National College of Ireland

Comparative Predictive Model for Road Traffic Accidents Involving Bicycles in Ireland

MSc Research Project
Data Analytics

Sergiy Mykhalevskiy
x14127717

School of Computing
National College of Ireland

Supervisor: Dr. Catherine Mulwa

National College of Ireland
Project Submission Sheet – 2015/2016
School of Computing

Student Name:	Sergiy Mykhalevskiy
Student ID:	x14127717
Programme:	Data Analytics
Year:	2016
Module:	MSc Research Project
Lecturer:	Dr. Catherine Mulwa
Submission Due Date:	12/12/2016
Project Title:	A Comparative Predictive Model for Road Traffic Accidents Involving Bicycles in Ireland
Word Count:	7623

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature:	
Date:	20th December 2016

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
3. Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

A Comparative Predictive Model for Road Traffic Accidents Involving Bicycles in Ireland

Sergiy Mykhalevskiy
x14127717
MSc Research Project in Data Analytics

20th December 2016

Abstract

Recently many people choose the bicycle as a means of transportation because of low costs, avoiding traffic, convenient traffic and improving health. But the increase of cyclists on the roads is leading to increase accidents involving them. Due the vulnerability of the cyclists it is significant to identify the main factors contributing to increase in severity of accident occurrence in Ireland. The research explores this factors using Linear Regression, Multinomial Logit Regression, Decision Tree and Clustering and compares severity of cyclists collision in Dublin verses Cork and rural areas. The Multinomial Logistic Model predicts the probability in three injury severity outcomes: fatal, serious and minor. The analyses based on Road Safety Authority data between 2004 and 2013 in Ireland. The largest increase of the probability of fatal injury increases is caused by speed. Poor visibility on the roads increases fatality due to serious accidents. Interesting facts were relieved in comparison of accidents occurrence in Dublin, Cork and rural areas.

1 Introduction

1.1 Problem Statement

"Recently there have been a rapid increase in road traffic accidents caused by bicycle drivers in different counties in Ireland. It is significant to critically analyse and identify the main factors contributing to these increases in severity of accident occurrences. This research proposes to identity factors contributing to these increases of accidents occurrence and to develop a prediction model of accident occurrence in order to support and enhance road safety awareness".

1.2 Research Objectives

Objective A: Literature review of road traffic accidents involving vehicles and bicycles in Ireland (2004-2013).
Objective B: To design, analyses, develop and evaluate the proposed prediction model for road traffic accidents involving vehicles and bicycles in Ireland.

Objective B1: Requirement specifications of the proposed prediction model.

Objective B2: Architectural and technical design of the prediction model.

Objective B3: To develop the prediction model.

Objective B4: Comparison of accidents occurrence in Dublin verses rural areas.

Objective C: The evaluation and results of the developed prediction model.
Objective D: Thesis technical report and configuration manual.

1.3 Research Methodology Used

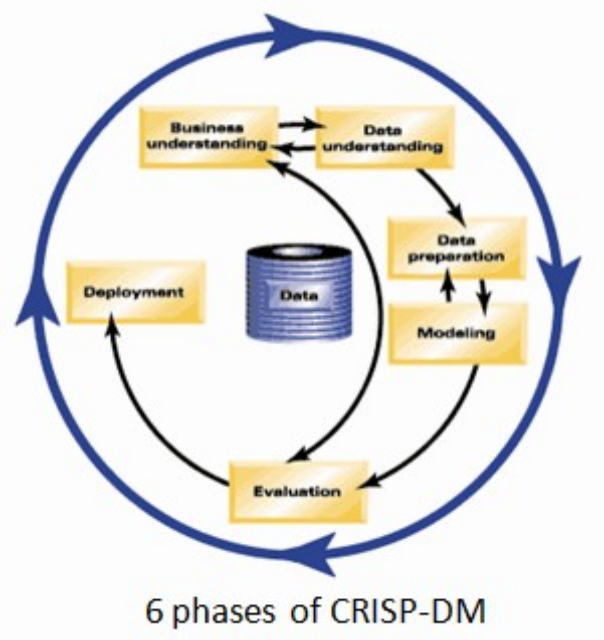

Figure 1: CRISP-DM

CRISP-DM or Cross-Industry Process of Data Mining was chosen because it is the most commonly used process for data miners (Figure 1). The six high-level phases are still a good description for the analytics process. The Business Understanding like the first phase helps focused analytics project on objectives that are most important for business. That is the main reason of choosing CRISP-DM for current research project.

1.4 Thesis Contributions

1. The results of literature review especially the identified gaps in the body of knowledge.
2. The developed, fully evaluated, and analysed results of the prediction model for road traffic accidents involving vehicles and bicycles in Ireland.
3. The results of comparison of accidents occurrence in Dublin verses rural areas.
4. Full documentation of the technical report and configuration manual on how to run the developed model and environment/software installs configurations.

2 Literature Review of Predictive Models for Road Traffic Accidents (2004-2013)

2.1 Introduction

In recent years because of heavy traffic in city urban areas, attempts to reduce the use of hydrocarbons and thus to reduce the emission of carbon dioxide European governments stimulate the development of cycling infrastructure. Dublin is no exception. Along with the construction of bicycle paths and parking lots the public bicycle rental scheme was launched June 13, 2009. Since the scheme was introduced bicycle rental as well as the number of the bicycles that belong to the scheme and the number of bike rental station has increased significantly. All this increased the number of bicycles on the city roads. But bicycle use has increased not only in Dublin but in Ireland as a whole. The objectives of this study to investigate whether of increasing using bicycles has led to increase in severity of accident occurrence involving both bicycle and other vehicles and identify the main factors contributing to these increases and to develop a prediction model of accident occurrence in order to support and enhance road safety awareness.

It is difficult to overestimate the importance of road safety. Many people are killed or seriously injured in road accidents every

year. The cyclist is one of the most vulnerable road users. On the other hand, using the bicycle as a transport is reducing the use of hydrocarbons and thus to reduce the emission of carbon dioxide, avoiding traffic congestions and improving health so it stimulates by the Irish government. It is significant to critically analyse and identify the main factors contributing to increase in severity of accident occurrences. This research proposes to identity factors contributing to these increases of accidents occurrence and to develop a prediction model of accident occurrence in order to support and enhance road safety awareness.

2.2 Investigation of Prediction Models for Road Traffic Accidents

Safety and efficiency are main objectives of transportation engineering. This paper will focus on safety. More than 1 million of people die and more than 50 million are injured within a year in road accidents all over the world (World Health Organization, 2009). Improving safety on the roads is very important but challenging task because of complexity of road traffic system. Road safety management dire needs of understanding and accurate prediction of traffic system safety. A lot of research has been conducted due to the importance of such problem in recent decades. Some of papers that describe different technics and methodologies to improve road traffic system safety accuracy will be shown below. In accident prediction methodology three different approaches were tried:

- Multiple Linear regression;
- Poisson regression;
- Negative Binominal regression.

