A framework for quantifying and characterising road accident risk: A data mining approach

by

Shane Andrew van Heerden

Dissertation presented for the degree of
Doctor of Philosophy in Industrial Engineering
in the Faculty of Engineering at Stellenbosch University

Promoter: Prof JH van Vuuren

Co-promoter: Prof SS Grobbelaar

March 2020

Declaration

By submitting this dissertation electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the sole author thereof (save to the extent explicitly otherwise stated), that reproduction and publication thereof by Stellenbosch University will not infringe any third party rights and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

Date: March 1, 2020

Copyright © 2020 Stellenbosch University

All rights reserved

Abstract

According to the World Health Organisation, road accidents account for approximately 1.25 million deaths annually — the eighth leading cause of death worldwide. With the enormous losses to society resulting from road accidents, the prevention and severity reduction of road accidents has been an active area of research focus for many decades. Researchers frequently employ a variety of statistical learning techniques in an attempt to understand the factors contributing to higher levels of road accident risk. Such insights provide vital direction for governments with respect to safer road designs and the establishment of countermeasures aimed at reducing the number of road accidents. Furthermore, recent advances in machine learning have presented exciting new machine learning possibilities that were deemed far out of reach just over a decade ago. The tasks associated with data pre-processing in this context are, however, often daunting and immensely time-consuming. Moreover, the adoption of machine learning models in the road accident analysis literature has been relatively limited due to the uninterpretable nature of the majority of these models.

A generic modular data mining framework is, therefore, proposed in this dissertation, aimed specifically at formalising and facilitating the tasks associated with road accident data preparation, and facilitating the interpretation of machine learning model output. This framework is designed to facilitate the *configuration*, *enhancement* and *transformation* of raw accident, vehicle, road and victim data into useful information which appropriately quantifies and characterises road accident risk. More specifically, this framework facilitates evaluation of road accident risk in terms of the rate and severity of being involved in a road accident along road segments and at road junctions based on historically recorded RAs. The configuration procedure in the proposed framework allows a user to format data attributes appropriately, as well as correct any missing or erroneous values that may exist in data sets. The enhancement procedure allows a user to merge vehicle and road records to a corresponding accident record for the purpose of creating an all-encompassing data set. It is also possible to construct new attributes based on current attribute values residing in the aforementioned data sets. After each of the individual data sets has been prepared appropriately and the data are deemed of a sufficiently high quality, they may be stored in a database. Finally, the transformation procedure exploits these high-quality data to quantify the rate and severity of road accidents along road segments or at road junctions. These results serve as input to a standard supervised learning procedure in which road characteristics are used to predict these rate and severity measurements.

In order to demonstrate the practical workability and usefulness of the proposed framework, a concept demonstrator of the framework is implemented in an existing data mining platform and applied to a real-world case study based on road accident data from Greater Manchester in the United Kingdom. Each of the individual data preparation components of the framework is tested in the context of this case study, while the effectiveness of the road accident risk evaluation approaches is demonstrated by means of multiple investigations.

Uittreksel

Volgens die Wêreldgesondheidsorganisasie veroorsaak padongelukke jaarliks sowat 1.25 miljoen sterftes — die agtste hoofoorsaak van sterftes wêreldwyd. Met die enorme verliese vir die samelewing wat padongelukke meebring, is die voorkoming en ernsvermindering van padongelukke vir dekades al 'n aktiewe navorsingsarea. Navorsers gebruik dikwels 'n verskeidenheid statistiese leertegnieke om die faktore wat tot hoër vlakke van padongelukrisiko bydra, te verstaan. Sulke insigte bied belangrike geleenthede vir regerings ten opsigte van die daarstelling van veiliger padontwerpe en teenmaatreëls wat daarop gemik is om die aantal padongelukke te verminder. Verder het onlangse vordering in masjienleer opwindende nuwe masjienleerbenaderings wat net meer as 'n dekade gelede as buite bereik beskou is, moontlik gemaak. Die take wat verband hou met datavoorbereiding in hierdie konteks, is egter dikwels uitdagend en baie tydrowend. Die aanvaarding van masjienleermodelle in die literatuur oor die ontleding van padongelukke was ook betreklik beperk vanweë die oninterpreteerbare aard van die meerderheid van hierdie modelle.

'n Generiese modulêre data-ontginningsraamwerk word dus in hierdie proefskrif voorgestel, wat spesifiek gemik is op die formalisering en fasilitering van die take wat verband hou met die voorbereiding van padongelukdata, asook die fasilitering van die interpretasie van kragtige masjienleer algoritme afvoer. Hierdie raamwerk is ontwerp om die *konfigurasie*, *verbetering* en *transformasie* van rou ongeluks-, voertuig- en paddata na betekenisvolle inligting te fasiliteer wat padongelukrisiko toepaslik karakteriseer en kwantifiseer. Meer spesifiek, hierdie raamwerk fasiliteer die evaluasie van die risiko van padongelukke ten opsigte van die tempo en erns van die betrokkenheid by padongelukke langs padsegmente en by padkruisings gebaseer op historiese padongelukke. Die konfigurasieprosedure in die voorgestelde raamwerk stel die gebruiker in staat om data-eienskappe op gepaste wyse te formateer, sowel as om ontbrekende of foutiewe waardes wat in datastelle mag bestaan, reg te stel. Die verbeteringsprosedure stel die gebruiker in staat om voertuig- en padrekords saam te voeg tot 'n oorkoepelende rekord van padongelukke met die doel om 'n allesomvattende datastel te skep. Die konstruksie van nuwe attribute gebaseer op huidige attribuutwaardes wat in bogenoemde datastelle voorkom, is ook moontlik. Nadat elk van die individuele datastelle toepaslik voorberei is en die inligting as van voldoende hoë gehalte beskou word, kan die data in 'n databasis geberg word. Uiteindelik maak die transformasieprosedure van hierdie hoë-gehalte data gebruik om die tempo en erns van padongelukke langs padsegmente of by padkruisings te bepaal. Hierdie resultate dien as insette tot 'n standaard toesigleerproses waarin padkenmerke gebruik word om hierdie tempo- en ernsmetings te voorspel.

