154	39.37214	22.936621	1610	39.36998	22.937968
372	39.37164	22.936369	1733	39.37113	22.935549
391	39.37209	22.937949	1811	39.36998	22.937968
421	39.37085	22.938195	1854	39.37075	22.935315
432	39.37085	22.938195	1911	39.37075	22.935315
448	39.37075	22.934999	1918	39.37077	22.935315
464	39.3703	22.936045	2027	39.37046	22.93705
609	39.36962	22.937982	2155	39.37213	22.934946
611	39.36962	22.937982	2235	39.37143	22.934895
612	39.36962	22.937982	2239	39.37046	22.93705
613	39.36962	22.937982	2242	39.37049	22.93705



636	39.37153	22.935099	2257	39.37045	22.935746
923	39.37017	22.937882	2305	39.37083	22.935748
1215	39.36986	22.93755	2328	39.37073	22.935093
1239	39.36967	22.934876	2572	39.37153	22.935099

Table 3.7.3.1-1: Crime incidents fall within sector p25

As it is obvious Table 3.7.3.1-1 contains incidents with same coordinates. This indicates that crime incidents occurred on the same address but in different time. For the purposes of median estimation the duplicate coordinates should not be taken into account. The incident points with unique coordinates are 25 in total. The coordinates of these 25 incident points are used as input for the Java code which calculates the distances between all unique incident points within sector p25. From the execution of the code a  $25 \times 25$  distance matrix is produced (and it can be found in the Appendix section 7.5). This distance matrix contains the minimum distances from and to all network points of sector p25 and it is used in order to estimate the median point by applying Single Median Algorithm.

Given the minimum distance matrix, each column of it is multiplied by the demand weight  $h_j$ . The demand weight for the purpose of this study is considered equal to 1 for all incident points. However, it can be equal to other values defined based on the type or level of the particular offence, i.e.  $h_j, \text{burglaries} = 3$ ,  $h_j, \text{car robberies} = 4$  and  $h_j, \text{motorcycle robberies} = 2$ . As  $h_j = 1$ , the values of minimum distance matrix are same. As next step of the algorithm, the sum of each row of the matrix is calculated and the outcomes are presented in Table 3.7.3.1-2.

Row number	Id of incident belongs to row number	Sum of row
1	132	14194
2	147	8551
3	154	18479
4	372	10953
5	391	11571
6	421	9656
7	448	7056
8	464	9739
9	609	11060
10	636	8275
11	923	9715
12	1215	10670
13	1239	17475

Row number	Id of incident belongs to row number	Sum of row
14	1248	9470
15	1610	10365
16	1733	7757
17	1854	7504
18	1918	7504
19	2027	7806
20	2155	9643
21	2235	8338
22	2242	8067
23	2257	<b>6956</b>
24	2305	7463
25	2328	7581

Table 3.7.3.1-2: Sum of each row of minimum matrix after the multiplication of each row with weight  $h_j$

As last step of this algorithm, is to identify the incident point that corresponds to the row with the minimum sum of terms. This point is the single median point and in sector p25 is incident point with id equals to 2257. Therefore, the median point of sector p25 is the point with coordinates (39.3704493, 22.9357456). In Figure 3.7.3.1-1 sector p25 is illustrated with its incident points occurred from 2010 until 2017 and the median point.

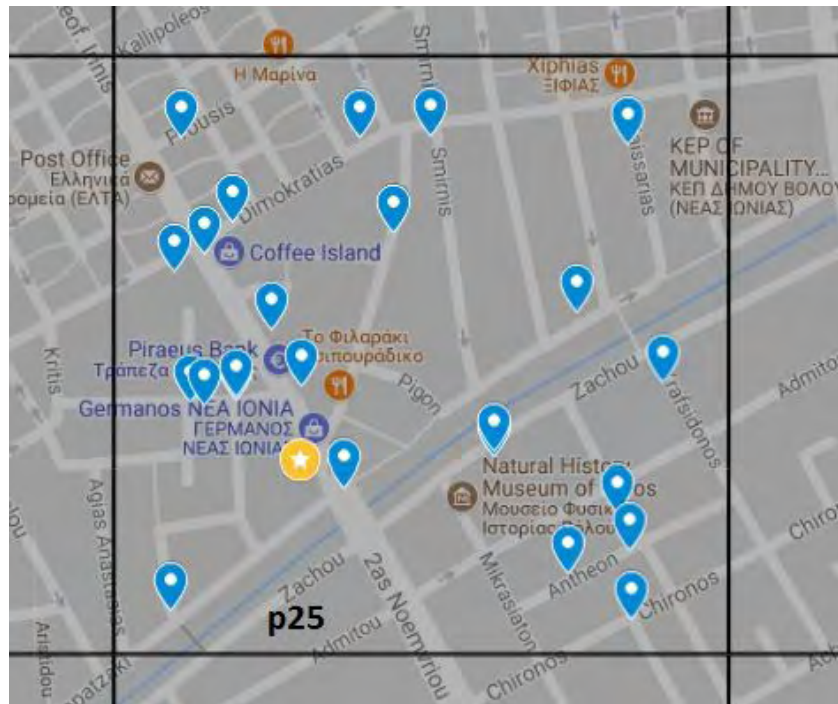


Figure 3.7.3.1-1: Incidents points and median point of sector p25

### 3.7.3.2 Median point of sector p38

The same procedure as the one followed to estimate the median point of network graph for sector p38 is applied for the estimation of the median point of sector p38. The median point of the incident network for sector p38 is located on (39.3671358, 22.9502072). In Figure 3.7.3.2-1 the incident points based on the historical data from 2010 until 2017 and the median point of sector p38 are depicted.

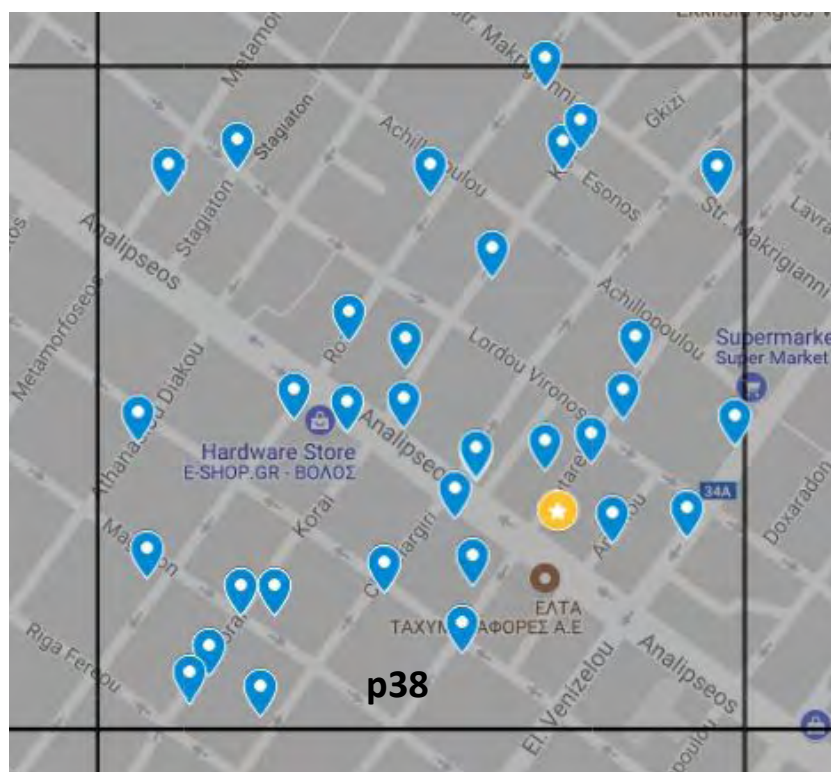


Figure 3.7.3.2-1: Incidents points and median point of sector p38

By applying the same procedure to each city sector and especially to those which have high volume of criminality the median points, which minimize the average distance to or from the incidents occurred in the sector based on the available historical data, can be estimated.

#### 4. CONCLUSIONS

The overarching goal of the work presented in this thesis is to study the data, provided by the Police Department of Volos (PDV), which refer to thefts as criminal offences, in the city of Volos, capital of the district of Magnesia, over the period 2010 - 2017 and to export results about the most common type of offense and the number of offenses that occurred every year, every month, every weekday, every hour and every important time period and day during the year.

For this purpose statistical experiments have been carried out and have been developed comparative charts, so as to compare the searches results and to come to conclusions. For the time period studied, the majority of offenses refer to Burglaries (43%) and Motorcycle Robberies (46%), with the later to be slightly more, while the Car Robberies refer to 11% of the total offenses. A general reduction of offenses was observed with only two years, the year 2011 and 2014, to show an increase over the previous year. From 2015 onwards, we have significant reduction of offenses with 2017 showing a 54.87% decrease over 2010. The highest reduction rate between the years has been observed in 2015. In fact, in 2015, there were 31.46% less offenses than in 2014. The total of offenses does not show any particular difference between the months. In fact the totals show a slight divergence of 1% and 2%. The months of September and November correspond to the lowest total of offenses, while the months in which the most offenses were committed are the months July, August, May, October, January and December.

From the above we infer that offenses are prone to occur during the summer months, where the schools are closed and most of the people leave for summer vacations, and the months that include important days such as Christmas. Statistical experiments related to those important days and specific time periods saw that the day with the display of most offenses is Tuesday. Then follow without much difference the day of Friday sharing the same rate with Wednesday, the day of Monday sharing the same rate with Thursday and last the day of Saturday sharing the same rate with Sunday. The percentage difference between the days is in the range of 1% to 2%. It appears

that the most offenses occurred mainly in the afternoon and evening hours. In particular, 22:00 hours seems to have the highest percentage of offenses occurrence, while 04:00 seems to have the lowest one. Also, the 12:00 turns out to be in the highest risk after the 22:00.

As for the specific weeks of the year, it appears that the most offenses occurred during Easter Week. Easter is usually celebrated in Greece in May or April, according to the Orthodox calendar. Furthermore by examining and comparing the important days of the year it is concluded that the majority of offenses took place in the 1<sup>st</sup> of January, New Year's Day , and then in the 15<sup>th</sup> of August, Celebration of Virgin Mary (a great religious feast for the country).

In regard to the location of incidents, it turned out that most of the incidents occurred nearby the main streets while they spread throughout the city. All the results are expressed in the forms of charts and are presented in the figures below.