In the University of Central Florida research was carried out by creating two mathematical models. One explains the relationships between the frequency of accidents and road geometrics and traffic characteristics, another is the model of accident involvement for different gender and age groups. Two models were built because two main factors are very important in occurrence of road accidents: one is related to the driver and another one is related to the road design. In this paper suggest that multiple linear regression is not the best for prediction of road accidents because the normal distribution that is the basic stone of it should be used in caution because of the problems with non-negativity and error terms. Poisson regression overcomes such a problem and can be used with smaller sample size but the limitation of Poisson regression model is that the variance of the accident frequency is constrained to be equal to the mean. The Negative Binominal regression model helps to avoid this problem. The results of this paper show that the road design and traffic factor affect road accidents occurrence as well as a driver factor. Increasing AADT (Annual Average Daily Traffic) per lane increases occurrence of road accidents. On the other hand, urban roadway segments have higher potential for accident occurrence than rural one. Male drivers have greater tendency to be involved in accidents while speeding but female drivers have higher probability of accidents during heavy traffic as well as young and older drivers experience higher probability to be involved in accident during heavy traffic than middle age drivers [1].

Scientists from Texas Transportation institute confirms that using Poisson regression and Negative Binominal models for road accident prediction is one of the best approaches. Bernoulli trials with independence among crashes and unequal crash probabilities across drivers, vehicles, roadways, and environmental conditions describe accident occurrence data in the best way. In its turn, Bernoulli trials can be well approximated as a Poisson trials. Poisson

regression and Negative Binominal regression models are statistical approximation to the accident occurrence process. The first one serves better under homogenous conditions while the second one serves better in other conditions [5].

But traditional accident occurrence prediction model (linear regression model) cannot take into account multilevel structure of input data. To overcome this problem 5xST-level hierarchy ((Geographic region level Traffic site level Traffic crash level Driver-vehicle unit level Occupant level) Spatiotemporal level) is proposed to present a multilevel structure of input data. Bayesian hierarchical approach is a dominant way to model the hierarchical structure. But it is required a large amount of computation so the cost should be taken into account [2].

Another study compares Bayesian Spatial Joint model with Poisson, Negative Binominal and Conditionally Autoregressive (CAR) models. The results show that Joint and CAR predictive performance is better than Poisson and Negative Binominal. On the other hand, the predictive performance of CAR and Joint models are equal. Joint model provides a new perspective at prediction of accident occurrence by taking into account spatial correlation of different types of road data input variables. But this model is very complex and need a significant computational power of software and hardware as well as it is a time-consuming process. The cost of Joint approach definitely should be taking into account [10].

There are different methodological approaches to study accident occurrence frequency:

- Accident occurrence frequency;
- Crash severities.

Accident occurrence frequency is the number of crashes occurring on the road segment over some time period and it is non-negative count data. Poisson regression variant models is very suitable for this approach. In some methodology crashes is not considered like a count data but as duration of time between accidents. For the severity of crashes prediction simple discrete logit and probit models evolved to consider multiple discrete outcomes (different injury-severity categories). In the recent years developing of new technologies opens up great opportunity for safety road prediction but many methodological issues such as risk compensation, missing data, special and temporal correlation etc. remains [6].

All the above models can be used to predict accident occurrence involved cyclist and another vehicle.

2.2.1 Critical Investigation of Prediction Models for Bicycles

Because of vulnerability of cyclist the road accidents are more dangerous for them in comparison with car drivers. It is important to identify the main causes of accidents occurrence involving both bicycle and vehicle as well as bicyclist injury severity in such accidents. A multinomial logit model predicting injury severity conditional on accident occurrence was estimated with the method of maximum likelihood. The results show the main factors that increase the probability of fatal injury for cyclist:

- High speed of vehicle;
- Truck involving accident;
- Vehicle driver or cyclist intoxication;
- Darkness;
- Head-on collision.

Further study needs to explore detailed road and bicycle lane geometry to assist in developing more specific engineering solutions to reduce injury severity in accidents involving both bicycle and vehicle [3].

To protect our environment taking into account global problem of climate change, traffic congestion and road safety Irish government is forcing to stimulate the increase

use of walking, cycling, public transport and green vehicles. The advantages of cycling are:

- Low cost;
- No emission and energy use;
- Avoiding traffic congestion;
- Convenient parking;
- Improving health.

But cyclist is more likely to be injured in road accident. To create a safe road conditions for them the engineers from University of British Columbia proposed negative binomial models for bicycleauto collisions to evaluate cyclist road safety. The result show that accident occurrence involving both bicycle and vehicle is positive correlated using the workload of vehicle lanes and cyclist lanes, traffic lights, bus stops and intersection density. To promote increasing cycling new models regarding the relationship between road safety and number of cyclists on the roads have to be developed [9].

Because cycling is promoted by public health and transportation specialist predicting cycling accident risk and identifying how road infrastructure influences cycling safety are very important. The group of scientist applied logistic and conditional autoregressive model on the Brussels-Capital Region. As a result of the research the following recommendations for safer urban cycling were made:

- Special attention should be paid for cyclist safety when designing on-road tram tracks;
- Bridges should be designed with a great care for cyclist;
- Cyclist facilities should also be designed with a care.

Inappropriate design of the cyclist facilities could lead to increase the risk of accident occurrences instead of decreasing it [8].

In the articles mentioned above the researches are mostly about road geometry and its impact of severity of road accident occurrence or correlation between accident occurrence and workload of vehicle and cyclist lines, traffic lights, bus stop and crossroads. The influence of environmental factors such as weather conditions, daylights hours and road surface conditions on severity of collision occurrence involving cyclists are studied not enough.

2.3 Identified Missing Gaps in Body of Knowledge

There are not too many papers about accident occurrence involving cyclists at all but even more difficult to find the research of such an accident occurrence in Ireland. The papers published by RSA on such a topic were found. But even in studies not related to Ireland are missing gaps. Most papers describe how to engineer the road for cyclist considering road geometry such as number of lines, special bicycle lines and particular dangerous areas like bridges, tram tracks etc. or concern the cyclist person: age, gender, alcohol level etc.

One of the papers about cyclist injury severities in bicycle accidents identifies the main factors that increase the probability of fatal injury for cyclist including high speed of vehicle, truck involved in accident, vehicle driver or cyclist intoxication by alcohol, head on collision and darkness [3]. In this paper, another factor like number of vehicles involved, time of day, weather condition, speed limit, surface condition was taken into consideration.