Ten einde die praktiese werkbaarheid en bruikbaarheid van die voorgestelde raamwerk te demonstreer, word 'n prototipe van die raamwerk in 'n bestaande data-ontginningsagteware omgewing geïmplementeer en op 'n werklike gevallestudie toegepas wat gebaseer is op padongelukdata in die Groter Manchester-area van die Verenigde Koninkryk. Elk van die individuele data-voorbereidingskomponente van die raamwerk word in die konteks van hierdie gevallestudie getoets, terwyl die doeltreffendheid van die padongeluk risiko-evalueringsproses deur middel van veelvuldige ondersoeke gedemonstreer word.

Acknowledgements

The author wishes to acknowledge the following people and institutions for their various contributions towards the completion of this work:

- To my supervisor, Prof JH van Vuuren, for the profound impact he has had on me over the past four years. Thank you, Prof, for your valuable guidance, immense patience, and unwavering belief in my abilities. I admire your incredible passion and constant pursuit of excellence, and I sincerely appreciate the momentous time and dedication you have invested in ensuring the work contained in this dissertation is of a high standard. I feel immensely privileged to call you my supervisor, mentor and friend.
- To my co-supervisor, Prof SS Grobbelaar, for initially providing the motivation for my research and consistently being a source of inspiration. I always look forward to our meetings knowing that I'll leave with a renewed ambition for my research.
- To the *Stellenbosch Unit for Operations Research in Engineering* (SUnORE) and the *Health Systems and Innovation* (HS&I) research groups for their financial support in the form of a post-graduate bursary.
- To my friends, Stephan and Ryan, for being my in-office source of motivation, my on demand thesauri, and the soundboard for all my weird and wonderful ideas. I truly value your unwavering friendship and I cannot wait to see what the future has in store for us.
- To my fellow SUnORE colleagues for their friendship, support and high-quality banter throughout the course of this research. I will cherish the incredible memories made during our office hours, tea times and social activities long after this dissertation has been submitted.
- To the Department of Industrial Engineering at Stellenbosch University, together with SUnORE, for the use of their stimulating office space and excellent computing facilities.
- To my fiancée's family for their constant love, support and prayers, especially during the more difficult times in my post-graduate studies.
- To my family for being my continuous source of motivation, moral support and unconditional love. Thank you for your unwavering belief in me throughout my post-graduate studies and constantly encouraging me to pursue my dreams.
- To my best friend and the most incredible fiancée in the whole world, Victoria. Thank you for your never-ending encouragement, for accompanying me during many late nights at the office, and for always looking after me. Thank you for always being my voice of reason, my everyday adventure and my number one supporter. Your beautiful heart and overflowing love is all a guy could ever ask for. I truly cannot wait for a lifetime of adventures with you by my side.

- To the King of Kings, Lord of Lords — for His continual guidance, abundant grace and love, and for blessing me with the ability and passion to seek His truth, as proclaimed in Psalm 111:2 — "Great are the works of the Lord, studied by all who delight in them."

Table of Contents

Abstract **iii**

Uittreksel **v**

Acknowledgements **vii**

List of Acronyms **xv**

List of Figures **xvii**

List of Tables **xxiii**

List of Algorithms **xxv**

1 Introduction **1**

1.1 Background 1

1.2 Problem statement 4

1.3 Dissertation objectives 5

1.4 Dissertation scope 7

1.5 Dissertation organisation 8

I Literature review **11**

2 Road safety modelling approaches **13**

2.1 Road safety data 13

2.2 Descriptive road safety modelling approaches 14

2.2.1 The fundamental components of a road accident 14

2.2.2 Measuring the risk of road accidents 15

2.2.3 Factors influencing the risk of road accidents 17

2.3 Predictive road safety modelling approaches 18

2.3.1 Data and methodological issues . . . 20
2.3.2 Predictive modelling methods for analysing accident frequency data . . . 22
2.4 Reflections on approaches adopted in road safety modelling . . . 24
2.5 Chapter summary . . . 24

3 Network spatial analysis 27
3.1 Constructing a GIS model of the real world . . . 27
3.1.1 Conceptual modelling approaches . . . 28
3.1.2 Data representation . . . 30
3.2 Modelling a network space mathematically . . . 31
3.2.1 An object-based model for networks . . . 31
3.2.2 A field-based model for networks . . . 32
3.2.3 The binomial point process . . . 32
3.2.4 The uniform network transformation . . . 34
3.2.5 Edge biases on bounded networks . . . 35
3.3 Network point pattern analysis methods . . . 37
3.3.1 Measuring uniformity of point distributions in a network . . . 37
3.3.2 Estimating point density in a network . . . 41
3.4 Chapter summary . . . 46

4 Machine learning 47
4.1 An overview of machine learning . . . 48
4.1.1 Machine learning paradigms . . . 48
4.1.2 The supervised learning procedure . . . 49
4.2 Data considerations . . . 50
4.2.1 Data taxonomy . . . 51
4.2.2 Data cleaning . . . 52
4.2.3 Data partitioning methods . . . 54
4.3 Model considerations . . . 55
4.3.1 Performance measures . . . 56
4.3.2 Overfitting *vs* underfitting the training data . . . 57
4.3.3 Model performance *vs* model interpretability . . . 59
4.4 Generalised linear models . . . 61
4.4.1 General linear models . . . 61
4.4.2 The exponential family . . . 62
4.4.3 Constructing a generalised linear model . . . 63

4.4.4 Logistic regression . . . 64
4.5 Decision tree learning . . . 65
4.5.1 Classification and regression trees . . . 66
4.5.2 Random forests . . . 68
4.5.3 The C4.5 algorithm . . . 69
4.6 The naive Bayes classifier . . . 70
4.7 The k-nearest neighbours algorithm . . . 71
4.8 Support vector machines . . . 72
4.8.1 Linear separability . . . 73
4.8.2 Linear inseparability . . . 76
4.8.3 The application of kernels . . . 77
4.9 The multi-layer perceptron model . . . 79
4.10 Chapter summary . . . 80