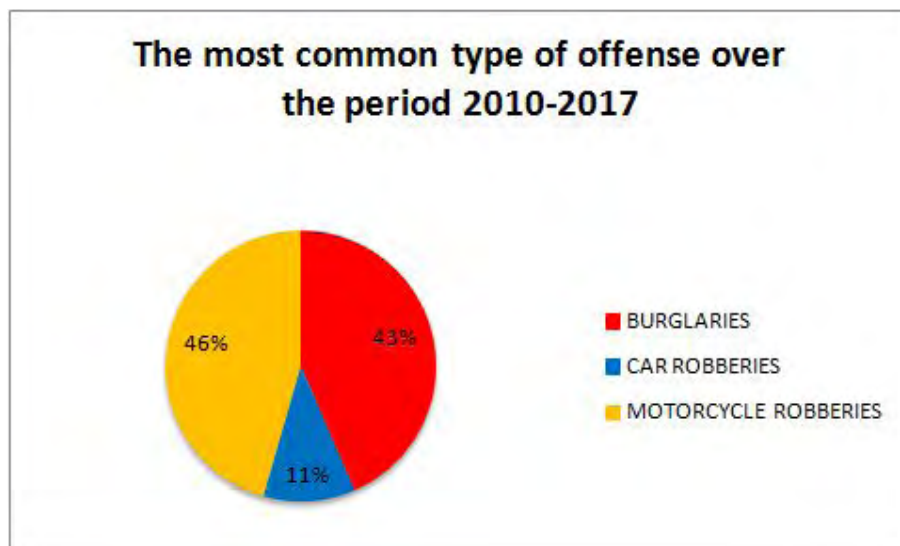


Figure 4.1: The most common type of offense over the period 2010-2017

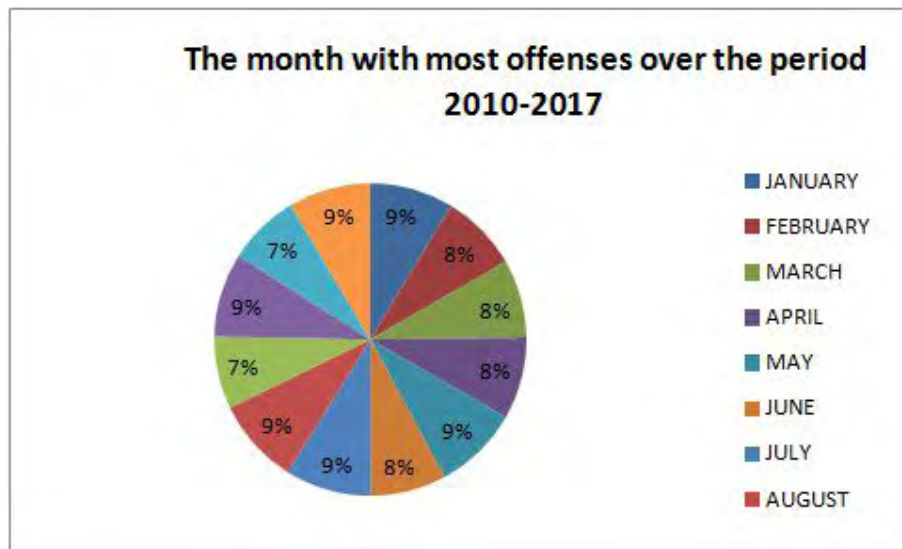


Figure 4.2: The month with most offenses over the period 2010-2017

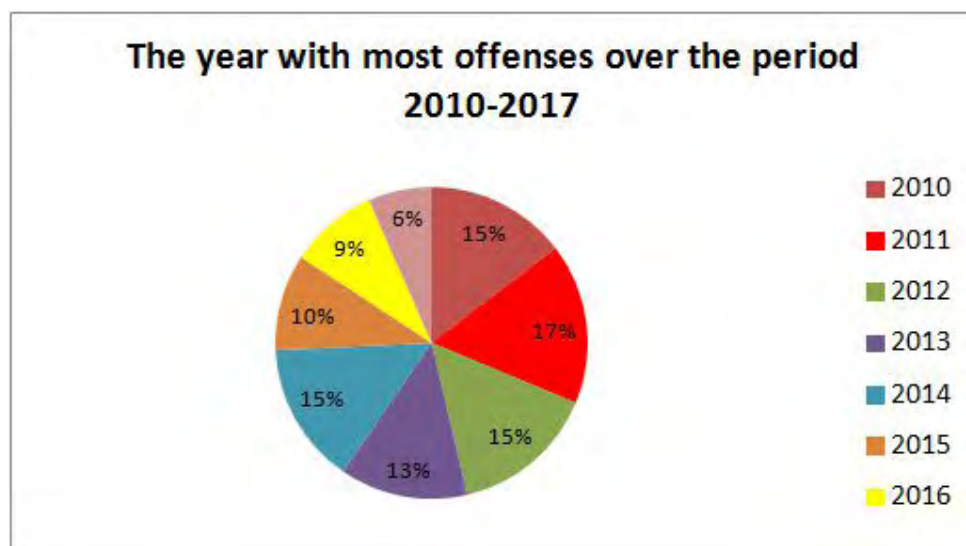


Figure 4.3: The year with most offenses over the period 2010-2017

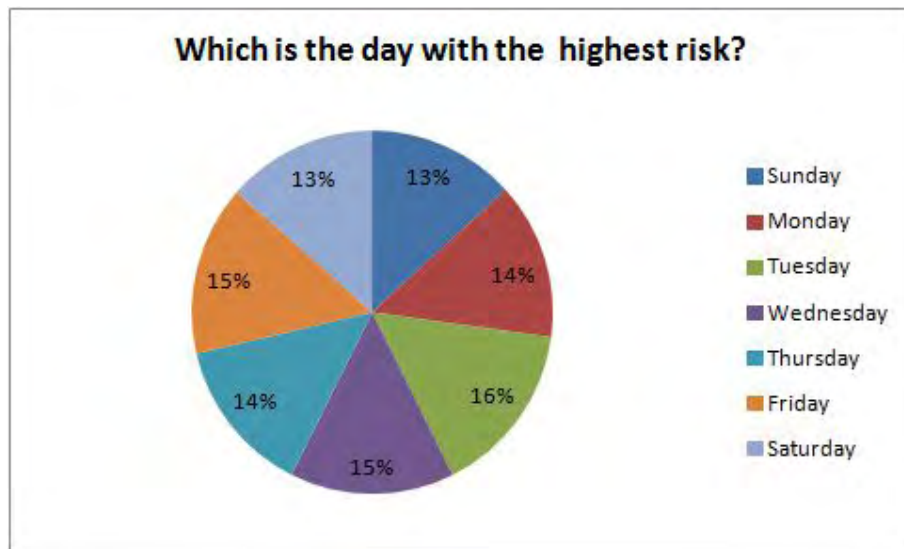


Figure 4.4: The weekday with most offenses over the period 2010-2017

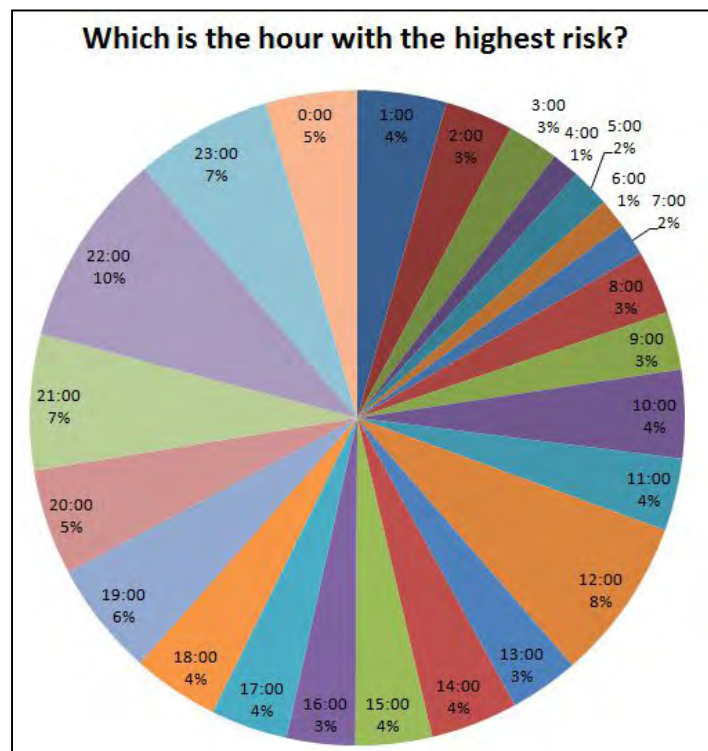


Figure 4.5: The hour with most offenses over the period 2010-2017



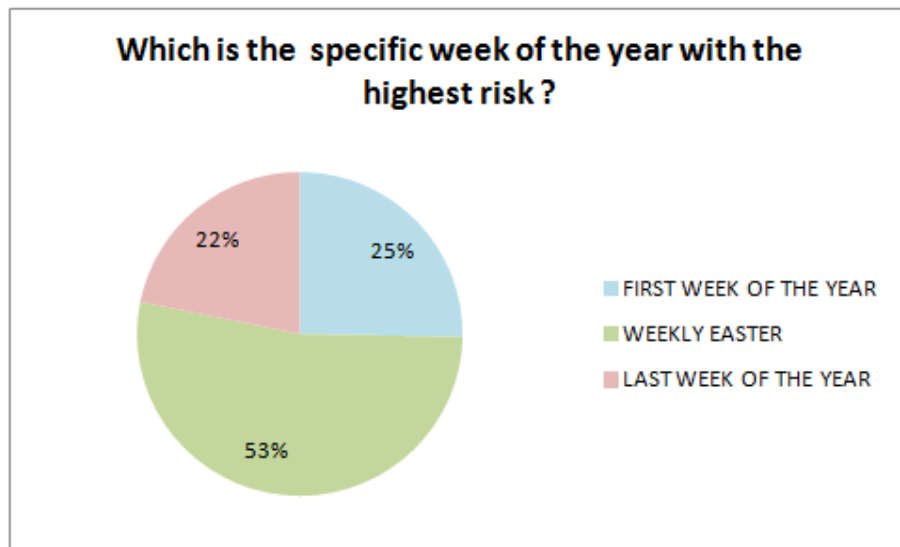


Figure 4.6: The specific week of the year with most offenses over the period 2010-2017

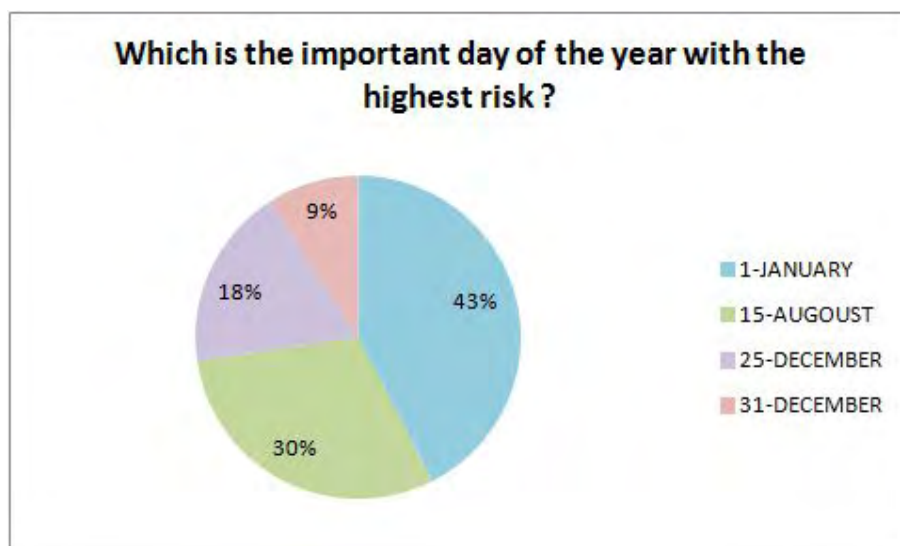


Figure 4.7: The important day of the year with most offenses over the period 2010-2017

It should be noted that the occurrence of offenses is based on many factors, both social and demographic. According to the discussion we had with the police we were informed that in the city of Volos the people related to thefts are known to the police. Therefore, their arrest has the effect of reducing offenses, whereas when they are left free they are again active and contribute to the increase of offenses. It is a fact that there are specific criminal gangs operating throughout the country and that if they are caught, there is a fall in the offenses for the following months. When important police successes are published, then a general fall of offenses is observed as the remaining perpetrators become alarmed and abstain from committing offenses. All have to do with the successes of the police both in the city of Volos and in the city of Athens (capital of Greece), and Thessaloniki as some gangs that act in Athena and Thessaloniki also act in the province. It is also important who is arrested by the police and whether they will impersonate others and testify movements, as well as and the offending punishment which affects the occurrence of offenses for the next period. In particular if the penalty is great then the guilty do not act for as long as they are imprisoned and those who are not arrested yet, abstain from committing offenses for the next period.

The Security Department of Volos, in order to be effective and reduce the offenses applied a predictive model based on the observation of offenses. Specifically if there are about twenty offenses with the same features, that is, a common day and time of action and certain areas and way of acting, then they are attributed to a particular gang. So it is possible for the police to create a map with predictions of the action of this particular gang,

Prediction of crime incidents is essential in order to prevent and decrease future occurrences. Therefore, Splunk Machine Learning Toolkit is used for prediction purposes in this dissertation. The city of Volos was divided into three big sectors indicating the west (Sector 1), the center (Sector 2) and the east (Sector 3) parts of the city. The implementations of the prediction

algorithms, for these 3 Sectors and for every type of offense led to the predictive results which are presented in the below comparative charts.

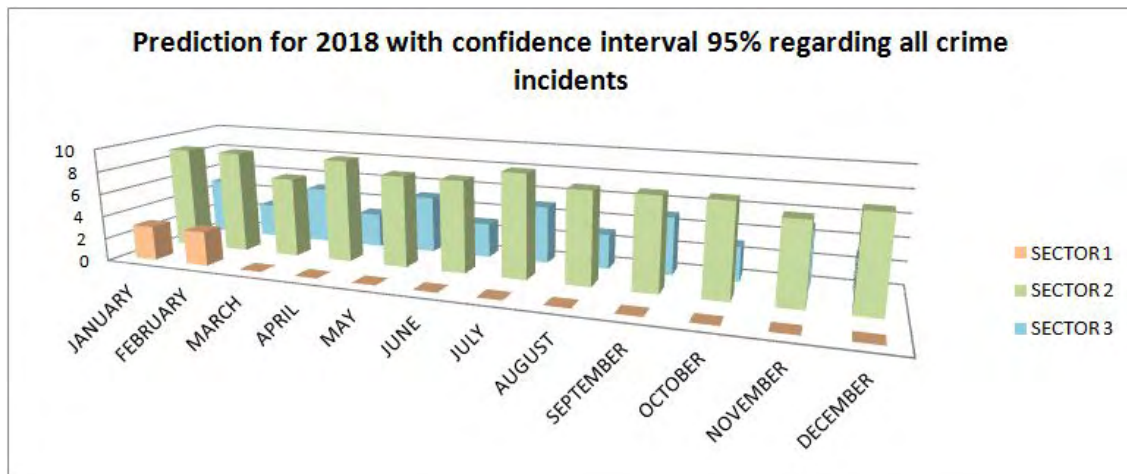


Figure 4.8: Prediction for 2018 with confidence interval 95% regarding all crime incidents