Due the fact that very little information of analysing severity of bicycle accident occurrence in Ireland has been found it is important to identify and investigate as many factors contributing to severity of cyclist accidents as possible and explore their relationship and influence on the severity of the accident. The paper that was published by RSA in 2014 presents the following finding:

- Trends in injuries among cyclist;
- Gender and age of cyclist causalities;
- When cyclists were injured;
- Light condition;
- Trip purpose;
- Injury breakdown by county.

Despite the authors above have done a great job there is still missing gaps in the body of knowledge. The impact of such factors as speed limit, weather condition, surface condition, number of vehicles involved in accident on the severity of accident have not been studied. They will be investigated in current study [7].

2.4 Conclusion

More and more people choose to use bicycle like a transport every year in Ireland according to statistics. But increasing the number of cyclists leads to the fact that the number of traffic accident occurrences increases too. So, it is significant to critically analyse and identify the main factors contributing to these increases in severity of accident occurrences. This research proposes to identity factors contributing to these increases of accidents occurrence and to develop a prediction model of accident occurrence in order to support and enhance road safety awareness.

Different approaches and models were discussed above. Some of them work better some worse, some is better for specific purposes but there is significant gap of knowledge in this field and offers a great scope for new research especially in Ireland. In this paper, new factors contributing to increase the severity of accident occurrence were identified. To create the prediction model Multiple and Linear Regression, Clustering and Decision Tree technics were used. The results were analysed and visualized.

3 Design, Implementation and Evaluation of the proposed prediction model

3.1 Introduction

This paper is mainly focused on identifying factors contributing to these increases of accidents occurrence and to developing a prediction model of accident occurrence in order to support and enhance road safety awareness. It is significant because of lack of information and gaps of knowledge in this particular field. Despite the fact that number of cyclist in Ireland is increasing year by year it is still not so much research examines the factors determine the severity of accident occurrence involved cyclist was carried out.

Vulnerability of cyclist on the one hand and the advantages of using this type of transport on another make the problem of developing a prediction model in this field actual in Ireland. In this study the attempt to create a model helps to predict the main factors influence the severity in road accident involving cyclist using Liner and Multiple Regression, Clustering and Decision Tree technics was made.

The scope of the project is to critically analyse and identify the main factors contributing to increase in severity of accident occurrences and to develop a prediction model of accident occurrence in order to support and enhance road safety awareness. In this study data was provided by RSA Road Safety Authority. The organization tries to make Irish roads safer for every road user, to save lives and to prevent injuries by reducing the number and severity of road accidents. RSA collects, analyse and provide free information about safety on Irish roads. After request was made the data were provided.

3.2 Specification and Architectural Design of Proposed Solution

3.2.1 Data Preparation and Requirement Specification

(1) Data Preparation

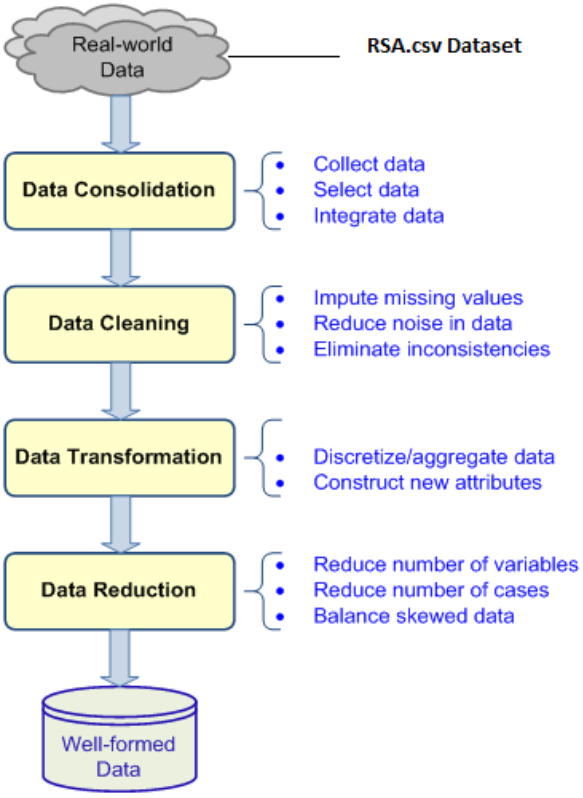

Figure 2: Data Preparation Diagram

The idea of project is to extract, clean, transform, analyse and implement the prediction model for enhancing road safety and supporting the RSA.

The data consists of ten different data sets containing data from 2004 to 2013, one data set per year. It contains information about all road users. The tables have from 5000 to 6000 observation of 241 variables (Refer to Appendix 1).

(2) Requirement Specification

A use case scenario was created that depicts three actors (data analyst, RSA company administrator and general user bicycle rider). This scenario is depicted in Appendix 2. Objective B1 is achieved.

(3) Data Warehouse

The candidate also specified, designed and created a data warehouse as part of the implementation of the prediction model but decided not to use when doing the computation of the regression, the decision tree and clustering (depicted in Appendix 3).

(4) Technologies Used

The main technologies used were a hybrid of data mining technologies: R Studio, Rapid Miner, Tableau and SQL server. This allow me to improve my computational skills.

3.2.2 Architectural and Technical Design of the Prediction Model

Architecture is designed specially so that the backend is only for data analytic use and the frontend is to provide the information to final user as cyclist or RSA staff. The data warehouse is in the backend and consists of database, SQL server and other technological tools such as R Studio, Rapid Miner, Tableau etc. which data analytics use to collect, clean, transform and finally prepare the data to the end user.

Rapid Miner and Tableau are in the backend of the system and available only data analytics engineers. Backend is created by data engineers and for data engineer use. This is done to separate the end user from business logic to protect primary data on the other hand to provide to decision maker the clear information which will facilitate his task to make a decision of the strategy of the company or in current case to make a decision to reduce road deaths and injuries by education, enforcement, engineering and evaluation. Information can be withdrawn by end user by sending the request.

Different algorithms are using to design a prediction model of traffic collision severity involving bicycle users. Multiple and Linear Regression were calculated to identify the factors that affect the severity of the acci-

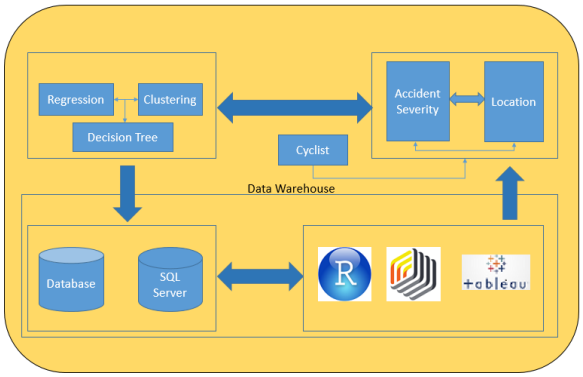

Figure 3: Architectural Design

dent. Computing the Decision Tree helps to make decision about importance of factors. Because it is not black box method it has open logic. According to the principle that records inside a cluster should be very similar to each other, but very different from those outside Clustering technique divides data into clusters, the set of close items. It helps to reduce complexity and make result more meaningful. Objective B2 is achieved.