II System design 81

5 Road accident risk evaluation framework 83
5.1 The data mining process . . . 83
5.2 A high-level modular data mining framework . . . 86
5.3 The proposed RARE framework . . . 88
5.3.1 Configuring the data appropriately . . . 89
5.3.2 Combining multiple data sets . . . 91
5.3.3 Creating new data attributes . . . 92
5.3.4 Converting data into information . . . 93
5.4 Chapter summary . . . 94

6 Framework module components 95
6.1 Formatting raw data . . . 95
6.2 Configuring data attributes . . . 97
6.3 Combining vehicle records to accident records . . . 103
6.4 Combining road records to accident-vehicle records . . . 107
6.5 Creating severity attributes . . . 113
6.6 Constructing a derived attribute . . . 115
6.7 Modelling road accident risk along a road segment . . . 116
6.8 Modelling road accident risk at a road junction . . . 118
6.9 Predicting road accident risk . . . 122

6.10 Chapter summary 124

III Case study **125**

7 Framework development and data preparation **127**

7.1 Development of the proposed RARE framework 127

7.2 Accident and road data 128

7.3 Configuring the data 130

7.4 Combining the data 134

7.5 Creating new data 138

7.6 Chapter summary 138

8 Data modelling and evaluation **139**

8.1 Analysis of road accident risk along road segments 140

8.1.1 Parameter choices for descriptive analysis of road segments 140

8.1.2 Spatial analysis of road segments 141

8.1.3 Non-spatial analysis of road segments 147

8.1.4 Parameter choices for predictive analysis of road segments 150

8.1.5 Predictive analysis of road segments 151

8.2 Analysis of road accident risk at road junctions 155

8.2.1 Parameter choices for descriptive analysis of road junctions 155

8.2.2 Spatial descriptive analysis of road junctions 156

8.2.3 Non-spatial descriptive analysis of road junctions 156

8.2.4 Parameter choices for predictive analysis of road junctions 160

8.2.5 Predictive analysis of road junctions 162

8.3 Predictive analysis of road accident severity 163

8.3.1 Road accident severity at crossroads controlled by traffic lights 166

8.3.2 Road accident severity on a motorway 169

8.4 Chapter summary 170

IV Conclusion **173**

9 Dissertation summary & appraisal **175**

9.1 Dissertation summary 175

9.2 Appraisal of dissertation contributions 178

10 Future work **181**

10.1 Improvements to the current RARE framework 181

10.2 Additions to the RARE framework 182

10.3 Promoting the RARE framework 185

References **187**

A GitHub repository contents **205**

B Accident, vehicle and victim categories **207**

C Road accident data collection form **215**

D UK Department for Transport terminology **221**

E Category spatial scores **227**

F Road segment spatial representation **231**

G Road segment SHAP values **233**

H Dissimilarity threshold and similarity measure empirical study **245**

I Road junction spatial representation **267**

J Road junction SHAP values **271**

K Road accident SHAP values at crossroads **283**

L Road accident SHAP values on motorways **291**

List of Acronyms

AADT: Average Annual Daily Traffic

AFA: Additive Feature Attribution

AI: Artificial Intelligence

API: Application Programming Interface

AUC: Area Under Curve

CA: Classification Accuracy

CART: Classification And Regression Trees

CRISP-DM: Cross-Industry Standard Process for Data Mining

CSR: Complete Spatial Randomness

DFD: Data Flow Diagram

EDA: Exploratory Data Analysis

FN: False Negatives

FP: False Positives

FPR: False Positive Rate

GIS: Geographical Information System

GLM: Generalised Linear Model

GBP: Great British Pounds

HCA: Hierarchical Cluster Analysis

IKL: Infinite Kernel Learning

KDE: Kernel Density Estimation

kNN: k-Nearest Neighbours

MLE: Maximum Likelihood Estimation

ML: Machine Learning

MLP: Multi-layer Perceptron

MKL: Multiple Kernel Learning

NND: Nearest Neighbour Distance

OHSAS: Occupational Health and Safety Assessment Series

OLSR: Ordinary Least Squares Regression

PPA: Point Pattern Analysis

RBF: Radial Basis Function

ReLu: Rectified Linear unit

RA: Road Accident

RAR: Road Accident Risk

RARE: Road Accident Risk Evaluation

ROC: Receiver Operating Characteristic

SANET: Spatial Analysis along NETworks

SVM: Support Vector Machine

TN: True Negatives

TP: True Positives

TPR: True Positive Rate

UK: United Kingdom

USA: United States of America

List of Figures

1.1 Henry Ford and his first automobile 1
1.2 The Ford Model T and the assembly line used for its mass production 2
1.3 A time line of the number of RA deaths in the USA 3

2.1 The three primary components of an RA 15
2.2 The proportion of RAs according to the three primary RA components 15
2.3 Comparison of the road safety problems in Sweden and South Africa 16
2.4 Annual fatalities according to the different transportation modes 17
2.5 Three dimensions of road safety using two different exposure measures 18

3.1 Modelling the real world in a geospatial sense 28
3.2 Object-based model framework 29
3.3 Vector and raster representations 30
3.4 A probability density function on a network 33
3.5 The probability integral transformation 35
3.6 The uniform network transformation 36
3.7 A naturally and artificially bounded network 36
3.8 The set of points within a distance k from a point p 38
3.9 The locations of various beauty parlours alongside roads 40
3.10 The number of beauty parlours satisfying the NND under the CSR hypothesis . 41
3.11 Five points distributed along a line segment 42
3.12 Multiple kernel density functions on a network embedded in a plane 42
3.13 A decomposition of a network into the sub-networks 43
3.14 An equal-split continuous kernel density function 46