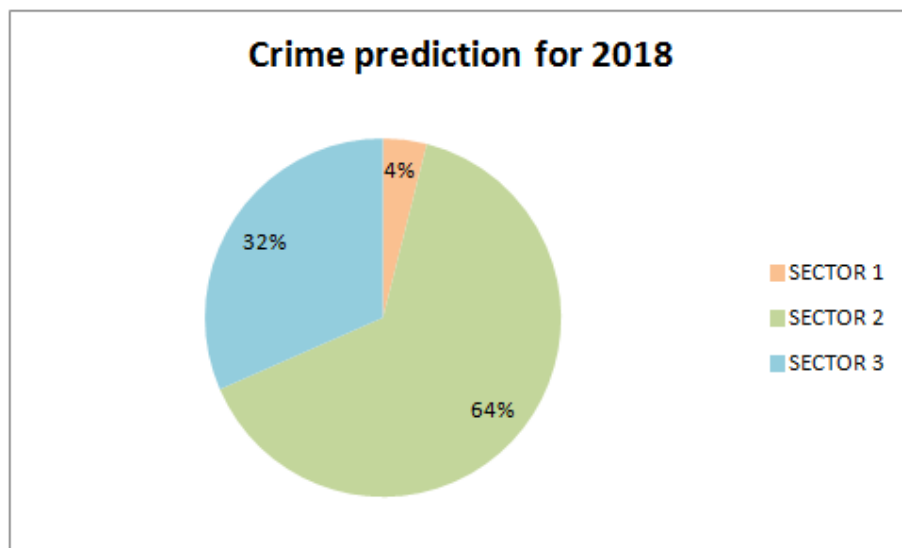


Figure 4.9: Crime prediction for 2018 with confidence interval 95%

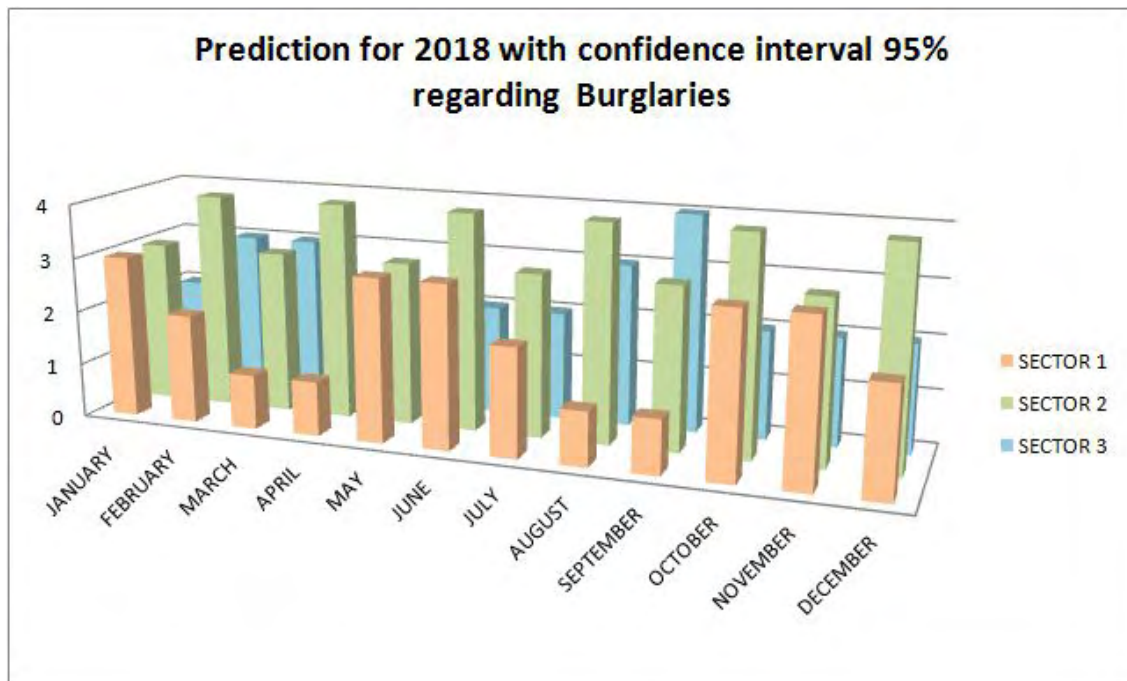


Figure 4.10: Prediction for 2018 with confidence interval 95% regarding Burglaries

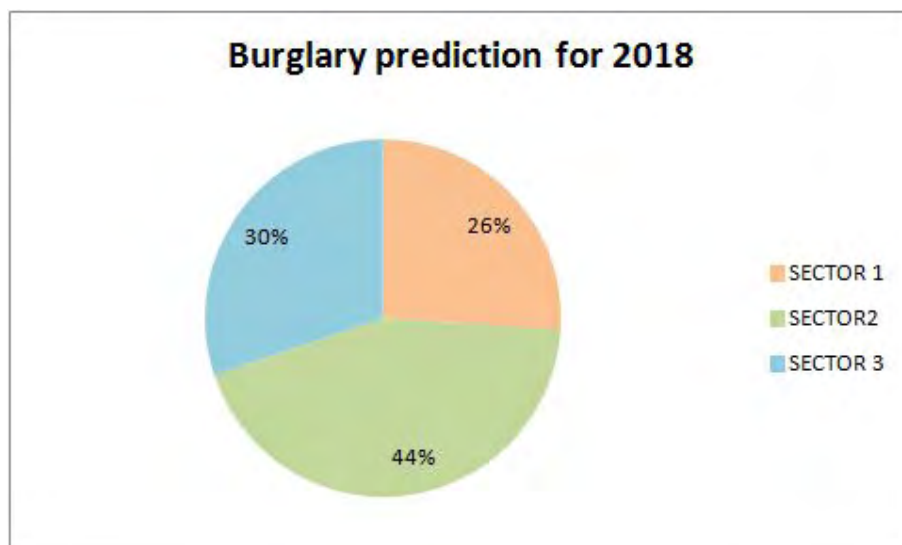


Figure 4.11: Burglary prediction for 2018 with confidence interval 95%

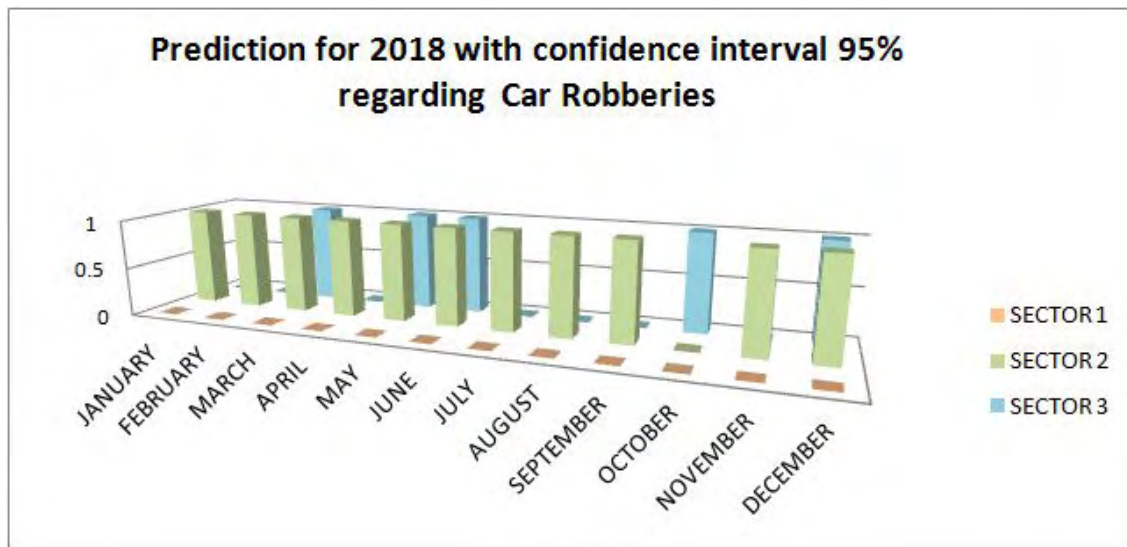


Figure 4.12: Prediction for 2018 with confidence interval 95% regarding Car Robberies

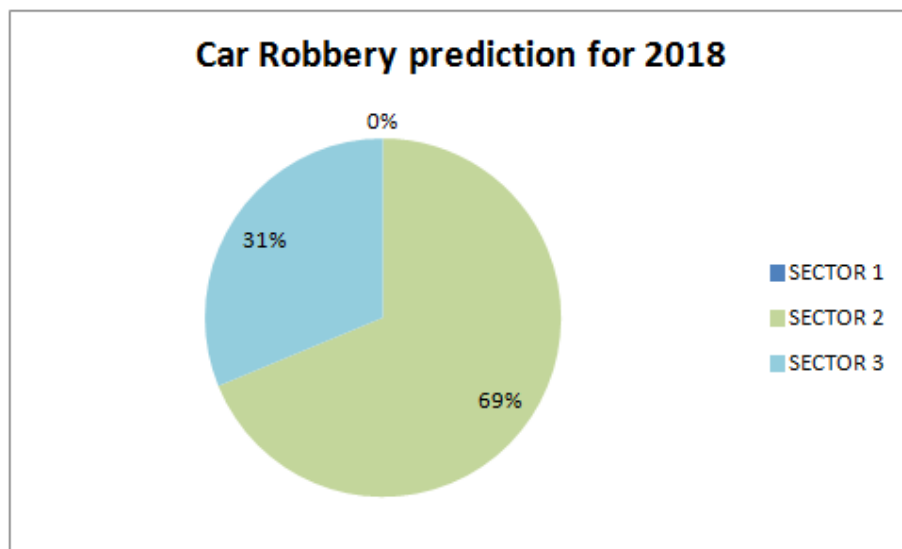


Figure 4.13: Car Robbery prediction for 2018 with confidence interval 95%

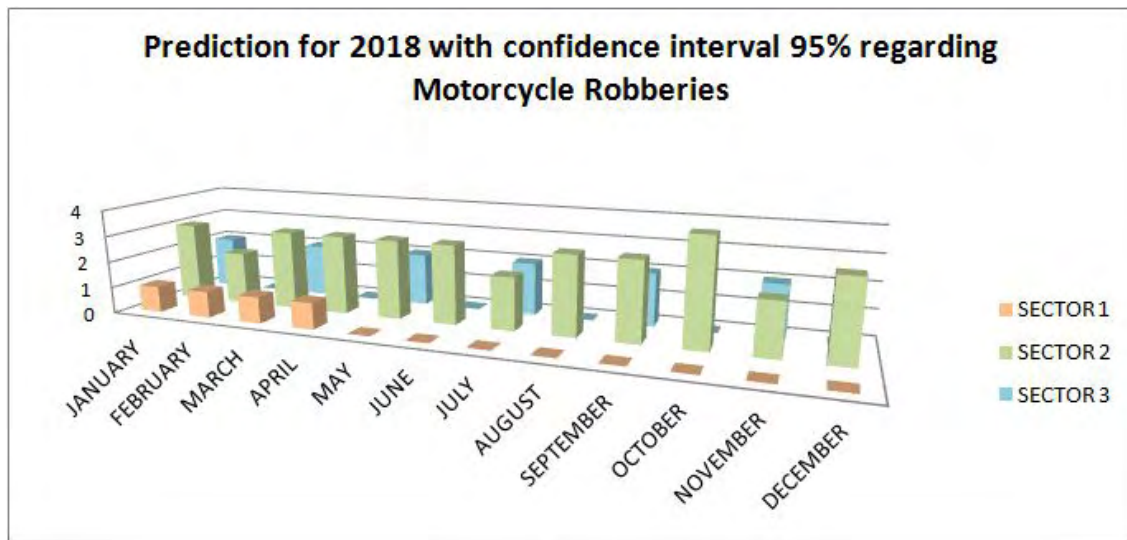


Figure 4.14: Prediction for 2018 with confidence interval 95% regarding Motorcycle Robberies