Some additional analyses were conducted. The relationships between gender, age and severity of cyclist collision occurrence were analysed. The severity of road accidents involving cyclists was compared in three different areas Dublin, Cork and the rest of Ireland.

3.3 Implementation and Interpretation of Results of the Prediction Model

To implement the prediction model Linear Regression and Multinomial Logistic Regression as well as Regression Tree will be compute using R Studio and Clustering will be compute using Rapid Miner. All data will be send to the data warehouse and then using Tableau will be visualized and finally can be reached by final user.

The prediction model should identify the main factors contributing to increase in severity of accident occurrence involving cyclist. For this purpose, will be used Ireland road collision statistics data provided by RSA (Road Safety Authority) that cover the period from 2004 to 2013. Linear Regression model and Multinomial Regression model will be used to predict the main factors of accident occurrence severity and Regression Tree technic to make recommendation how to make the consequences of the accident not so serious.

Section 3.3.1 to 3.3.3 presents the implementation of the regression, decision tree and clustering algorithms which were specified as the main components of the proposed prediction model (Objective B). The results of this implementations (depicted in section 3.5) have enabled us to solve the research problem statement.

3.3.1 Regression Computations

Linear

In this research, Linear and Multinomial Logistic Regressions were computed. In the logit model the log odds of the outcome is modelled as a linear combination of the predictor variables. Simple Linear Regression formula is

$$y = \alpha + \beta x, \tag{1}$$

where α describes where the line crosses y axe and β describes the change in y given by increase of x. R provides the function to estimate α and β automatically. To create the Logit Model *glm* (generalised linear model) function was used (Table 1).

Call:
glm(formula = type hour + splimit + weat1, family = "binomial", data = bic)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-3.0165	0.1507	0.1887	0.2289	1.0674

Table 1: glm

To fit linear model *type* as the response and *weekday*, *speed limit*, *county*, *number of vehicle*, *weather*, *surface*, *collision type* and *hour* as predictors *lm*

function was used (Table 2).

Call:
lm(formula = type weekday + splimit + county + noveh + weat1 + surface + prcoltyp + hour, data = bic)
Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.95116	0.05807	0.09609	0.13919	0,76871

Table 2: lm

The Linear Regression Model was executed using Rapid Miner with setting role for *type* attribute as a label and *county*, *hour*, *light*, *month*, *noveh*, *prcoltyp*, *splimit*, $surface$, $weat1$ and *weekday* attributes as predictors. The process is shown in Figure 4.

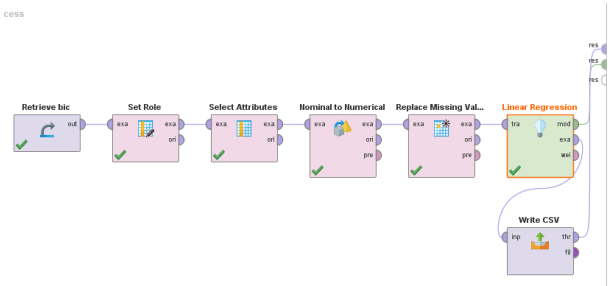

Figure 4: Linear Regression Model

Multiple The outcomes of Multinomial Logistic Regression are modelled as linear combination of predictors. This model is better to predict type than Logit Model because this dependant variable has three levels: fatal, serious and minor. Multinom function is used to estimate this model.

Call:
multinom(formula = type light + hour, data = bic)

Most analysis are computed using Multiple Regression. The formula is

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_i x_i + \varepsilon, \quad (2)$$

where α is an intersection with y axe and β_i is change in y given by increase of each x_i and ε is error. The same can be expressed in another formula

$$Y = X\beta + \varepsilon, \quad (3)$$

where Y, β andε are vectors, X is matrix with a column for each feature and additional column with 1 value for interception [4].

Polynomial Regression was executed using Rapid Miner with using *type* as dependant variable and *county*, *light*, *noveh*, *prcoltyp*, *splimit*, $surface$, $weat1$ attributes as predictors. The process is shown in Figure 5.

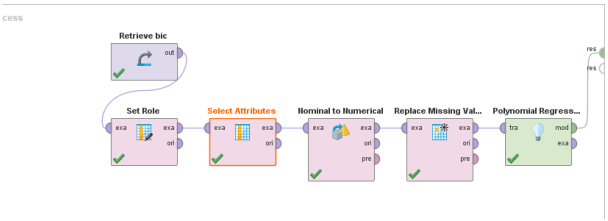

Figure 5: Plynominal Regression Model

3.3.2 Decision Tree Computations

Regression Tree is used to predict severity of collision occurrence involving bicycle users. Using this algorithm makes easier to build the model and to read and understand the results. The data set was divided on training and testing sets and used to train a Regression Tree Model.

Recursive Partitioning package was used as R implementation of CART (Classification and Regression Tree). Computing *rpart* function type attribute was specified as outcome and all others attributes from *bic train* data set were used as predictors.

Predict function was used for prediction on test data frame. Summary of prediction is show in Table 3.

Min	1Q	Median	Mean	3Q	Max
2.385	2.852	2.852	2.860	2.939	3.000

Table 3: Summary of Prediction

3.3.3 Clustering

Unsupervised machine learning Clustering is used in this study to reduce complexity and to find the hotspots of severity of accident occurrence involving bicycle users in Dublin and in the country.

To compute Clustering k-means algorithm was used like one of the most common methods. The *kmeans* function requires a data frame containing only numeric data and a parameter specifying the desired number of clusters. 10 attributes were chosen for this task. It is necessary to determine the number of clusters k. Through experimentation process with different values of k four clusters have been selected. The size of all four clusters is shown below in Table 4.

357	89	138	334

Table 4: Size of Clusters

3.4 Comparison of Accidents Occurrence in Dublin verses Rural Areas and Cork

Computing Clustering is the step to reduce complexity and to find the hotspots of severity of accident occurrence involving bicycle users in Dublin and in the country. $K-means$ algorithm is used for this task. The value of k was determined as a result of the experiments. Four clusters are most suitable. As shown in Figure 6 there are four clusters two smaller and two bigger. Two bigger clusters look nearly the same and mostly consist of minor accidents occurred in Dublin area. They are cluster1 and cluster4. Cluster 2 is smaller and consist of serious and fatal accidents occurred all over Ireland. Cluster 3 is bigger than cluster 2 but smaller than clusters 1 and 4 and consist mostly of minor accidents occurred in the country. Information provided in Figures 7 and 8. Analysing two scatterplots in this figures the conclusion can be made that vast majority of accidents and minor severity accidents happened in Dublin area (county number 6). It will be proved later in this study. But more serious accidents occurred more in Dublin too but that is not the vast majority of it. Will continue to analyse frequency and severity of accident occurrence involving cyclists in three different areas:

- County Dublin;
- County Cork;
- The rest of Ireland.