4.1 The supervised learning process 50
4.2 A taxonomy of the various data types 52
4.3 Sample data partitioned into the training, validation and testing sets 54

4.4 The k-fold cross-validation procedure 55
4.5 The bootstrap re-sampling procedure 55
4.6 Two distribution separations used to produce an ROC curve 58
4.7 The notion of underfitting or overfitting 58
4.8 SHAP values example 60
4.9 Splitting a CART parent node into corresponding child nodes 66
4.10 The change in misclassification rate according to a tree's size 67
4.11 A two-dimensional Euclidean distance-based Voronoi tessellation 72
4.12 Support vectors and the corresponding induced hyperplanes 73
4.13 Vectors of the SVMs learning model 74
4.14 A set of linear non-separable observations 76
4.15 A demonstration of the kernel trick 78

5.1 The six phases of the CRISP-DM reference model 84
5.2 Schematic representation of a generic data mining framework 86
5.3 An entity-relationship diagram between accident, vehicle and road entities 88
5.4 The Level 0 Diagram of the proposed RARE framework 90
5.5 The flow of data between modules in the Configure class 91
5.6 The flow of data between modules in the Combine class 92
5.7 The flow of data between modules in the Create class 93
5.8 The flow of data between modules in the Convert class 94

6.1 The flow of data between sub-modules in Module 1.0 96
6.2 The flow of data between sub-modules in Module 2.0 98
6.3 The locations of accidents occuring on a Monday and at a roundabout 100
6.4 The observed and expected network G-function 101
6.5 The flow of data between sub-modules in Module 3.0 104
6.6 The flow of data between sub-modules in Module 4.0 107
6.7 Defining the region $\mathcal{D}^{(k)}$ 109
6.8 The eight possible orientations of a line segment 110
6.9 A scenario where a point $\boldsymbol{p}^{(i+1)}$ does not fall in regions $\mathcal{D}^{(k)}$ or $\mathcal{D}^{(k+1)}$ 111
6.10 The flow of data in Module 5.0 113
6.11 A representation of nominal and ordinal attribute categories 114
6.12 A plot of the function x^{η} 115
6.13 The flow of data in Module 6.0 115
6.14 The flow of data between sub-modules in Module 7.0 117

6.15 The flow of data between sub-modules in Module 8.0 119
6.16 A hypothetical example of accident points occurring at two road juncions 119
6.17 Steps for defining a road junction 121
6.18 The flow of data between sub-modules in Module 9.0 123

7.1 The RARE framework concept demonstrator 129
7.2 The Greater Manchester accidents and roads data set 131
7.3 The spatial scores of all accident attributes 134
7.4 The frequency of collapses in the recursive ordering algorithm 136
7.5 The roads classified according to accident count 137

8.1 The effects of the maximum subinterval length choice 140
8.2 Accident rate and severity along road segments 142
8.3 A zoomed-in view of accident severity in the Manchester city centre 143
8.4 The Bridge Street and Deansgate intersection in Manchester 144
8.5 The road segment attribute values in Greater Manchester 147
8.6 Accident rate *vs* severity scatter plot of the `RoadType` categories 148
8.7 Accident rate *vs* severity scatter plot of the `SpeedLimit` categories 148
8.8 Accident rate *vs* severity scatter plot of the `UrbanOrRuralArea` categories 149
8.9 SHAP values for the MLP trained to predict RA rate along segments 153
8.10 SHAP values for the MLP trained to predict RA severity along segments 154
8.11 Accident rate and severity at road junctions 157
8.12 The road junction attribute values in Greater Manchester 159
8.13 Accident rate *vs* severity scatter plot of the `JunctionDetail` categories 159
8.14 Accident rate *vs* severity scatter plot of the `JunctionControl` categories 160
8.15 Accident rate *vs* severity scatter plot of the `PedestrianCrossing` categories . . . 161
8.16 SHAP values for the MLP trained to predict RA rate at junctions 164
8.17 SHAP values for the MLP trained to RA predict severity at junctions 165
8.18 ROC plot of seven learning models predicting accident severity at crossroads . . 167
8.19 SHAP values of RF model predicting accident severity at crossroads 168
8.20 ROC plot of seven learning models predicting accident severity on motorway . . 170
8.21 SHAP values of RF model predicting accident severity on motorway 171

10.1 Lotteries associated with choosing road A or road B 183
10.2 A hypothetical network composed of eight links and nine possible routes 184
10.3 A land cover map in a raster format 185

C.1 Page 1 of the STATS19 accident report form 216
C.2 Page 2 of the STATS19 accident report form 217
C.3 Page 3 of the STATS19 accident report form 218
C.4 Page 4 of the STATS19 accident report form 219

F.1 The road segment attribute values in Greater Manchester 232

G.1 SHAP values for the LR model trained to predict RA rate along segments 234
G.2 SHAP values for the DT model trained to predict RA rate along segments 235
G.3 SHAP values for the RF model trained to predict RA rate along segments 236
G.4 SHAP values for the kNN model trained to predict RA rate along segments . . . 237
G.5 SHAP values for the SVM model trained to predict RA rate along segments . . . 238
G.6 SHAP values for the LR model trained to predict RA severity along segments . . 239
G.7 SHAP values for the DT model trained to predict RA severity along segments . . 240
G.8 SHAP values for the RF model trained to predict RA severity along segments . . 241
G.9 SHAP values for the kNN model trained to predict RA severity along segments . 242
G.10 SHAP values for the SVM model trained to predict RA severity along segments . 243

H.1 Junction clusterings for the Lin similarity measure 252
H.2 Junction clusterings for the Goodall similarity measure 259
H.3 Junction clusterings for the Occurrence Frequency similarity measure 266