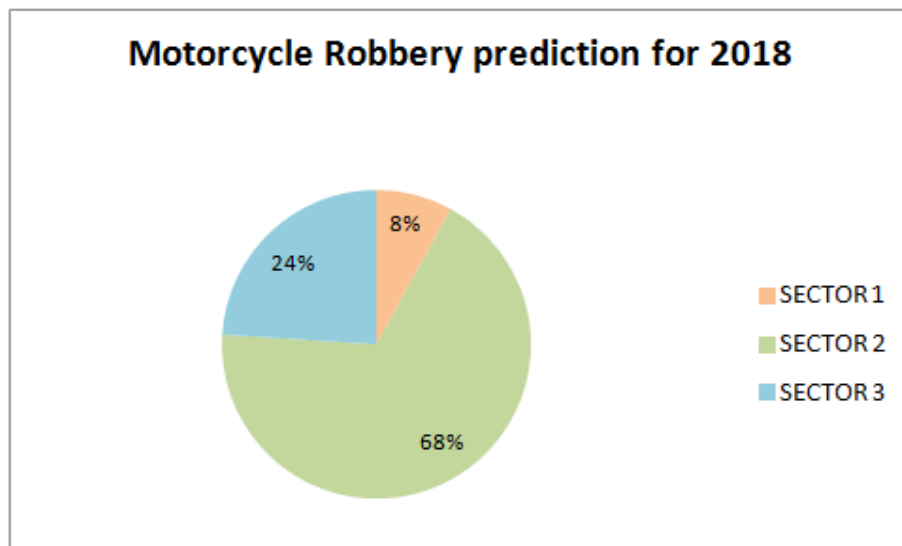


Figure 4.15: Motorcycle Robbery prediction for 2018 with confidence interval 95%

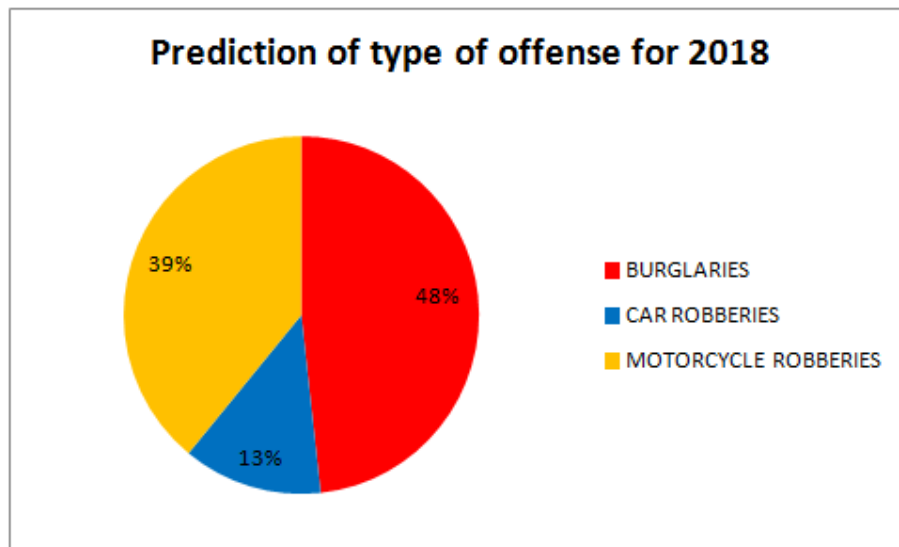


Figure 4.16: Prediction of type of offense for the year 2018

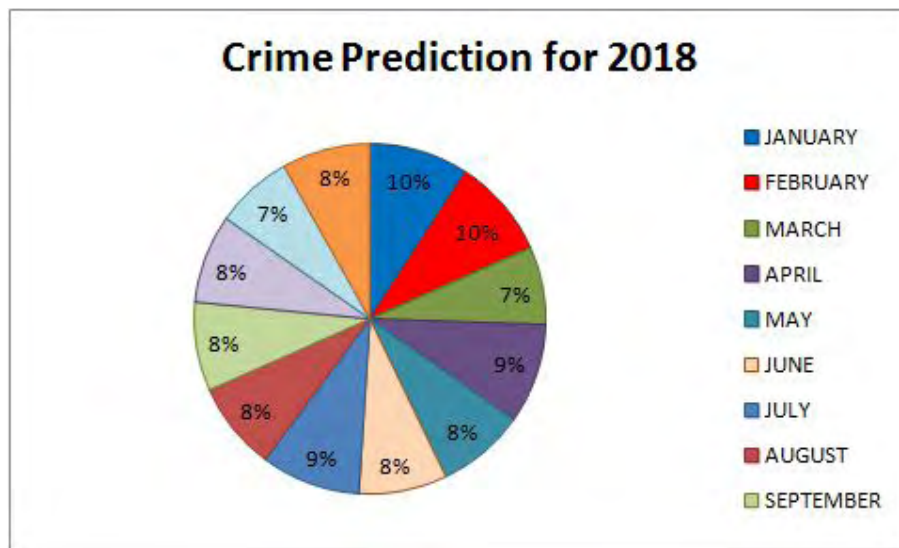


Figure 4.17: Crime Prediction for 2018 over the months

It is concluded from the above charts that the Sector 2, the center part of the city, is the one with the most offenses and that the months where the largest number of offenses will appear in the 2018 are the months of January and February. Also the most common type of offense for the 2018 is expected to be the Burglaries.

Subsequently was defined the Busiest day, the day with the largest number of crime incidents. This day is Saturday 1st February of 2014 and 8 incidents have been occurred. By applying the Absolute Center Algorithm for the busiest day was determined the absolute center which indicates the point where the police car can be parked for stakeout. In fact the police car was located so as to minimize the maximum distance to or from the crime incidents points. The coordinates of the absolute center are: 39.37391, 22.93782.

In order to study how the police cans serve all the incidents occurred in the busiest day was applied the DARP problem and was concluded that the police vehicle can reach all the crime incidents points on time. In Section 3.6 is presented the aggregate table of the DARP results (table 3.6.2).

Furthermore the city of Volos was divided in 68 sectors and the median points for sectors with high criminality were estimated by applying the Single Median Algorithm. The median points are considered valuable for police officers, as they indicate which minimize the average distance to or from the incidents occurred in the sectors. These points can be considered as hot spots across the city and police officers can do patrol around them. Sectors p.25 and p.38 were the ones with the highest volume of crime incidents. In particular in Sector p.25 occurred 34 crime incidents and in Sector p.38 occurred 42 crime incidents. For Sector p.25 the median point is the point with coordinates 39.3704493, 22.9357456 (Leoforos Eirinis 1, N. Ionia) and is the incident point with id equals to 2257. The median point of the incident network for sector p38 is located on 39.3671358, 22.9502072 (Koutarelia 120, Volos).



## **5 FUTURE RESEARCH (ARTIFICIAL INTELLIGENCE)**

This dissertation set out to predict crime incidents in the city of Volos and propose methods to suppress offences effectively. Data analysis, prediction techniques and optimization algorithms were studied and applied for this purpose. However, the contribution of other scientific disciplines could be also useful and enlightening. Therefore, in this final chapter, the directions for future research of crime analysis, forecasting and prevention based on computer science branches, such as artificial intelligence, will be discussed.

Prior to analyzing the potential future contribution of artificial intelligence in the field of crime analysis and predictive policing, it will be sensible to provide detailed information about this area of computer science.

### **5.1 Definition of artificial intelligence**

Artificial intelligence (AI) is the simulation of human intelligence processes by machines, especially computer systems. These processes include learning (the acquisition of information and rules for using the information), reasoning (using the rules to reach approximate or definite conclusions) and self-correction. (TechTarget, 2016)

### **5.2 Types of artificial intelligence**

AI can be categorized in many and different of ways based on researcher's perspective and the area of science that is applied. Two are the most commonly accepted ways to classify AI.

Beginning with a more aggregate classification, an AI system can be classified as either weak AI or strong AI. Weak AI, also known as narrow AI, is an AI system that is designed and trained for a particular task. Virtual personal assistants, such as Apple's Siri, are a form of weak AI. Strong AI, also known as artificial general intelligence, is an AI system with generalized human cognitive abilities so, that when presented with an unfamiliar task, it has enough intelligence to

find a solution. In 1950, mathematician Alan Turing developed the Turing Test, which is a method used to determine if a computer can actually think like a human.

An alternative and more detailed way which also indicates the barriers that separate machines from human beings is the categorization of AI based on four different types. Arend Hintze, an assistant professor of integrative biology and computer science and engineering at Michigan State University, classified AI into four types, from the kind of AI systems that exist today to sentient systems, which do not yet exist. The four types of AI are as follows:

➤ ***Type I AI: Reactive machines***

The most basic types of AI systems are purely reactive. They have the ability neither to form memories nor to use past experiences to inform current decisions. They perceive the world directly and acting on what they see. A well-known example is Deep Blue, the IBM chess program, which beat Garry Kasparov in the 1990s. Deep Blue can identify pieces on the chess board and make predictions, but it has no memory and cannot use past experiences to inform future ones. It analyzes possible moves – its own and its opponent – and chooses the most strategic move only based on current situation.

➤ ***Type II AI: Limited memory***

AI systems of type II can use past experiences to inform future decisions. Some of the decision-making functions in autonomous vehicles have been designed this way. Observation are added to the self-driving cars' preprogrammed representations of the world, which also include lane markings, traffic lights and other important elements, like curves in the road. They are included when the car decides when to change lanes, to avoid cutting off another driver or being hit by a nearby car. However, these observations are only transient. They are not saved as part of the car's library of experience it can learn from.

➤ ***Type III AI: Theory of mind***

This is a psychological term. It refers to the understanding that others have their own beliefs, desires and intentions that impact the decisions they make. This kind of AI does not exist yet. If AI systems are indeed ever to walk among human beings, they will have to be able to understand that each of human beings has thoughts, feelings and expectations for how he will be treated. AI systems will have to adjust their behavior accordingly.

➤ ***Type IV AI: Self-awareness***

This final AI type derives from type III “theory of mind” and implies that AI systems should be able to have a sense of self, have consciousness. Machines with self-awareness understand their current state and can use the information to infer what others are feeling. As an extension of type III “theory of mind”, AI systems with “self-awareness” do not exist yet. While creating machines that are self-aware seems beyond reach, the important step is to understand memory, learning and the ability to base decisions on past experiences, all these parts that constitute human intelligence. (Arend Hintze, 2016)

### **5.3 Historical review of artificial intelligence**

AI has its origins in the ancient Greece. Many myths in antiquity, such as the Hephaestus, the blacksmith who manufactured mechanical servants, and the bronze man Talos incorporate the idea of human-like artifacts and intelligent robots. Many mechanical toys and models were actually constructed, such as the steam-powered pigeon by Archytas of Tarentum. However, the earliest research into thinking machines began in the 1940s and 50s when a handful of scientists from various fields, such as mathematics, psychology, engineering, economics and political science began to discuss the possibility of creating an artificial brain.

It was summer of 1956 at a conference of Dartmouth College where John McCarthy introduced the term of AI and the discipline was born. Those who attended would become the

leaders of AI research for the next decades. Many of them predicted that a machine as intelligent as a human being would exist in no more than a generation and lot of money invested to make this vision come true.

Eventually, it became obvious that they had underestimated the difficulty of the project due to computer hardware limitations. In 1973, the U.S. and British governments stopped funding research into AI, as a response to the criticism of James Lighthill and ongoing pressure of congress. Lighthill criticized in his report (known as Lighthill report) the failure of AI to achieve its grandiose objectives. He specifically mentioned the problem of “combinatorial explosion” or “intractability” of the discourse universe, which implied that many of AI’s most successful algorithms would grind to a halt faced with real world problems and were only suitable for solving toy versions.

The following years were difficult and are known as an "AI winter". Seven years later, a visionary initiative by the Japanese Government inspired governments and industry to provide AI with billions of dollars. However, by the late 80s the investors became disillusioned by the absence of the needed computer hardware and withdrew funding again.

Investment and interest in AI boomed in the first decades of the 21st century, when machine learning was successfully applied to many problems in academia and industry thank to the presence of powerful computer hardware. As in previous "AI summers", some observers, such as Ray Kurzweil, predicted the imminent arrival of artificial general intelligence: a machine with intellectual capabilities that exceed the abilities of human beings. (M. Tim Jones, 2008)

#### **5.4 The connection between artificial intelligence, machine learning and data science**

As it is stated above, AI aims at the infusion of intelligence to machines. This includes visual perception, understanding human text and speech, responding to human beings in a more natural way and, also self-correction. In order to acquire these attributes of intelligence, machine learning

(ML) helps the AI systems to learn and improve their performance, just like how humans do through the process of learning and assessing their mistakes.

For this purpose, computer science tools and statistical analysis of large datasets are required. Figure 5.4-1 illustrates the connection between AI, ML and data science. Data science is the cornerstone; ML is the core and the combination of these two disciplines lead to AI implementation.

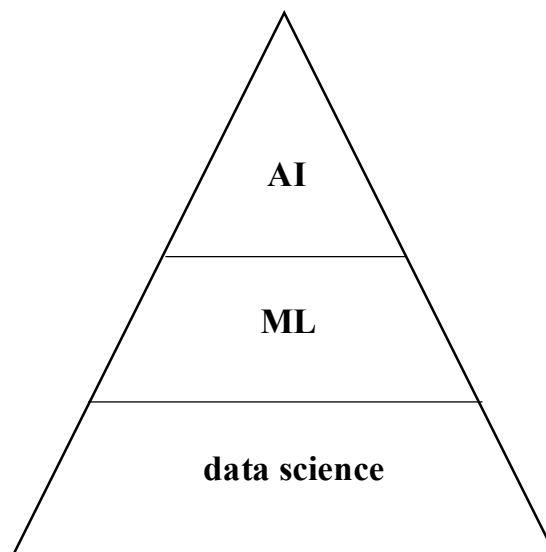


Figure 5.4-1: Connection between AI, ML and data science

ML can provide not only recognition but also prediction of new patterns and, possibility to improve performance with feedback. Given a training sample of  $n$  observations on a class variable 'Y' that takes values  $(1, 2, \dots, k)$  and  $p$  predictor variables  $(X_1, \dots, X_p)$ , ML aims to find a model for predicting the values of 'Y' from new 'X' values.

Among the large range of ML methods applied for the purpose of imitating the human logic, it is worth citing neural networks, Gaussian mixture models, classification tree and random forest. Neural networks are based on modeling the neurons and feeding the network a set of training data

to find patterns. A Gaussian mixture model is a probabilistic model that assumes all the data points are generated from a mixture of a finite number of Gaussian distributions with unknown parameters. Classification tree consists in modeling by recursively partitioning the data space and fitting a simple prediction model within each partition. As a result, the partitioning can be represented graphically as a decision tree. Based on the classification trees theory, random forests try to classify a new object from an input vector, put the input vector down each of the trees in the forest. Each tree gives a classification, and we say the tree ‘votes’ for that class. The forest chooses the classification having the most votes over all the trees in the forest. (Breiman, 2001)

The machine learning theory is based on four different phases, in order to provide accurate outcomes. These four phases are: feature extraction, training, test and evaluation.

Feature extraction is a type of dimensionality reduction that efficiently represents interesting parts of a query object as a feature vector. In the training phase, the features of an object are stored as reference features to generate numerical templates for future comparisons. The numbers of reference templates, which are required for efficient recognition, depend upon the kind of features or techniques that the system uses for object recognition. In the recognition phase, features similar to the ones that are used in the reference template are extracted from an input object whose identity is required to be determined. The recognition decision depends upon the computed distance between the reference template and the template devised from the input utterance. Therefore, in order to achieve better pattern recognition, big training data sets are required. In Figure 5.4-2, system based on a phase of enrolment of data to create model, a phase of pattern recognition and a phase of decision is depicted.