Due to the task the necessary data was extracted and new table created. In Table 5 frequency of accidents occurrence with different severity per county is presented. Most accidents happened in Dublin. It is normal considering that Dublin is the capital and the most populated area. But if compare accident occurrence taking in account population it is not so simple. Frequency of minor and serious accident occurrence per 1.000 people the biggest but number of fatal cases per population the smallest. It may be due the ability to provide the necessary medical assistance to the victims of accident. On the other hand, in Cork percentage of fatal cases from all accident the highest (Figure 9) even higher than in country but the lowest percentage of serious cases. In the country percentage of fatal cases the second highest but it can be explained by the facts that speed limit higher on the country roads, medical infrastructure is not as good like in the city (Refer to Appendix 4). .

It is impossible to analyse the severity of road accidents occurrence involving cyclist in each county separately due to lack of

data. For example, County Carlow has only 2 records, County Cavan 1record, County Leitrim 1 record. For the majority of counties analysis is possible and shown in Figure 11. According to diagram can be concluded that percentage of fatal and serious cases in Dublin one of the lowest in Ireland. Objective B4 was achieved.

3.5 Results of the Implement Prediction Model

After extracting, cleaning, preparing, or after ETL processes the final data was delivered in SQL server and was analysed using Rapid Miner, R Studio, Tableau and Excel. In this chapter the results, tables, diagrams and graphics are shown and discussed.

A correlation test between dependent and independent variables did not produce the expected result. Only low negative correlation between dependent variable type of severity of accident and independent variable speed limit has been found. But correlation does not mean causation and probably it is the reason of such a result.

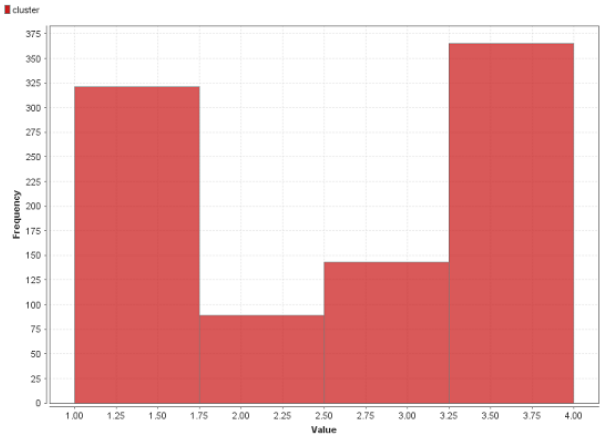

Figure 6: Clusters Histogram

By creating a table of frequency of fatality in accident occurrence was found

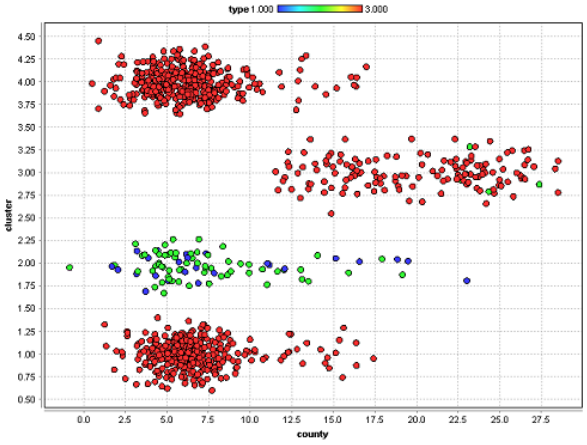

Figure 7: Scatterplot 1

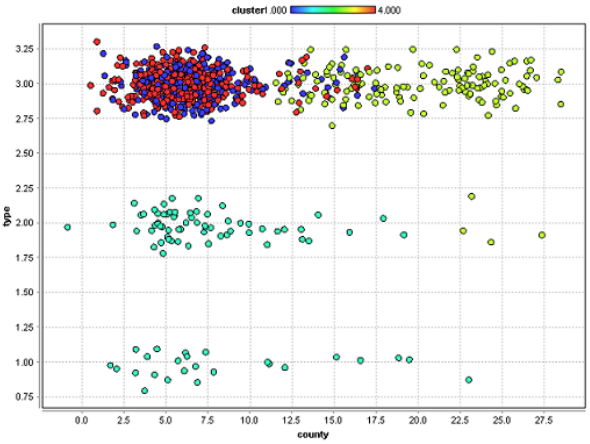

Figure 8: Scatterplot 2

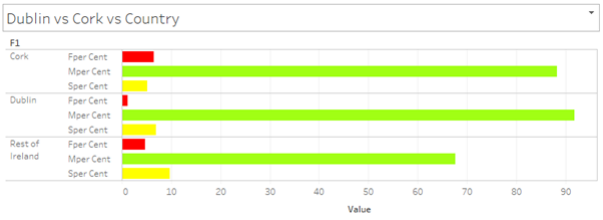

Figure 9: Accident Severity

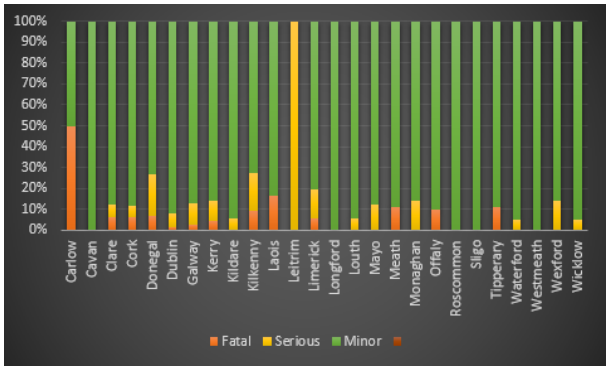

Figure 10: Accident Severity per County

that overwhelming number of accidents involving cyclists is minor severity accidents as shown in its turn can affect the correlation.

We can identify some relationship by using simple aggregation and creating a cross table with relationship between dependent variable type with one of each independent such as light and weather. It is simple to spot the lowest value of severity type attribute belongs to category 6 of light attribute corresponding to dark no lighting, to category 6 of weather attribute corresponding to high winds weather conditions and to category 3 corresponding to frost/ice weather conditions. The lower value of type of severity the closer it to serious (2) and fatal (1) severity of accident occurrence. But in case with high winds weather conditions, only two accidents were registered. It is not enough to make any conclusions. From the above data, it can be preliminary concluded that the most unfavourable for bicycle users to cycle with high winds in the dark. It looks logical.