I.1 The road junction attribute values in Greater Manchester 269

J.1 SHAP values for the LR model trained to predict RA rate at junctions 272
J.2 SHAP values for the DT model trained to predict RA rate at junctions 273
J.3 SHAP values for the RF model trained to predict RA rate at junctions 274
J.4 SHAP values for the kNN model trained to predict RA rate at junctions 275
J.5 SHAP values for the SVM model trained to predict RA rate at junctions 276
J.6 SHAP values for the LR model trained to predict RA severity at junctions 277
J.7 SHAP values for the DT model trained to predict RA severity at junctions . . . 278
J.8 SHAP values for the RF model trained to predict RA severity at junctions 279
J.9 SHAP values for the kNN model trained to predict RA severity at junctions . . . 280
J.10 SHAP values for the SVM model trained to predict RA severity at junctions . . 281

K.1 SHAP values for the OR model trained to predict RA severity at crossroads . . . 284
K.2 SHAP values for the NB model trained to predict RA severity at crossroads . . . 285
K.3 SHAP values for the CT model trained to predict RA severity at crossroads . . . 286

K.4 SHAP values for the kNN model trained to predict RA severity at crossroads . . 287
K.5 SHAP values for the SVM model trained to predict RA severity at crossroads . . 288
K.6 SHAP values for the MLP model trained to predict RA severity at crossroads . . 289

L.1 SHAP values for the OR model trained to predict RA severity on motorways . . 292
L.2 SHAP values for the NB model trained to predict RA severity on motorways . . 293
L.3 SHAP values for the CT model trained to predict RA severity on motorways . . 294
L.4 SHAP values for the kNN model trained to predict RA severity on motorways . . 295
L.5 SHAP values for the SVM model trained to predict RA severity on motorways . 296
L.6 SHAP values for the MLP model trained to predict RA severity on motorways . 297

List of Tables

2.1 The categorisation of various risk factors according to the Haddon matrix 19

3.1 The most popular choices of kernel functions 45

4.1 Examples of erroneous and missing data values 53
4.2 A confusion matrix 56

6.1 The set of compulsory attributes to be defined by the user 99
6.2 Categorising attributes according to the Haddon matrix 103

7.1 The attributes and data types of the accident data set 132
7.2 The attributes and data types of the vehicle data set 132
7.3 The attributes and data types of the road data set 133
7.4 The attributes and data types of the victim data set 133
7.5 Classification of attributes according to the Haddon matrix 135

8.1 Predictive performance of six learning models for road segments 151
8.2 Predictive performance of six learning models for road junctions 162
8.3 Performance of seven learning models predicting accident severity at crossroads . 167
8.4 Performance of seven learning models predicting accident severity on motorways 169

B.1 The categories of the various accident attributes 207
B.2 The categories of the various vehicle attributes 209
B.3 The categories of the various victim attributes 213

D.1 Terminology adopted by the UK Department for Transport 221

E.1 The spatial scores associated with the various attribute categories 227

List of Algorithms

6.1 Ordering vehicle records 106
6.2 Projecting a point onto its closest line segment 112

CHAPTER 1

Introduction

Contents

1.1 Background 1
1.2 Problem statement 4
1.3 Dissertation objectives 5
1.4 Dissertation scope 7
1.5 Dissertation organisation 8

1.1 Background

Henry Ford, depicted in Figure 1.1(a), was arguably the greatest and most influential industrialist of his time. Ford rose to fame in 1896 after unveiling his gasoline-powered "Quadricycle," shown in Figure 1.1(b), which was one of the first self-propelled, horseless carriages [73]. After receiving the support of Thomas Edison and the capital investment of a wealthy Detroit lumber merchant, Ford went on to found the Ford Motor Company in 1903 [32] with the vision of producing an automobile for the masses — an automobile that would stand outside every American home.

(a) (b)

FIGURE 1.1: *(a) A portrait of Henry Ford [81] and (b) his first automobile, the Quadricycle [52].*

(a) (b)

FIGURE 1.2: *(a) The Ford Model T [265] and (b) a moving assembly line used for vehicle mass production [128].*

After five years of perfecting his design, Ford would realise his dream through the creation of the Ford Model T, shown in Figure 1.2(a) — an automobile that would signify an attainable status for the rising middle class and become a powerful symbol of the age of modernisation in America [80]. The immediate commercial success of this automobile was predominantly due to its ease of operation, maintenance and handling, together with its astonishing affordability for the average American citizen [81]. In order to meet the overwhelming demand, Ford introduced revolutionary mass-production methods which included the use of large production plants, standardised and interchangeable parts, significantly high employee wages, and a moving assembly line, like that shown in Figure 1.2(b) [109]. Years after Ford's death in 1947, he is still widely remembered as the man who, more than any other, initiated a new era in personal transport and ushered in the rapid advancements of the 20th century.

Nowadays, the global automobile industry is worth approximately 1.9 trillion dollars [253]. Annual vehicle production has grown rapidly since the introduction of the Ford Model T, reaching just under 90 million vehicles world-wide in 2018 [251]. In 2015, it was estimated that approximately 1.3 billion commercial and passenger automobiles are on the roads world-wide [252], and this number is only expected to increase as a result of newly industrialised countries like China and India [206]. This significant increase in vehicle ownership has led to expansive developments in the global road transportation network. To date, approximately 39.5 million kilometres of roadway have been constructed world-wide [39]. To put this number into perspective, the distance between the earth and the moon is only 384 400 kilometres [182]. Although human life has been improved considerably by the automobile, one should also be aware of the immediate consequences associated with this fast-paced world humanity has created. According to the World Health Organisation, *road accidents* (RAs) account for approximately 1.25 million deaths world-wide and an additional 50 million injuries annually [272]. Road accidents are ranked as the 8th leading cause of death world-wide [295] and the leading cause of death among people aged 15–29 [272]. It is perhaps surprising to note that approximately 94% of all RAs are attributed to driver error [273]. In the *United States of America* (USA) alone, the annual cost of RAs amounts to a staggering $836 billion [282].