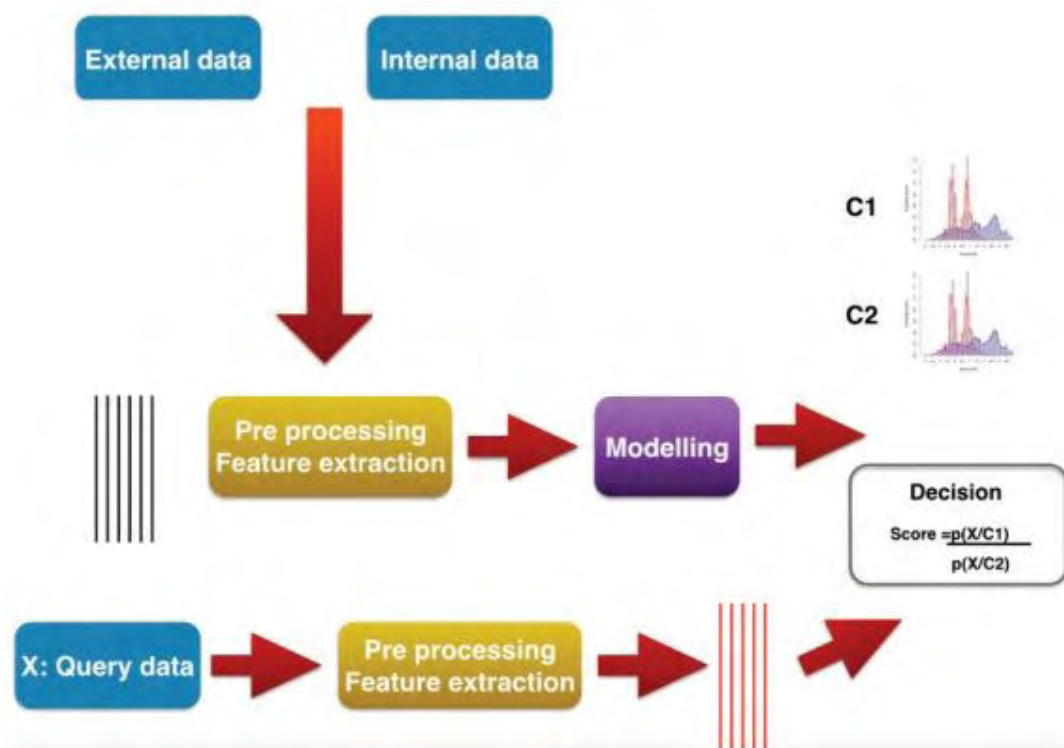


Figure 5.4-2: System principle

The level of performance of a system is quantified most typically by a receiver operating characteristic, called ‘ROC curve’ or a detection error tradeoff curve called ‘DET curve’. This curve reveals the compromise between the ‘false acceptance rate’ and the ‘false rejection rate’. The false acceptance rate is the frequency with which query data from different sources are erroneously assessed to be from the same source. The false non-match rate is the frequency with which query data from the same source are erroneously assessed to be from different sources. The performance of a system falls on a point on the ROC curve whose location is a function of the matching ‘threshold’ applied. A higher match threshold reduces false acceptance rate and increases false rejection rate. On the contrary, a lower match threshold reduces the false rejection rate but increases false acceptance rate.

The decision step is a binary hypothesis testing problem expressed by:

$H_0$  : impostor object

$H_1$  : client (real) object

The  $P_{fa}=P(x=1|H_0)$  is the probability of false acceptance and  $P_{fr}=P(x=1|H_1)$  is the probability of false rejection.

According to the Bayesian theory, these two kinds of errors are weighted by costs and summed into a single cost function, the Bayesian risk function:

$$B_{risk} = P(H_0) \times C_{fa} \times P_{fa} + P(H_1) \times C_{fr} \times P_{fr} \quad (5.4.1)$$

with  $P(H_0)$ : probability of an impostor object,  $P(H_1)$ : probability of a genuine object,  $C_{fa}$  cost of false acceptance and  $C_{fr}$ , cost of false rejection.

$P(x|H_0)$  is compared to a threshold that divides the decision region between a region of acceptance and a region of rejection. If an object's matching score happens to fall above the thresholds, it is considered as genuine, if it is below as imposter.

One of the main advantages of an artificial intelligence system is its capacity to continuously learn any kind of information increasing the efficiency of a decision. Such a system takes into account different views, different perspectives and thus is proposing a more complete analysis. However, it is important to consider that AI is first an empirical science. AI follows a hypothesis-and-test research paradigm. The performance of these systems are very linked to the databases and to the algorithm used.

AI has developed rapidly over the past few years and computers can now outperform humans in some cases. This is the case in the field of object recognition, face recognition, facial expression, speaker recognition and even emotion identification.

## 5.5 Contribution of artificial intelligence in the field of predictive policing

The first step before the development of AI could be the predictive analysis development. Spatial and temporal methods appear as a very good opportunity to model criminal acts. Common sense reasoning about time and space is fundamental to understand crime activities and to predict some new occurrences. The principle is to take advantage of the past acknowledgment to



understand the present and explore the future. Based on multiple methods including exponential smoothing algorithms (simple, double, triple), autoregressive integrated moving average techniques and neural networks, a prediction analysis is carried out on different offences. These models are fitted to offence time series data either to better understand the data or to predict future points in the series. Results are derived from two different sets of past data. The first one is used to train the algorithm and build the model and the second one is used to evaluate the performance of the model. In this study, data from 2010 to 2017 constitutes the training and evaluation sets. Based on the prediction model, future evolutions are proposed for 2018 per sector and per sector and type of offence. Based on the prediction outcomes, it is possible to optimize the allocation of resources during some specific periods but also to evaluate the performance of some modus operandi used in the past.

The main interest is to identify trends, patterns, or relationships among data, which can then be used to develop a predictive model and propose short, medium and long-term trends (Hoaglin et al., 1985) in order to inform police service at different levels. A map visualisation is very interesting and relevant to anticipate crime or criminal moving by evaluating the places of concentration or the dispersion movement. For instance, based on a regression model it is possible to explain crimes like burglaries from social and economic data like urban development and population growth.

Because crime is neither a random nor a deterministic process, some features exist to characterize crime and perhaps offenders or police officers. Based on this assumption, it is possible to mechanize some tasks and upgrade predictive policing by applying AI techniques on the question of crime and to get benefits from machine intelligence.

In this direction many global corporates, such as the GAFAM (Google, Apple, Facebook, Amazon, Microsoft) and the NATU (Netflix, Air BNB, Telsa, Uber), collect more and more data every day and have a real capacity of analysis and produce objective results. In 2015, Bill Gates

expressed that AI is entering a period of rapid advances. AI will fundamentally change how humans move, communicate and live. Today personal data and private information cannot be fully controlled. GAFAM are at the forefront of innovation in artificial intelligence, with active research exploring virtually all aspects of machine learning, including deep learning and more classical algorithms. They gather large volumes of direct or indirect evidence of relationships of interest, applying learning algorithms to understand and generalize. In the field of crime analysis it is easy to imagine some concrete applications:

- To recognise a known criminal in a specific area and send an email on a personal smartphone
- To identify geographical and time hotspot areas of crime
- To make a profile of criminal based on massive data
- To indicate the level of multiple offences in a specific area, and
- Why not to replace a police officer by a virtual agent in specific tasks

It could be very exciting to visualise on one's own smartphone a risk of theft or aggression on a specific area or to get an estimation of the number of pickpockets around us. It might provide the citizen with a sensation of control of his or her own security — but isn't it an illusion? Indeed, all these applications cause a risk for privacy and for the power to decide. In addition, without any control of the data, it will be very difficult to evaluate the reliability of the information. The risk is real and the best way to protect people from abuses and to avoid a police driven by ROI (Return On Investment), is to allow and rely on the development of AI applications by law enforcement.

In an era of accountability, law enforcement cannot rest on past accomplishments against crime for very long. Law enforcement must go one step further but step by step by rigorously respecting privacy. It is more and more important to be flexible and creative to face a criminality in constant and quick evolution. This new vision for future is a great and exciting challenge. Using computer to analyze how offenders react to questions combined with the ability to identify

what sort of people they are, could also provide new opportunities to help investigators.

An expert system is able to autonomously learn crime activities and behaviors which otherwise would be masked in a global environment. From a theoretical point of view, AI can be used in three different cases:

- To model criminal acts
- To model behaviour and criminal way of reasoning
- To model behaviour and investigator way of reasoning.

The objective is to extract knowledge from these three sources and why not from a fusion of these sources.

A possible use of AI is to model specific profiles of criminals. The principle of this approach consists in evaluating the possibilities that a suspect relates to an a priori class. The advantage of an AI is to train the model from criminological theory and from real case reports. These kinds of applications could be realized to build a class model for specific criminals but also for victims. Analysis of the patterns formed by prolific offenders could be built on many elements, like their movements, their area of living, advertising or working, their habits, their type of crime, their previous convictions, their home, their daily activities, their social networks, the offence locations, etc.

Based on the same principle and in the case of financial crimes linked to sensible companies, a profile based on social engineering could define an evaluation risk of attacks. A first condition is to get a history of past cases in order to build a dataset of victims and another one of non-victims. The competitive hypothesis developed in the decision process is:

$H_0$  : non-victim company

versus

$H_1$  : victim company

and the risk for a query company to be victim is calculated by (5.4.1).

This kind of analysis starts to be used in order to find the most probable possibility. In future, it could also be possible to model the behavior and the investigator's reasoning. In many cases, the investigation process is a logical enterprise in a logical environment, formed by the legal procedure. In addition, an investigator uses his own experience to increase his relevance. These different aspects can be modeled by an expert system based on the principle of training. One source of the investigator reasoning is the results of interviews carried out with the agents that could be used to train a model. A police officer on patrol most likely uses deductive reasoning and learns everything by experience (Bosio, 2011). So, the challenge for an AI system is to be able to incorporate experience and a way of reasoning. Yet, knowledge and intuition of the police officer play a central role. All the process is not logical and an investigator in front of a situation needs to keep an open and adaptive attitude. Technical and logical knowledge, although necessary, is not sufficient to account for the global process of investigation.

Currently, a virtual agent cannot be able to provide objective help to a real investigator, because of the heterogeneity and the complexity of the situation that is not uniquely logical. Formal or informal perception plays an important role in the grip on reality for investigators. Understanding criminal investigations also requires inferring a hidden factor, namely, the intention of the police officer. But the extension of AI in this field is based on an analysis of police officer patterns. Analyses could for instance include experience, age of investigators, trajectories, modus operandi of investigations, crime type, and so on.

In conclusion machines are learning to see in increasingly reliable and useful ways, opening up a wide range of opportunities and perspectives for law enforcement. AI can increase the capacity to receive real-time alerts of abnormal behavior and quickly respond to time-sensitive and critical events. Indecision and delays are the parents of failure, the aim is to upgrade human decision-making thanks to AI. The risk is to see these perspectives developed by private societies or industrial groups instead of law enforcement.

## 6. BIBLIOGRAPHY

### 6.1 Greek Bibliography

- Government Gazette, Presidential Decree No 7 Articles 95 and 97 first sheet number 14, February 2017

### 6.2 Foreign Bibliography

- Agrawal, R., Imielinski, J. & Swami A. (1993). *Mining Association rule between sets of items in large databases*, Proceedings of the ACM SIGMOD International Conference of Management of Data, New York: Association for computer machinery, 207-216.
- Chamikara M.A.P., Yapa, Y.P.R.D., Kodituwakku, S.R. & Gunathilake, J. (2012). SLSecureNet: *Intelligent Policing Using Data Mining Techniques*. International Journal of Soft Computing and Engineering (IJSCE),2(1), 175-180.
- Chen, H., Chung, W., Xu, J. J., Wang, G., Qin, Y. & Chau, M. (2004). *Crime data mining: a general framework and some examples*. Computer, 37(4), 50-56.
- Giles Oatley & Brian Ewart (2011). *Data mining and knowledge discovery*. Wiley Interdisciplinary Reviews, 1(2), 147-153.
- Koperski, K. & Han, J. (1995). *Discovery of spatial association rules in geographic information databases*, Proceeding of the 4th International Symposium on Spatial Databases, Advances in Spatial Databases, 47-67.
- International Association of Crime Analysis (IACA) (2014). *Definition and types of crime analysis*. Standard Methods,& Technology (SMT) Committee. White paper 2014-02
- **Srikanta Mishra & Akhil Datta-Gupta (2017).** *Applied Statistical Modeling and Data Analytics*, 1<sup>st</sup> ed. A Practical Guide for the Petroleum Geosciences
- Kuchipudi Sravanthi et al, (2015) *Applications of Big data in Various Fields*.