The above conclusion was done without doing any modelling or even any visualization and even if it looks logical it doesnt mean it is correct. To learn data deeper some algorithms were computed and some models created. Linear Regression Model was computing using Rapid Miner software andit helps to identify the main factors contributing to increase of severity in cyclist collision occurrence. As shown in Table 5 the most significant are weat =1 and surface =1 attributes that corresponding to dry weather condition and dry surface. All tables are provided in Appendix 1.

This maybe several reasons. Firstly, much more people prefer cycling in nice sunny day and secondly, good driving conditions makes people more careless. The significance of weat = 6, light =6 and surface = 2 corresponding to high winds, dark time and wet surface looks absolutely logical to increase severity of collision as well as speed limit. Speed limit is high positive correlated with real speed of the vehicles on the road and speeding leads to increase severity of accident occurrence especially if most vulnerable road user involved.

Logit Regression Model is not very suitable in current case because the dependent variables type has three levels and not binominal. Using Multinomial Logistics Regression is more useful. The results are shown below in Table 6 and Figure 11. Percent of minor severity of cyclist collision does not deviate too much according to light conditions but percent of fatal and percent of serious do. While light conditions change from 1 to 8 (from daylight with good visibility to dark) the percentage of fatal cases of accidents involving cyclists increased by reducing the percentage of serious cases. In other words, the deterioration of visibility leads to the fact that serious accidents are fatal longer. Objective B3 is achieved.

Attribute	Coefficient	Std. Error	Std. Coefficient	Tolerance	t-Stat	p-Value	Code ↓
surface = 1	-0.133	0.030	-0.146	1.000	-4.365	0.000	****
weat1 = 1	0.098	0.029	0.115	0.978	3.408	0.001	****
splimit	-0.003	0.001	-0.129	0.991	-3.853	0.000	****
surface = 2	-0.099	0.037	-0.090	0.984	-2.679	0.008	***
weat1 = 6	-0.572	0.207	-0.092	0.997	-2.762	0.006	***
light = 6	-0.178	0.091	-0.072	0.815	-1.958	0.051	*
light = 3	0.074	0.039	0.066	0.940	1.919	0.055	*
light = 1	0.030	0.030	0.034	0.994	1.003	0.316	
light = 7	0.082	0.070	0.039	0.977	1.166	0.244	
light = 4	0.014	0.068	0.007	0.996	0.207	0.836	
light = NA	-0.128	0.131	-0.033	0.995	-0.976	0.329	
light = 2	0.106	0.091	0.039	0.997	1.166	0.244	
surface = 7	-0.073	0.072	-0.034	0.992	-1.008	0.314	
surface = NA	0.029	0.071	0.014	0.980	0.408	0.683	
surface = 3	0.107	0.169	0.021	0.994	0.631	0.529	

Table 5: Linear Regression.

Regression Tree is calculated to make result more meaningful and easier to read. To make the results Even easier to understand they were visualized. In the Figure 11 plot starts with speed limit variables like the most important factor. If speed limit more than 70 km per hour it is increased a chance that an accident with more serious consequences will be occurred. Lower the number of dependent type variables more serious severity of collision. The second predictor involved is number of vehicles of accident occurrence. If speed limit is less than 70 km per hour and number of vehicles less than 2.5 the severity of accident more serious. Than number of vehicles appears again. If two vehicles involved it leads to less serious consequences than one vehicle involved. It probably means that collision between cyclist and pedestrian more traumatic. Using county variable is not very meaningful because it is discrete variables assigned to each county in alphabetical order. But type of collision, less than 1.5 means pedestrian involving that leads to more serious severity of collision occurrence.

Row No.	att1	Fatal	Serious	Minor
1	1	0.025	0.073	0.902
2	2	0.026	0.072	0.902
3	3	0.026	0.071	0.902
4	4	0.027	0.070	0.903
5	5	0.028	0.069	0.903
6	6	0.029	0.068	0.903
7	7	0.029	0.067	0.903
8	8	0.030	0.066	0.903

Table 6: Multinomial Regression.

To evaluate a model performance the mean absolute error (MAE) was used. The difference between our model's predictions and the true quality score was about 0.23. On a quality scale from 0 to 10, this seems to suggest that this model is doing good. Objective B3 is achieved.

Comparing accident occurrence distribution by gender and age can be concluded that vast majority of accidents occurred involved man cyclist (Figure 13). Men from 25 to 54 years old as well women of same age categories are mostly in the group of risk to be involved in road collision that involved bicycle users (Figures 14 and 15).