With the enormous losses to society resulting from RAs, the prevention and severity reduction of these occurrences has been an active area of research focus for many decades [144]. The fruits of these tireless research efforts can be clearly seen in Figure 1.3, where the total annual RA

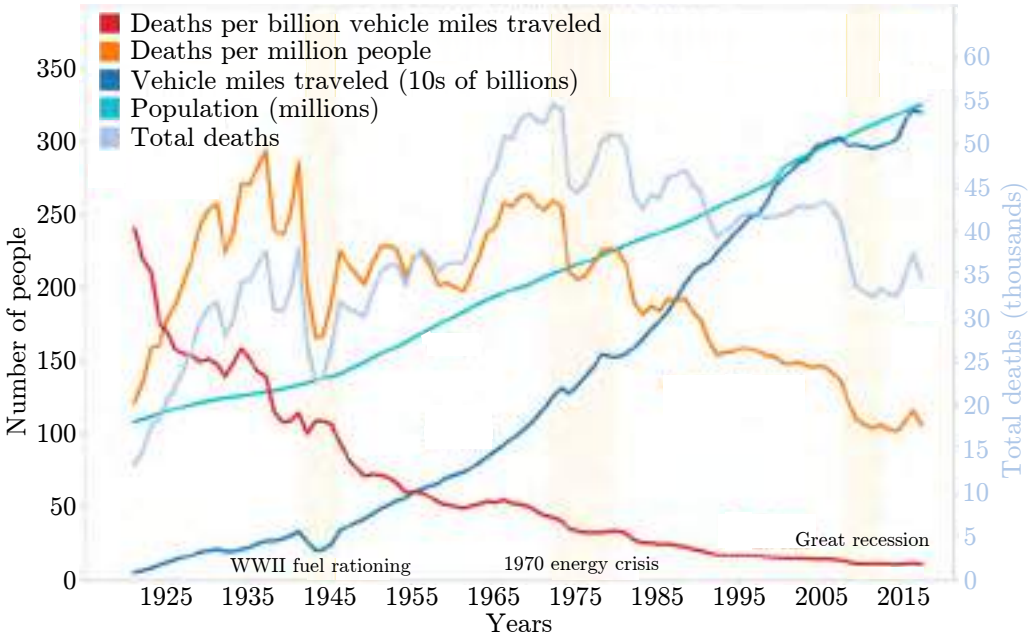


FIGURE 1.3: *A time line showing the number of RA deaths in the United States of America during the period 1921–2017 [24].*

deaths in the USA is plotted for the years 1921–2017. As may be seen in this figure, the total RA deaths per vehicle mile travelled (indicated by the red curve) has substantially decreased ever since the introduction of the automobile. The reason for this significant progress is twofold: First, researchers have provided automobile manufactures with valuable recommendations over the years as to how they can improve the safety design of manufactured vehicles, and, secondly, researchers have been able to draw insights from RA data which has placed them in a position to provide direction for governments with respect to policy formulation, road design and countermeasures aimed at reducing the number of RAs [274]. Recent advancements in *Artificial Intelligence* (AI), together with a substantial increase in the computational power and memory storage capacity of computers, has presented researchers pursuing these two intertwined fields of research with exciting opportunities which were deemed far out of reach just over a decade ago. In the case of the former, researchers are attempting to completely remove the human driving component by developing fully autonomous vehicles using powerful reinforcement learning techniques [61]. In the case of the latter, researchers are now able to build better, more powerful analytical models, utilising the large amounts of RA data gathered over the years, to gain a better understanding of where and how RAs occur. Although the introduction of commercially available fully autonomous vehicles is fast approaching [282], it is predicted that complete adoption of this revolutionary technology is expected (optimistically) to take approximately five decades [141]. Consequently, there is a significant need for researchers to develop more advanced analytical models which may ultimately supplement the advanced autonomous vehicle systems of the future.

Researchers pursuing this line of work are typically referred to as *data scientists* who commonly engage in the art of *data mining*[1]. Data mining is best described as the process of identifying and analysing patterns and relationships in large complex data sets using quantitative tools and methods residing in the intersection between the research areas of database systems, statis-

[1]The term "data mining" is, in fact, a misnomer as it is often confused with the extraction (*i.e.* mining) of data itself [100]. It is also a buzzword [262] frequently misused to describe large-scale data or information processing.

tics and *machine learning* (ML) [40, 47]. This interdisciplinary sub-field of computer science ultimately aims to extract useful information through various data manipulation tasks and intelligent methods from one or more sets of data so that it may be used to inform decision making [263]. This typically requires the design and development of descriptive, predictive and/or prescriptive analytical models which attempt to answer the questions "*what has happened*," "*what could happen*" and/or "*what should happen*," respectively [53]. Apart from analytical modelling, the data mining process comprises a variety of sub-processes concerned with data preparation, database management and information visualisation [40]. Data mining is, therefore, often described as a creative process as it requires the successful integration of analytical tools and domain knowledge for the purpose of answering (often challenging) questions [290].

In the road transportation domain, one such question is how one can analytically model the risk presented by RAs along a road network. This is not entirely obvious as the notion of risk is, for the most part, a highly subjective concept which leads to a wide variety of interpretations and definitions, and depends largely on the a specific context under consideration. As a result, there is no universally accepted definition of risk [98]. Nonetheless, the *Occupational Health and Safety Assessment Series* (BS OHSAS 18001), a British standard for occupational health and safety management systems, defines risk as a *"combination of the likelihood of a hazardous event or exposure and the severity of injury or ill health that can be caused by the event of exposure"* [27]. Echoing this definition, Aven and Renn [7] argue that all expressions of risk can be partitioned into one of two categories: Either as (1) a probability or expected value, or (2) as an event, consequences or uncertainty. This view of risk aligns with the two central burdens associated with RAs, namely: (1) The damage or loss of property, and (2) the threat to or loss of human life.