International Journal of Computer Science and Information Technologies (IJCSIT),  
Vol. 6 (5) ,4629-4632

- Chainey Spencer & J. H. Ratcliffe. 2005. *GIS and crime mapping*. Hoboken, NJ: John Wiley.
- Migrant & Seasonal Headstart Technical Assistance Center (2006) *Introduction to data analysis handbook*. Academy for educational development contract with DHHS/ACF/OHS/Migrant and Seasonal Program Branch
- Manfred M. Fischer & Jinfeng Wang (2011). *Spatial data analysis models, methods and techniques*. Springer Heidelberg Dordrecht London New York.
- Shannon E. Reid & George Tita (2017). *The mapping and spatial analysis of crime*
- D. Fisher, R. DeLine, M. Czerwinski, and S. Drucker. *Interactions with big data analytics*. Interactions, vol. 19 (3), 50–59
- Laney D. (2001) *3D Data Management: Controlling Data Volume, Velocity, and Variety*, Technical Report
- Asma Hassani & Sonia Ayachi Gahnouchi (2017). *A framework for Business Process Data Management based on Big Data Approach*. Procedia Computer Science 121, 740-747
- Vashisht P, & Gupta V. (2015) *Big data analytics techniques: A survey*. International Conference In Green Computing and Internet of Things (ICGCIoT), .264-269.
- Gandomi A. & Haider M. (2015). *Beyond the hype: Big data concepts, methods, and analytics*. International Journal of Information Management, vol 35(2),137-144.
- Anselin L.(2009) *Spatial regression*. In: Fotheringham AS, Rogerson PA (eds) The SAGE handbook of spatial analysis. SAGE, Los Angeles, 255–275

- Anselin L. & Rey S.J. (2009). *Perspectives on Spatial Data Analysis*. Berlin: springer -verlog
- Mizoram University (20015). *Research Methodology and Data Analysis*. (NME-ICT) Vanlalhiviati & Sign
- Piquero Alex R., & Weisburd David (2010). *Handbook of Quantitative Criminology*. 1<sup>st</sup> ed. Springer-Verlag New York
- Di Ciaccio Agostino, Coli Mauro & Angulo Ibanez Jose Miguel (2010). *Advanced Statistical Methods for the Analysis of Large Data-Sets*. 1<sup>st</sup> ed. Springer-Verlag Berlin Heidelberg
- Brotcorne, L., Laporte, G., & Semet, F. (2003). *Ambulance location and relocation models*. European Journal of Operational Research, 147, 451–468.
- Jean-François Cordeau and Gilbert Laporte (2007). *The dial-a-ride problem: models and algorithms*. Ann Oper Res:153, 29–46
- Molenbruch, Y., Braekers, K. & Caris, A. (2017). *Typology and literature review for dial-a-ride problems* Ann Oper Res 259: 295-325
- Bartolini Enrico, Algorithms for Network Design and Routing Problems, Marzo 2009
- [http://web.mit.edu/urban\\_or\\_book/www/book/chapter6/6.4.html](http://web.mit.edu/urban_or_book/www/book/chapter6/6.4.html)
- Hoffman Karla L., Padberg Manfred, Rinaldi Giovanni, Traveling Salesman Problem, Kluwer Academic Publishers 2001
- Cordeau Jean-Francois, Laporte Gilbert, Salvesbergh Martin W.P., Vigo Daniele, Vehicle Routing, Elsevier B.V. 2007
- Laporte Gilbert, The Vehicle Routing Problem: An overview of exact and approximate algorithms, European Journal of Operational Research 1992
- Caric Tonci, Gold Hrvoje, Vehicle Routing Problem, In-Tech September 2008

### 6.3 Electronic Bibliography

- <http://www.justiceacademy.org/iShare/LibraryCrimeAnalysis/historyofcrimeanalysis.pdf> (2011)
- <http://www.oxfordbibliographies.com/view/document/obo-9780195396607/obo-9780195396607-0123.xml>
- <https://searchbusinessanalytics.techtarget.com/definition/predictive-analytics>
- <http://fhyzics.com/predictive-analytics-process-and-its-applications.html>
- <http://www.fico.com/en/predictive-analytics/understanding-predictive-analytics/what-are-the-main-types-of-predictive-analytics>
- <https://www.redpixie.com/blog/predictive-modelling-techniques-infographic>



## 7. APPENDIX

### 7.1 Calculation of the Rates of Reduction / Increase of offenses for a specific year compared to the previous year

Table 7.1.1: Aggregate table of offenses over the period 2010-2017

YEAR	2010	2011	2012	2013	2014	2015	2016	2017	TOTAL
BURGLARIES	149	175	205	174	135	107	125	92	1162
CAR ROBBERIES	60	73	42	32	18	29	26	13	293
MOTORCYCLE ROBBERIES	181	197	155	149	238	132	93	71	1216
TOTAL	390	445	402	355	391	268	244	176	2671

The rates of reduction / increase of offenses for a specific year, in relation to the previous year were produced from the above table as follows:

2010 → 390 offenses were occurred

2011 → 445 offenses were occurred

Difference between the two years:  $445 - 390 = 55$  (more offenses in the 2011 compared to 2010)

#### ***Rate of Increase for 2011:***

$$\frac{55(\text{more offenses in the 2011 compared to 2010}) * 100}{390(\text{total of offenses in 2010})} = 14.1\%$$

The rates of reduction / increase for the rest years (2012,2013,2014,2015,2016,2017) were similarly calculated and presented in the table 3.3.2 of Chapter 3.