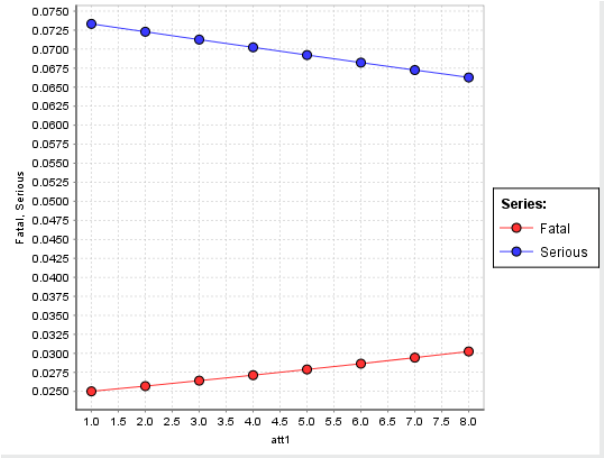

Figure 11: Severity Changing by Light Conditions

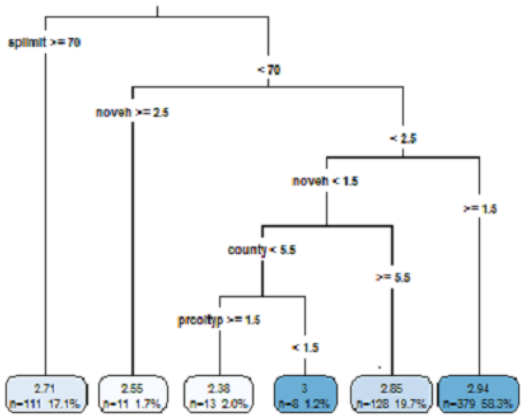

Figure 12: Plot

3.6 Testing and Evaluation of Implemented Components and Results

(a) Data Preparation Testing

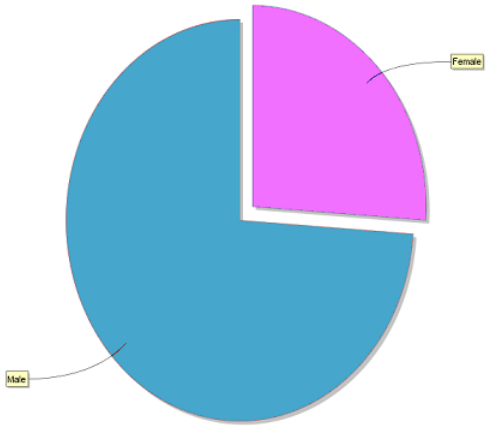

Figure 13: Accident Occurrence

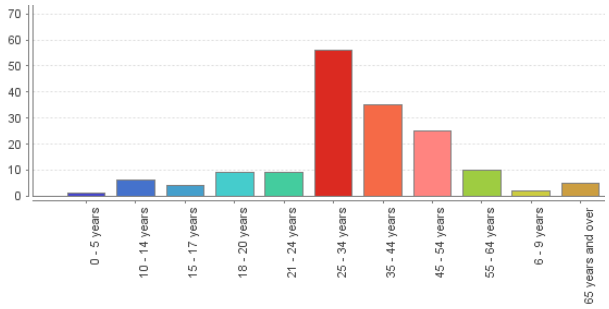

Figure 14: Accident Occurrence According to Female Age

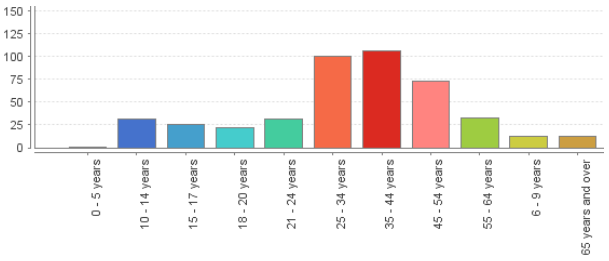

Figure 15: Accident Occurrence According to Male Age

Data from separated csv files provided by RSA was extracted and combined using rbind function. White spaces and were removed using *duplicated* and *na.omit* functions. Correct formulas were used to return correct results.

(b) Data Preparation Testing

(i) Regression Linear Model was created using Rapid Miner software and to fit linear model type as the response and weekday, speed limit, county, number of vehicle, weather, surface, collision type and hour as predictors lm function was used and Multinomial Model was created using R Studio and executed multinom function. Correct formulas were used to return correct results.

(ii) Computing rpart function type attribute was specified as outcome and all others attributes from *bic train* data set were used as predictors. Predict function was used for prediction on test data frame. To evaluate Decision Tree Model performance the mean absolute error (MAE) was used. Correct formulas were used to return correct results.

(iii) To compute Clustering k-means algorithm was used. Through testing process with different numbers of k four clusters have been selected. Correct formulas were used to return correct results.

3.7 Conclusion

In this paper the main factors contributing to increase in severity of road accident occurrence involving cyclist were identified and analysed and the prediction model of severity in accident occurrence was developed in order to support and enhance road safety awareness. Data of Ireland road collisions from 2004 to 2013 was provided by RSA. After cleaning, avoiding duplicate and not available data the data set was used to create the prediction model and compare result in different areas of Ireland. The

most vulnerable groups of cyclists have been identified by age and gender.

Some interesting facts about correlation between the severity of the road accidents involving bicycle users and weather conditions, daylight, road surface conditions, number of vehicles involved, type of collision and road speed limit have been discovered. But due to missing gaps in body of knowledge on this topic there still much to do to improve our roads for bicycle user and to make their trips more safety.

4 Conclusion and Future Work

Characteristics of severity of road accidents occurrence involving bicycle user have received little attention in safety research, especially in rural areas where cyclists represent a small share of all transport modes, only two or three records are registered in some counties of Ireland for ten years. It doesnt look like an actual number of accidents have occurred in these counties. Probably not all of them have been reported and registered. Lack of information does not allow to analyse the full picture of accidents involving cyclist occurred in Ireland in general. Development of the new technologies will help to improve accident data recording system particularly involving cyclists and will provide more accurate data.

This paper hence focused on identifying and analysing the main factors contributing to increase in severity of cyclist accidents occurrence and developing the prediction model of severity in accident occurrence involving bicycle users. Some relationships have been identified by using simple aggregation and creating a cross table with relationship between dependent variable type with one of each independent. It was preliminary concluded that the most unfavourable for bicycle users to cycle with high winds in the dark. To verify this and to predict the risk of having a cycling accident using Ireland road collision data for the years 2004 2013 different models were used.

Linear Regression Model that was applied on Ireland helped to identify the main factors that affect the risk of serious injury of victims in the accidents. Analysing the accident data for the years 2004 2013 providing by RSA somewhat unexpected conclusion was that the most of accidents occurred in good weather and good visibility has been made. Firstly, it because people prefers to cycle in good weather in daytime, secondly, it because people are losing their alertness and become more careless in sunny day.

Using Multinomial Logistics Regression Model the changes in predicted probability associated with light condition were examined. Deterioration of visibility on the roads leads to the fact that the number of minor accidents is not change but a greater number of serious accidents leading to death of victim of accident. Computing Regression Tree revealed that the most important factor that affects the severity of the cyclist accidents is the speed limit. It is evidenced by its correlation with type of severity too (Table 7). It because the speed limit and actual speed of vehicles on the roads is highly positive correlated. In other words, actual speed of vehicle is very close to the speed limit. The next factor is number of vehicles involved in accident, then county where accident has happened, then collision type.

Clustering Model is playing important role in analysing and comparison of severity of cyclist accidents in Dublin and rural areas. Plotting the results of Clustering and analysing it the conclusion can be made that vast majority of accidents and minor severity accidents happened in Dublin area. But more serious accidents occurred more in Dublin too but that is not the vast majority of it. Further research showed that frequency of minor and serious accident occurrence per 1.000 people (Table 6) is the biggest but number of fatal cases per population the smallest. It may be due the

ability to provide the necessary medical assistance to the victims of accident. On the other hand, in Cork percentage of fatal cases from all accident the highest (Figure 8) even higher than in country but the lowest percentage of serious cases. In the country percentage of fatal cases the second highest but it can be explained by the fact that speed limit higher on the country roads, not as good medical infrastructure like in the city.

Relationships between different age and gender groups of cyclists and the risk of getting into the accident were conducted. The result is coincided with results conducted by the RSA analysts two years earlier [7] This research is not without weaknesses and limitations. There are not too many records of severity in road accident occurrence involving cyclist for the years 2004 2013. Because of lack or sparse of data some analyses cannot be conducted. It would be interesting to test the impact of the road width, road type, presence of bicycle lanes, road marking, purposes of trip on severity of cyclists collision. All these analyses are planning to be conducted in future.

Acknowledgements

The author wish to thanks Dr Catherine Mulwa for a great help. The author gratefully acknowledges the assistance of Road Safety Authority (RSA) which provided the data for this study.

References

[1] Mohamed A Abdel-Aty and A Essam Radwan. Modeling traffic accident occurrence and involvement. *Accident Analysis & Prevention*, 32(5):633–642, 2000.