Furthermore, in the wake of the 4th industrial revolution, there exists significant opportunity in the road transportation domain for the development of disruptive technologies which may be realised through the process of data mining. A data mining solution which is able to answer the question "*To how much and what type of risk are individuals exposed on the roads?*" is one which would likely appeal to a wide variety of stakeholders. Governments may, for example, benefit from a system which is able to identify those points within its road networks that exhibit high accident frequencies/rates. The identification of these so-call accident "hotspots" may provide vital direction with respect to safer road designs and countermeasures aimed at reducing the number of RAs. Such a system is, however, not anticipated only to be limited to government decision support. Many vehicle insurance companies are already installing tracking devices in their customers' vehicles and quoting premiums based on how well their customers drive [48]. Consequently, vehicle insurance companies may also benefit from a system that is able to quantify and characterise the risk experienced by each of their customers on the roads so that this can be translated into monthly premiums. The application of such a system may also stretch far beyond insurance. The public at large typically relies on routing applications (*e.g.* Google Maps [89]) to prescribe the fastest or shortest route from a specified starting point to an intended destination. Hence, society may also benefit from a system that is able to provide navigation along routes which minimise the risk of being involved in an RA. Moreover, the endorsement of such a system by insurance companies may lead to incentive schemes which persuade customers to drive along safer routes.

1.2 Problem statement

The primary research aim in this dissertation is to design, develop and demonstrate the practical workability of a generic framework for producing risk evaluations in the context of RAs. This

road accident risk evaluation (RARE) framework should be able to take as input raw accident and road data supplied by the user and facilitate the configuration, enhancement and transformation of these data sets into useful information which can quantify and characterise the *road accident risk* (RAR) experienced by road users appropriately. More specifically, this framework should evaluate RAR in terms of the rate and severity of being involved in an RA when travelling along road segments or passing through road junctions based on historically recorded RAs. In other words, the aim of the RARE framework is to *quantify* and *characterise* the *rate* and *severity* of an RA along road *segments* and road *junctions*, as visually illustrated in Figure 1.4(a), where eight distinct sectors are formed by a binary partitioning complex according to the six aforementioned terms, as shown in Figure 1.4(b). This framework is to be designed by following a data mining approach through the incorporation of various techniques and methods residing in the fields of network spatial analysis and ML, while also drawing from traditional approaches employed in the road accident analysis literature.

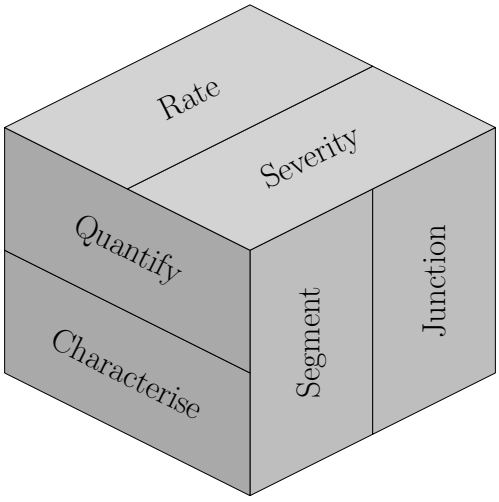


FIGURE 1.4: *An illustration of the elements of the RARE framework for quantifying and characterising the rate and severity of an RA along road segments and road junctions.*

1.3 Dissertation objectives

The following eight objectives are pursued in this dissertation:

I To *conduct* a comprehensive investigation into the literature related to this study. More specifically:

(a) To *understand* the various components and factors of a road transport system which contribute to the occurrence of RAs,

(b) To *identify* the most notable descriptive and predictive risk evaluation frameworks and models applied in the literature in respect of road safety,

(c) To *review* the approaches employed in the field of *geographical information systems* (GIS) when constructing and representing models of reality,

(d) To *describe* the mathematical techniques and considerations applied when modelling a network space, together with the statistical methods employed during network spatial analysis investigations,

(e) To *report* on the common ML paradigms, together with practical considerations associated with preparing data and evaluating the performance of ML models, and

(f) To *document* well-known ML models that may be used to draw knowledge and insights from data.

II To *design* a generic RARE framework that is modular in nature and capable of configuring, enhancing and transforming data provided by a user into a useful risk evaluation. This framework should be designed in accordance with the major stages associated with data mining processes.

III To *develop* and *document* the individual data preparation and risk evaluation module components of the RARE framework of Objective II by borrowing from the techniques and methods uncovered in the literature review in pursuit of Objective I. More specifically:

(a) To *establish* various components for facilitating data configuration based on the literature reviewed in fulfilment of Objectives I(a), (c) and (e),

(b) To *devise* specialised components capable of combining information from multiple independent data sets,

(c) To *include* components for creating new data according to specific algorithmic procedures,

(d) To *design* specialised components for defining an appropriate set of road segment and road junction observations,

(e) To *develop* a quantitative risk evaluation of RAs by employing notions from the existing RA analysis literature uncovered in pursuit of Objective I(b), and the network spatial analysis methods of Objective I(d) for estimating point density along a given road network, and

(f) To *implement* a standard supervised learning procedure, using the results of Objective II(e) and the techniques and methods identified in pursuit of Objective I(f), to characterise the RAR experienced by road users.

IV To *put forward* a computer implementation of a concept demonstrator of the proposed RARE framework of Objective II by implementing existing and specially developed components, as described in Objective III.

V To *obtain* and *document* an adequate set of real-life accident and road data which is deemed appropriate for conducting a realistic case study in the context of the proposed RARE framework of Objective II.

VI To *prepare* the data obtained in pursuit of Objective V by means of the data preparation components developed in fulfilment of Objective IV. Additionally, this step should serve as a form of verification of the components designed in pursuit of Objectives III(a)–(c).

VII To *verify* and *validate* the proposed working of the RARE framework of Objective II by conducting eight RAR investigations (corresponding to the eight sectors of Figure 1.4) using the data prepared in pursuit of Objective VI and the individual components of Objectives III(d) and (e), developed in fulfilment of Objective IV.

VIII To *identify* further additions or improvements to the RARE framework which may be pursued in the form of follow-up future work.