## 7.2 Charts for the type of the offenses for every month over the period 2010-2017

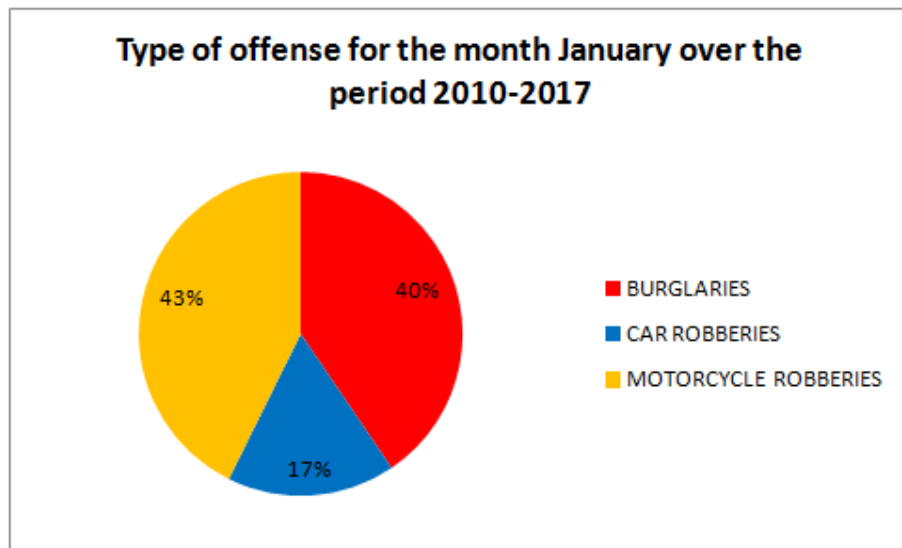


Figure 7.2.1: Type of offense for the month January over the period 2010-2017

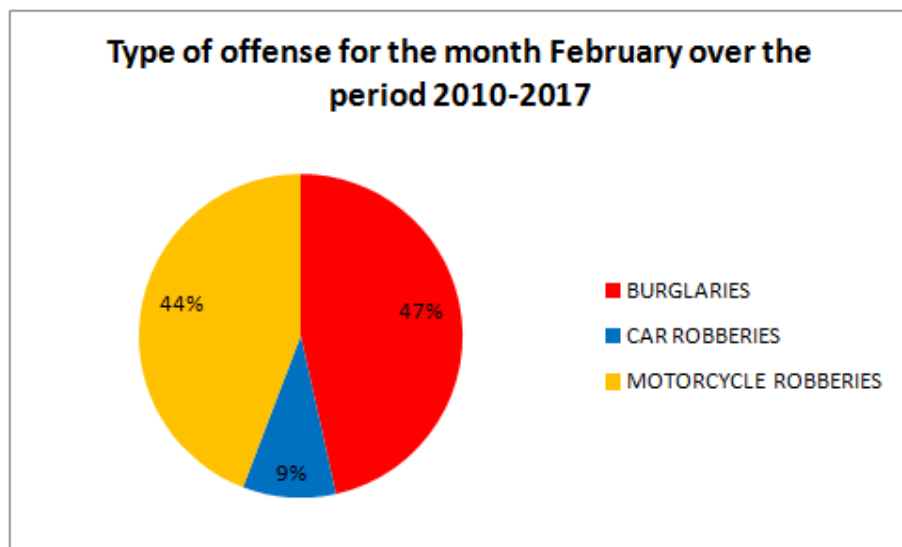


Figure 7.2.2: Type of offense for the month February over the period 2010-2017

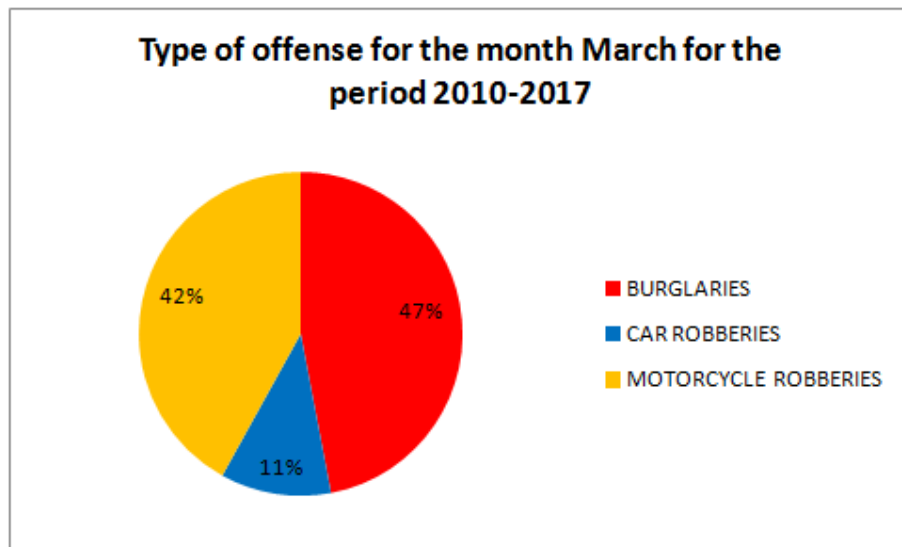


Figure 7.2.3: Type of offense for the month March over the period 2010-2017

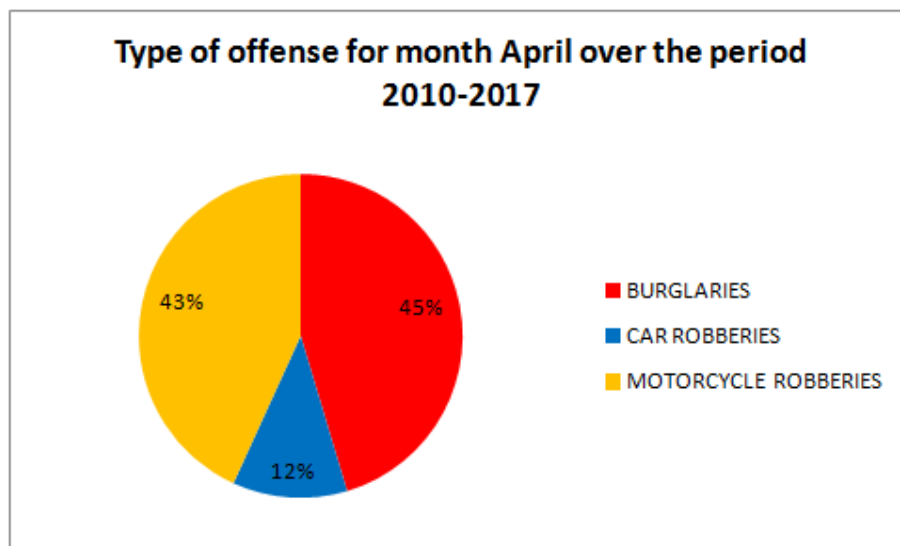


Figure 7.2.4: Type of offense for the month April over the period 2010-2017

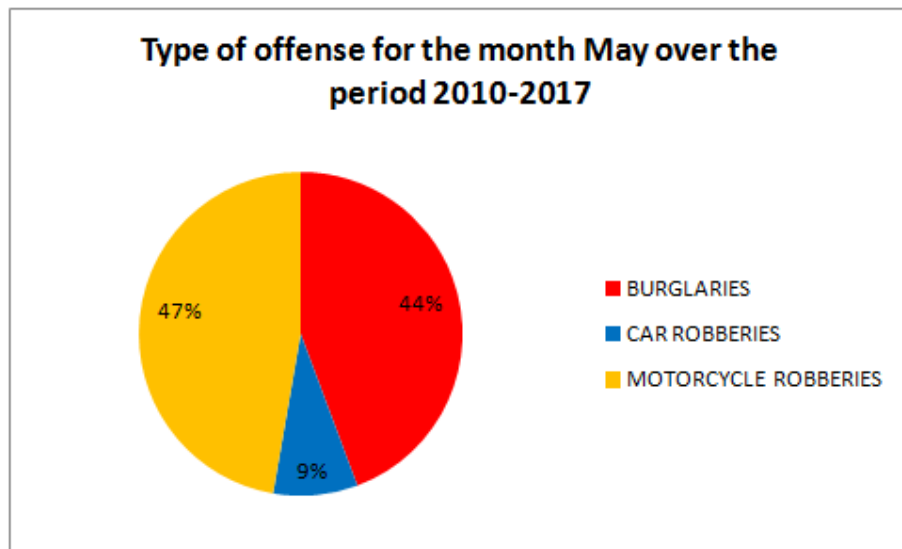


Figure 7.2.5: Type of offense for the month May over the period 2010-2017

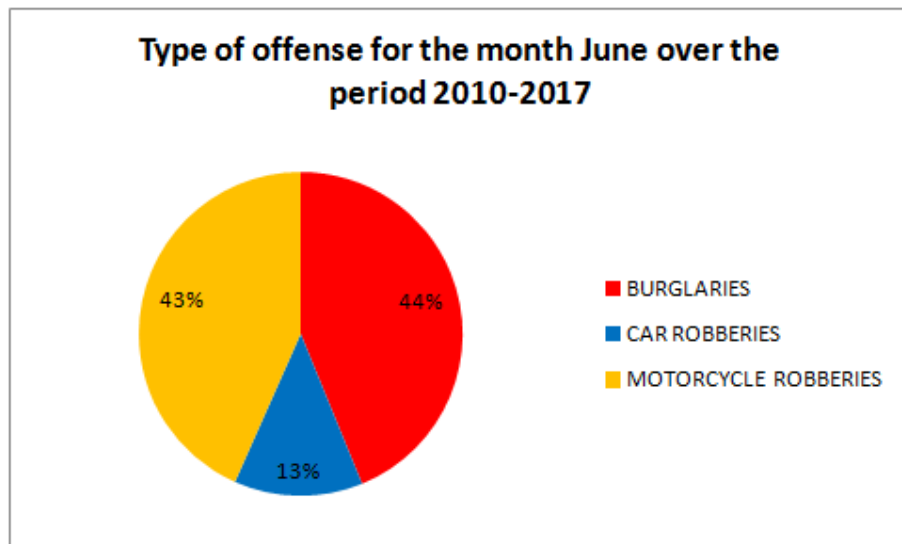


Figure 7.2.6: Type of offense for the month June over the period 2010-2017

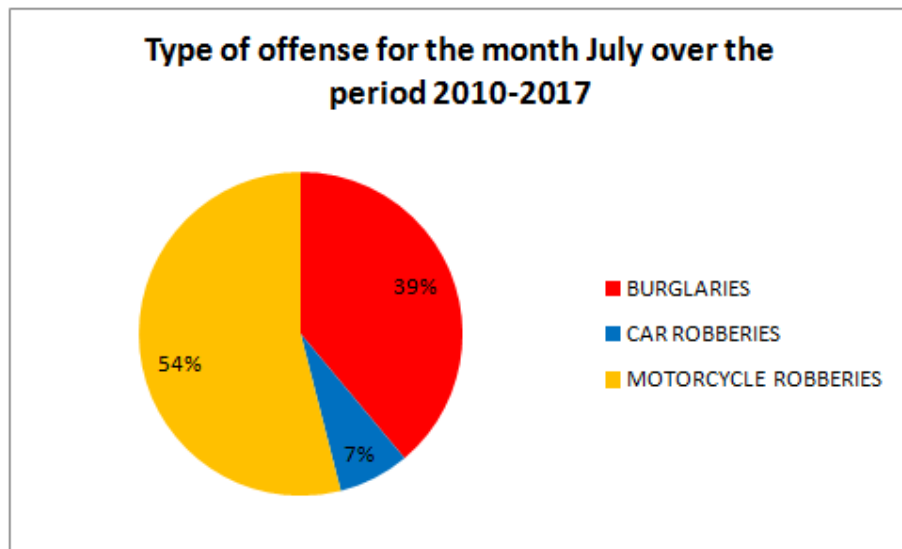


Figure 7.2.7: Type of offense for the month July over the period 2010-2017

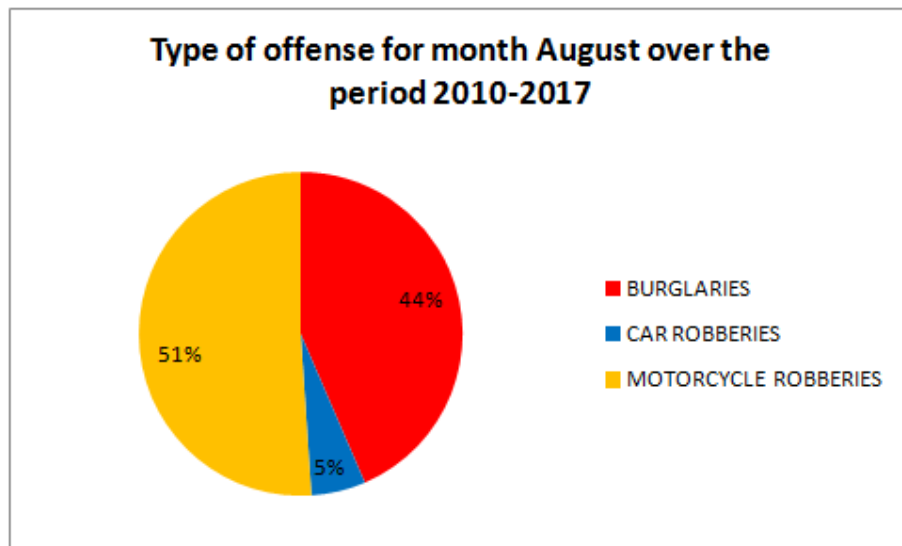


Figure 7.2.8: Type of offense for the month July over the period 2010-2017

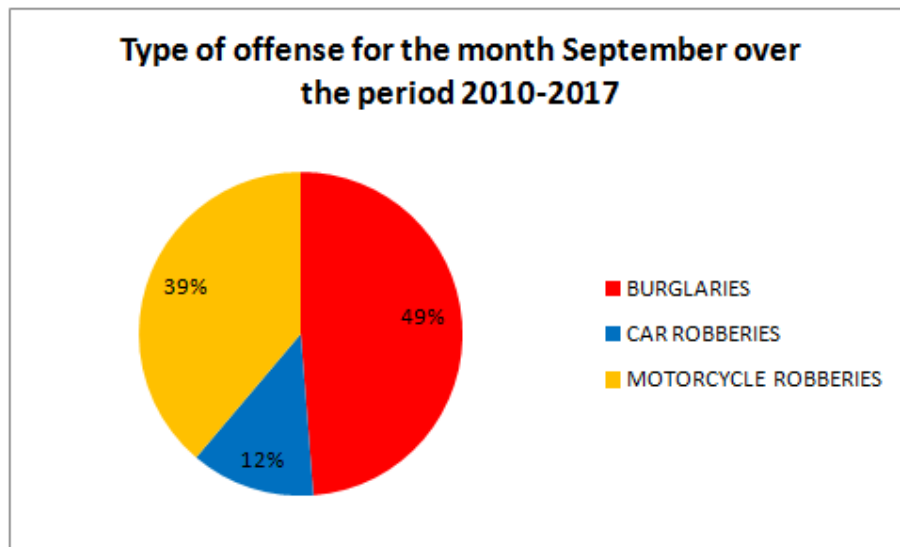


Figure 7.2.9: Type of offense for the month September over the period 2010-2017

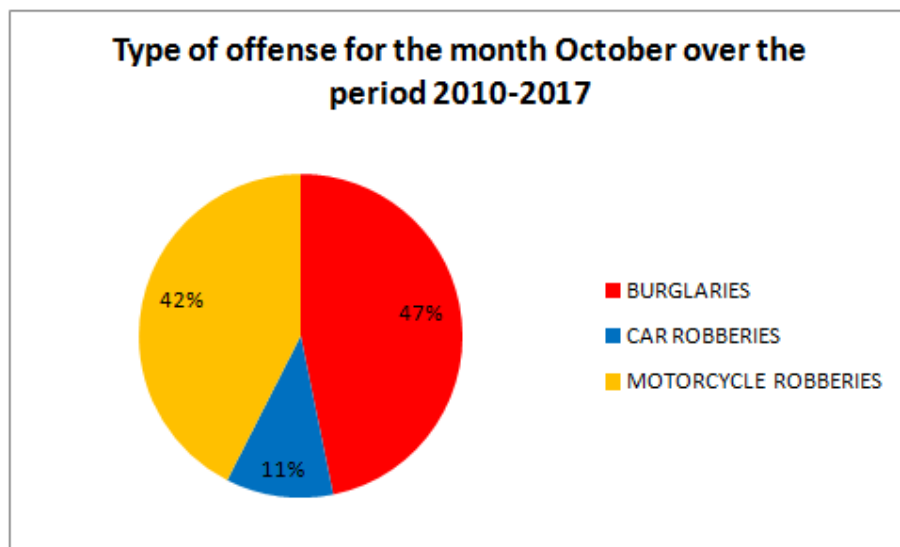


Figure 7.2.10: Type of offense for the month October over the period 2010-2017

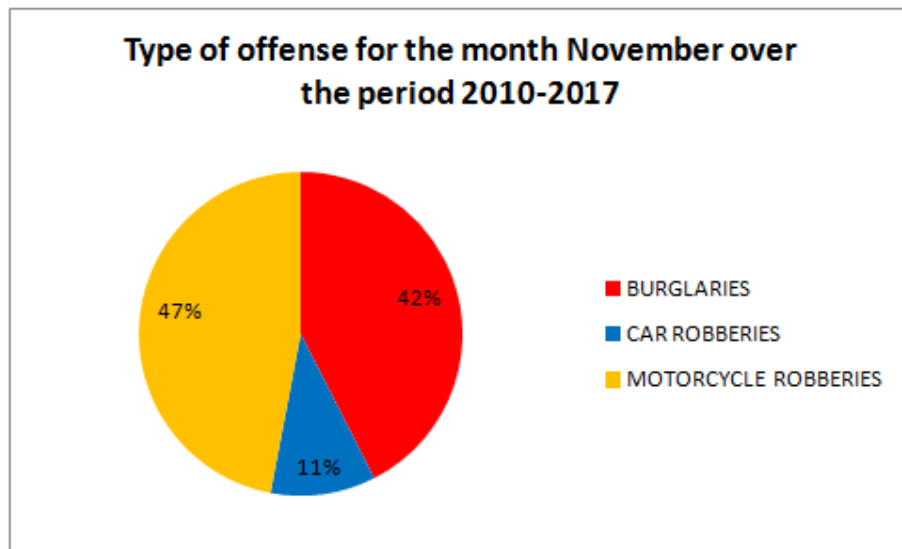


Figure 7.2.11: Type of offense for the month November over the period 2010-2017

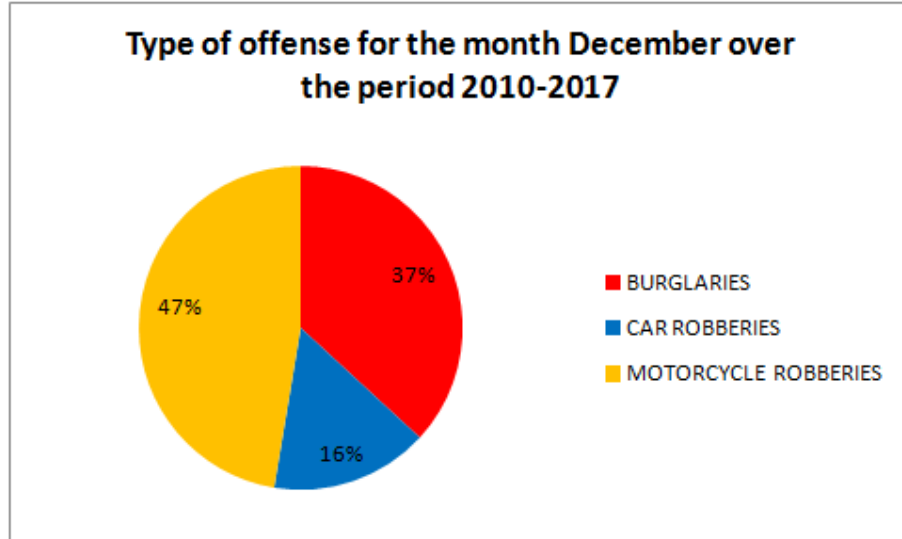


Figure 7.2.12: Type of offense for the month November over the period 2010-2017

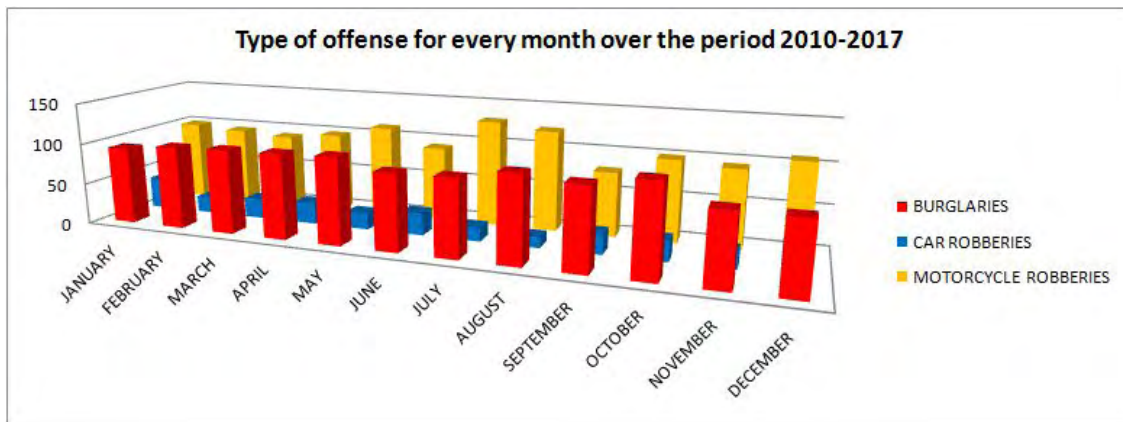


Figure 7.2.13: Comparative chart for the type of offense for all the months over the period 2010-2017

From the above charts it is concluded that the most common type of offenses over the months is the Motorcycle Robberies. The Burglaries are second in the row with not a big difference while the Car Robberies are the type of offense the less frequently occurring, with a great deal of difference from the other two types.