[2] Helai Huang and Mohamed Abdel-Aty. Multilevel data and bayesian analysis in traffic safety. *Accident Analysis & Prevention*, 42(6):1556–1565, 2010.

[3] Joon-Ki Kim, Sungyop Kim, Gudmundur F Ulfarsson, and Luis A Porrello. Bicyclist injury severities in bicycle–motor vehicle accidents. *Accident Analysis & Prevention*, 39(2): 238–251, 2007.

[4] Brett Lantz. *Machine learning with R*. Packt Publishing Ltd, 2013.

[5] Dominique Lord, Simon P Washington, and John N Ivan. Poisson, poisson-gamma and zero-inflated regression models of motor vehicle crashes: balancing statistical fit and theory. *Accident Analysis & Prevention*, 37(1):35–46, 2005.

[6] Fred L Mannering and Chandra R Bhat. Analytic methods in accident research: methodological frontier and future directions. *Analytic Methods in Accident Research*, 1:1–22, 2014.

[7] RSA. Review of cyclist injuries in 2012, 2012. URL `url{http://www.rsa.ie/Documents/Road%20Safety/Crash%20Stats/Review_of_Cyclist_Injuries_2012.pdf}, urldate={2016-12-06}`.

[8] Grégory Vandenbulcke, Isabelle Thomas, and Luc Int Panis. Predicting cycling accident risk in brussels: a spatial case–control approach. *Accident Analysis & Prevention*, 62: 341–357, 2014.

[9] Feng Wei and Gordon Lovegrove. An empirical tool to evaluate the safety of cyclists: Community based, macro-level collision prediction models using negative binomial regression. *Accident Analysis & Prevention*, 61:129–137, 2013.

[10] Qiang Zeng and Helai Huang. Bayesian spatial joint modeling of traffic crashes

on an urban road network. *Accident Analysis & Prevention*, 67:105–112, 2014.

A Appendix 1. Data Preparation

Provided by RSA data consists of one per year table. After ETL process the final table for cyclists consist of 918 observation of 17 variables.

	type	splimit	county	hour	light
type	1.000000000	-0.12778870	0.01211049	0.01750459	-0.006033807
splimit	-0.127788703	1.00000000	0.19292170	0.03571114	-0.022795778
county	0.012110488	0.19292170	1.00000000	0.04898046	0.013444918
hour	0.017504589	0.03571114	0.04898046	1.00000000	0.039107543
light	-0.006033807	-0.02279578	0.01344492	0.03910754	1.000000000

Figure 1: Correlation.

	Var1	Freq
1	Fatal	25
2	Serious	68
3	Minor	825

Figure 2: Severity.

	light	type
1	1	2.866460
2	2	2.904762
3	3	2.929577
4	4	2.842105
5	6	2.615385
6	7	2.945946

Figure 3: Severity and Light.

	weat1	type
1	1	2.890411
2	2	2.850575
3	3	2.500000
4	4	3.000000
5	5	3.000000
6	6	2.250000
7	7	3.000000
8	8	2.857143
9	9	2.000000

Figure 4: Severity and Weather.

B Apendix 2. Requirement Specification

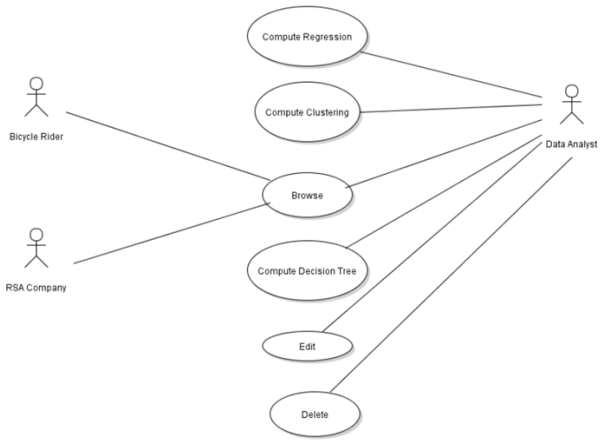

Figure 5: Use Case Diagram

C Apendix 3. Data Warehouse

In the process of Dimensional Modelling the Facts, Dimensions and Grain were chosen and Star Schema was created. Clean data is delivered to the SQL server.

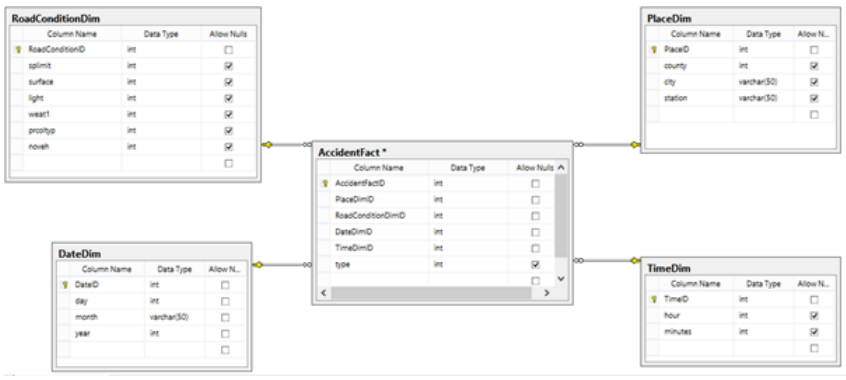

Figure 6: Star Schema

D Apendix 4. Comparison of accidents occurrence in Dublin verses rural areas

Bicycle		Carlow	Cavan	Clare	Cork	Donegal	Dublin	Galway	Kerry	Kildare	Kilkenny	Laois	Leitrim	Limerick	Longford	Louth	Mayo	Meath	Monaghan	Offaly	Roscommon	Sligo	Tipperary	Waterford	Westmeath	Wexford	Wicklow
	Fatal	1	0	1	5	1	6	1	2	0	1	1	0	2	0	0	0	2	0	1	0	0	1	0	0	0	0
	Serious	0	0	1	4	3	35	4	4	1	2	0	1	5	0	2	1	0	1	0	0	0	0	1	0	2	1
	Minor	1	1	14	68	11	457	33	36	17	8	5	0	29	4	33	7	16	6	9	5	3	8	19	4	12	19

Figure 7: Severity of Accidents in Conties of Ireland

	Fatal	FperCent	Serious	SperCent	Minor	MperCent	All	Population	FatalPerPop	SerPerPop	MinorPerPop	AllPerPop
Dublin	6	1.2	35	7	457	91.8	498	1345	0,004	0,026	0.34	0.37
Cork	5	6.5	4	5.2	68	88.3	77	542	0,009	0,007	0.13	0.14
Rest of Ireland	14	4.7	29	9.7	400	67.7	443	2687	0,005	0,011	0.16	0.16

Figure 8: Severity of Accidents in Ireland.