### 7.3 Command used for prediction in Splunk Machine Learning Toolkit

In chapter 3.4, for the prediction of future incidents using the Splunk Machine Learning Toolkit the following command and the parameter values used in each case are:

#### Command

```
| inputlookup<file_name>.csv
| eval _time=strptime(Month, "%m/%d/%Y")
| timechart span=1mon values(<incidents_type_name>) as <incidents_type_name>
| predict "incidents" as prediction algorithm=<algorithm_name> lower"95"=lower"95"
future_timespan="<number_of_periods>" holdback="<number_of_periods>"
upper"95"=upper"95"
| `forecastviz(<number_of_periods>, <number_of_periods>, "incidents", 95)`
```

where:



Parameters	Values
file_name	area_1, area_2, area_3
incidents_type_name	incidents, burglaries, car robberies, motorcycle robberies
algorithm_name	LL, LLT, LLP, LLP5
number_of_periods	12, 24, 36, 48, 60, 72

#### 7.4 Number of incidents in city sectors

In this section the table containing the number of crime incidents (occurred from 2010 until 2017) fall into each city sector is presented.

Sector	Sector coordinates				Number of Incidents
	Lat1	Lon1	Lat2	Lon2	
p1	39.38267	22.92168	39.37922	22.92587	14
p2	39.38267	22.92587	39.37922	22.930183	12
p3	39.38267	22.930183	39.37922	22.934496	0
p4	39.38267	22.934496	39.37922	22.938637	9
p5	39.37922	22.91737	39.375678	22.92168	4
p6	39.37922	22.92168	39.375678	22.92587	38
p7	39.37922	22.92587	39.375678	22.930183	22
p8	39.37922	22.930183	39.375678	22.934496	2
p9	39.37922	22.934496	39.375678	22.938637	25
p10	39.37922	22.938637	39.375678	22.942778	15
p11	39.375678	22.91737	39.37256	22.92168	8
p12	39.375678	22.92168	39.37256	22.92587	32
p13	39.375678	22.92587	39.37256	22.930183	36
p14	39.375678	22.930183	39.37256	22.934496	42
p15	39.375678	22.934496	39.37256	22.938637	28
p16	39.375678	22.938637	39.37256	22.942778	31
p17	39.375678	22.942778	39.37256	22.947118	47
p18	39.375678	22.947118	39.37256	22.951458	28
p19	39.375678	22.951458	39.37256	22.955728	9
p20	39.375678	22.955728	39.37256	22.959998	21
p21	39.375678	22.959998	39.37256	22.964236	17
p22	39.37256	22.92168	39.369441	22.92587	30
p23	39.37256	22.92587	39.369441	22.930183	20
p24	39.37256	22.930183	39.369441	22.934496	28
p25	39.37256	22.934496	39.369441	22.938637	34
p26	39.37256	22.938637	39.369441	22.942778	49
p27	39.37256	22.942778	39.369441	22.947118	42
p28	39.37256	22.947118	39.369441	22.951458	25
p29	39.37256	22.951458	39.369441	22.955728	29

p30	39.37256	22.955728	39.369441	22.959998	29
p31	39.37256	22.959998	39.369441	22.964236	13
p32	39.369441	22.92168	39.365999	22.92587	69
p33	39.369441	22.92587	39.365999	22.930183	53
p34	39.369441	22.930183	39.365999	22.934496	17
p35	39.369441	22.934496	39.365999	22.938637	54
p36	39.369441	22.938637	39.365999	22.942778	52
p37	39.369441	22.942778	39.365999	22.947118	42
p38	39.369441	22.947118	39.365999	22.951458	42
p39	39.369441	22.951458	39.365999	22.955728	47
p40	39.369441	22.955728	39.365999	22.959998	28
p41	39.369441	22.959998	39.365999	22.964236	17
p42	39.365999	22.92168	39.362556	22.92587	19
p43	39.365999	22.92587	39.362556	22.930183	21
p44	39.365999	22.930183	39.362556	22.934496	21
p45	39.365999	22.934496	39.362556	22.938637	25
p46	39.365999	22.938637	39.362556	22.942778	47
p47	39.365999	22.942778	39.362556	22.947118	33
p48	39.365999	22.947118	39.362556	22.951458	89
p49	39.365999	22.951458	39.362556	22.955728	68
p50	39.365999	22.955728	39.362556	22.959998	31
p51	39.365999	22.959998	39.362556	22.964236	20
p52	39.362556	22.92587	39.35923	22.930183	21
p53	39.362556	22.930183	39.35923	22.934496	17
p54	39.362556	22.934496	39.35923	22.938637	19
p55	39.362556	22.938637	39.35923	22.942778	3
p56	39.362556	22.942778	39.35923	22.947118	85
p57	39.362556	22.947118	39.35923	22.951458	91
p58	39.362556	22.951458	39.35923	22.955728	64
p59	39.362556	22.955728	39.35923	22.959998	43
p60	39.362556	22.959998	39.35923	22.964236	27
p61	39.362556	22.964236	39.35923	22.96842	22
p62	39.35923	22.947118	39.35581	22.951458	20
p63	39.35923	22.951458	39.35581	22.955728	26
p64	39.35923	22.955728	39.35581	22.959998	18
p65	39.35923	22.959998	39.35581	22.964236	24
p66	39.35923	22.964236	39.35581	22.96842	19
p67	39.35581	22.955728	39.35255	22.959998	32
p68	39.35581	22.959998	39.35255	22.964236	0

Table 7.4.1: Number of incidents in each city sector

## 7.5. Minimum distances matrices for sectors p25 and p38

From the execution of the Java code scripted by Professor Athanasios Lois, the minimum distance matrices for the network graph of sectors p25 and p38 are produced. These matrices are presented in this chapter of the Appendix analytically.

- The minimum distance matrix for the network graph of sector p25 is:

	132	147	154	372	391	421	448	464	609	636	923	1215	1239	1248	1610	1733	1854	1918
132	0	328	203	621	436	612	872	819	769	358	723	735	691	240	756	416	666	666
147	343	0	139	292	443	443	543	825	862	29	437	406	362	89	444	87	337	337
154	509	501	0	794	609	785	1045	992	942	531	896	908	864	413	929	589	839	839
372	289	178	85	0	389	621	721	771	721	207	615	584	540	89	622	265	515	515
391	198	271	146	563	0	399	814	606	556	300	510	581	633	182	543	358	608	608
421	545	436	493	407	645	0	608	248	201	406	160	185	426	530	147	349	401	401
448	520	177	316	227	620	378	0	853	797	148	372	341	186	266	379	90	27	27
464	449	188	245	160	549	781	360	0	881	159	775	744	178	249	782	101	153	153
609	702	441	498	413	802	439	613	253	0	412	327	305	431	502	343	354	406	406
636	372	29	168	263	472	414	513	889	833	0	408	377	332	118	415	58	307	307
923	656	429	604	400	756	113	601	241	87	399	0	70	419	641	32	342	394	394
1215	727	456	675	428	827	185	628	268	113	427	70	0	446	712	38	369	421	421
1239	952	610	749	855	1052	944	245	759	1363	592	938	907	0	698	945	650	272	272
1248	254	89	50	381	354	532	632	737	687	118	526	495	451	0	533	176	426	426
1610	689	462	637	433	789	147	634	274	75	432	32	38	452	674	0	375	427	427
1733	430	87	226	205	530	356	456	831	775	58	350	319	275	176	357	0	250	250
1854	492	149	288	199	592	350	391	825	769	120	344	313	159	238	351	62	0	0
1918	492	149	288	199	592	350	391	825	769	120	344	313	159	238	351	62	0	0
2027	638	302	358	273	738	133	474	114	552	272	126	158	292	362	159	215	267	267
2155	449	106	245	353	549	503	449	946	922	90	497	466	422	195	504	147	397	397
2235	386	44	183	289	486	440	540	869	859	26	434	403	359	132	441	84	334	334
2242	636	305	584	276	736	131	477	117	550	275	124	160	295	365	157	218	270	270
2257	421	160	217	132	521	276	376	752	695	131	270	239	195	221	277	73	170	170
2305	464	121	260	168	564	318	418	794	737	92	312	281	237	210	319	34	212	212
2328	512	169	308	219	612	370	410	845	789	140	364	333	178	258	371	82	19	19

Table 7.5.1: Minimum distance matrix of sector p25

- Minimum distance matrix for sector p38:

	30	118	140	289	374	454	455	559	656	947	1348	1522	1564	1728	2031	2194	2221	2260	2269	2280	2353	2362	2394	2412
30	0	229	470	1374	574	235	441	819	244	1414	116	1355	1007	1431	211	249	220	282	798	190	743	233	898	511
118	132	0	603	857	707	368	574	158	15	897	249	838	340	914	241	20	353	312	931	323	79	365	1031	641
140	181	50	0	910	110	417	623	207	64	950	298	891	389	967	31	70	402	102	980	372	129	414	1080	691
289	963	396	637	0	741	402	608	151	411	40	283	1066	320	57	378	416	386	449	965	356	308	399	232	681
374	222	349	692	800	0	458	663	245	364	840	338	781	433	857	331	369	442	114	1020	412	169	455	1032	621
454	1085	809	759	1663	863	0	730	1108	824	1703	1201	1644	1296	1720	791	829	569	862	1245	1275	1032	143	1345	801
455	211	79	29	940	140	447	0	237	94	980	328	921	419	997	61	99	432	132	1010	402	158	444	1110	721
559	919	352	593	933	697	359	564	0	367	973	239	914	416	990	334	372	343	405	921	313	155	356	80	631
656	117	244	588	842	692	353	559	143	0	882	234	823	325	899	226	5	338	297	916	308	64	350	1016	631
947	923	356	597	1045	701	362	568	111	371	0	243	1026	280	17	338	376	347	409	925	317	268	359	192	641
1348	680	404	354	1258	458	479	325	703	419	1298	0	1239	891	1315	386	424	103	457	682	73	627	485	782	391
1522	982	415	656	19	760	421	627	170	430	59	302	0	339	76	397	435	406	468	984	376	327	418	251	701
1564	1014	451	692	324	796	458	663	101	466	364	338	305	0	381	433	471	442	504	1020	412	255	455	180	731
1728	1008	445	686	317	790	452	657	94	460	357	332	298	263	0	427	465	436	498	1014	406	248	449	173	731
2031	150	18	621	875	725	386	592	176	33	915	267	856	358	932	0	38	371	330	949	341	97	383	1049	661
2194	112	239	583	837	687	348	554	138	254	877	229	818	320	894	221	0	333	292	911	303	59	345	1011	621
2221	812	245	486	1098	590	305	457	399	260	1138	132	1079	581	1155	227	265	0	298	814	206	320	302	245	531
2260	158	285	628	906	106	394	599	184	300	946	274	887	366	963	267	305	378	0	956	348	105	391	1056	671
2269	1178	823	1064	427	1168	829	1035	576	838	467	710	408	476	484	805	843	813	876	0	783	735	826	659	1101
2280	606	330	280	1184	384	405	251	629	345	1224	722	1165	817	1241	312	350	174	383	608	0	553	411	708	321
2353	1008	431	672	315	776	438	643	79	446	355	318	296	261	372	413	451	422	484	1000	392	0	435	160	711
2362	941	459	615	622	719	68	586	450	474	662	346	603	671	679	441	479	425	512	196	419	371	0	296	651
2394	956	830	780	846	734	1248	751	148	844	886	1129	827	330	903	811	850	1232	848	1914	1202	69	1245	0	671
2412	282	409	752	860	60	518	723	305	424	900	398	841	493	917	391	429	502	174	1080	472	229	515	1092	
2430	53	142	523	1427	627	288	494	872	157	1467	169	1408	1060	1484	124	162	273	195	851	243	796	285	951	561
2440	51	139	521	1425	625	287	492	870	154	1465	167	1406	1058	1482	121	159	271	192	849	241	794	284	949	561
2473	751	184	425	1329	529	550	396	774	199	1369	71	1310	962	1386	166	204	175	237	753	145	698	556	853	461
2500	1183	513	754	432	858	519	725	268	528	157	400	413	481	173	495	533	504	566	124	474	425	516	349	791
2589	1059	704	945	308	1049	710	916	457	719	348	591	289	357	365	686	724	694	757	1440	664	616	707	540	981
2601	972	409	650	279	754	416	621	57	424	319	296	260	225	336	391	429	400	462	978	370	214	413	139	691
2634	879	312	553	1001	657	208	524	302	327	1041	199	982	484	1058	294	332	302	365	881	272	223	205	148	591
2759	484	208	158	1062	262	720	129	507	223	1102	600	1043	695	1119	190	228	704	261	1282	674	431	717	1294	201
2763	811	244	485	933	589	610	456	235	259	973	131	914	417	990	226	264	234	297	813	204	156	616	80	521
2783	590	314	264	1168	368	389	235	613	329	1208	706	1149	801	1225	296	334	158	367	592	780	537	395	692	301
2814	601	325	275	963	379	334	246	624	340	1003	687	944	812	1020	307	345	103	378	537	760	548	340	637	311

Table 7.5.2: Minimum distance matrix of sector p38