1.4 Dissertation scope

Due to the complexities associated with developing a completely generic framework, together with the wide variety of modelling approaches which may be adopted when attempting to model RAR, the scope of the RARE framework proposed in dissertation is limited by the following assumptions:

Road accident data. RA data are captured in many countries world-wide [296]. Upon the occurrence of an RA, legislation typically requires the parties involved to report the accident, in which case an accident report form is usually completed, detailing the events of the accident. These data are commonly used to conduct statistical analyses for the purpose of identifying areas that require the implementation of more effective safety control policies and informing the general public of the road safety situation faced on their country's roads. The aforementioned data commonly comprise multiple historical RA records with specific attributes describing three key entities: (1) Details of the accident, (2) details of the vehicles involved and (3) details of the resulting victims. These data are typically organised in the form of a relational database in which an accident record may comprise one or more vehicle records which may, in turn, be associated with one or more victim records. It is assumed, in this dissertation, that accident and corresponding vehicle and victim data are logicically associated and structured according to a tabular format. It is also assumed that the victim data set contains an attribute describing the severity of the RA victims involved.

Road network data. Road data describing the geographical layout of the road network of a country are often freely available. In the majority of traditional road safety approaches for describing RAR, however, the frequency of accidents on a road have to be normalised according to some measure of vehicle exposure. It is, therefore, assumed that the set of road data contains an annual measure of the traffic volume counts experienced on each road segment.

Geographical representation. Attributes describing the physical representation or location of an object in the real world are essential in the field of GIS. Since the possibility of applying various network spatial analysis methods to accident and road data is explored in this dissertation, it is assumed that, in the case of RA data, the geographical position of each accident record is known. It is, furthermore, assumed that, in the case of road data, each road record is geographically approximated by a single straight line segment at an appropriate level of abstraction, and the spatial position of each RA is defined by a single geographical point.

Modelling approach. Over the years, researchers have proposed a wide variety of modelling approaches in an attempt to qualify and/or quantify road safety. Kikkonen and Peltola [169] classified these modelling approaches into four distinct categories, namely: (1) *Descriptive models*, (2) *predictive models for aggregated data*, (3) *risk models for non-aggregated data* and (4) *accident consequence models*. In this dissertation, a descriptive modelling approach is pursued for the purpose quantifying RAR, which aims to unify ideas from the literature on traditional road safety analysis and network spatial analysis. Furthermore,

a predictive modelling approach is adopted for the purpose of characterising RAR, which draws from both the road safety and ML literature. Consequently, the investigation of models developed according to other scientific disciplines, such as psychology, sociology, ergonomics, medicine, biometrics and physics, are considered outside the scope of this dissertation.

1.5 Dissertation organisation

Excluding the present introductory chapter, this dissertation contains a further nine chapters which are partitioned into four separate parts. Part I, comprising Chapters 2–4, contains a comprehensive literature review in which a thorough investigation of the topics stated in Objective I of §1.3 is conducted. More specifically, Chapter 2 is devoted to a review of the pertinent literature related to the traditional approaches employed when attempting to model road safety. A high-level perspective of the road transport system is first described so as to identify the major components that contribute to RAs, in fulfilment of Objective I(a). Thereafter, existing road safety frameworks and models are reviewed in pursuit of Objective I(b) in order to gain a better understanding of the most popular approaches adopted by researchers when attempting to conduct a risk assessment in the context of RAs. The focus then shifts in Chapter 3 to a review of the methodologies typically employed in the field of GIS so as to present the reader with an overview of conceptual modelling approaches and data representations typically employed when conducting spatial analyses, in fulfilment of Objective I(c). In that same chapter, Objective I(d) is also pursued by investigating techniques, methods and considerations employed when conducting a spatial analysis in the context of a network space. The focus again shifts in Chapter 4 to the literature pertaining to ML techniques in general so as to inform the reader about common learning paradigms, as well as practical considerations associated with appropriately preparing data and evaluating the performance of ML models within these paradigms, in fulfilment of Objective I(e). Finally, the literature study concludes with a description of some well-known ML models commonly adopted in the paradigm of supervised learning, as per Objective I(f).

Part II of this dissertation comprises Chapters 5 and 6, and is concerned with the design of the generic RARE framework according a top-down design methodology. In Chapter 5, the major stages concerned with the data mining process are first established according to a standard data mining reference model. Thereafter, this reference model is used to formulate a high-level data mining framework encompassing all the major stages of the data mining process. This data mining framework is subsequently used as the basis for designing a high-level architectural overview of a generic RARE framework, in fulfilment of Objective II. As previously eluded to, this RARE framework comprises multiple modules which are collectively responsible for configuring, enhancing and transforming RA data into a useful RAR assessment. The formulation and integration of the individual module components concerned with data preparation, in fulfilment of Objectives III(a)–(c), and risk evaluation, in pursuit of Objectives III(d)–(f), are formally described and documented in Chapter 6. The required data flows between these individual modules and sub-module constituents of the various framework components are explained by means of various diagrams.

Part 3 of this dissertation comprises Chapters 7 and 8, and is a concept demonstration in which the RARE framework of Objective II is applied to a case study involving real-life RA data. In Chapter 7, Objective IV is first pursued by developing a concept demonstrator of the proposed RARE framework for the purpose of demonstrating the practical workability of the system, which is achieved by using existing and specially developed implementations of the module components of Objective III. This software implementation is developed in the Python 3.7.0

programming language, and formalised in the Orange Data Mining software suite. The focus of Chapter 7 then shifts to a description and assessment of the real-life data obtained for the case study investigation, in fulfilment of Objective V. Thereafter, the data are appropriately configured and enhanced according to the processes set out in the RARE framework, in pursuit of Objective VI. The data are then used in Chapter 8 to conduct eight case study investigations (corresponding to the eight sectors of Figure 1.4) whereby the RAR pertaining to various data attributes are appropriately assessed in pursuit of Objective VII.

The final part of this dissertation, Part IV, comprises Chapters 9 and 10. Chapter 9 contains a summary and critical appraisal of the contributions of the dissertation, while possible improvement opportunities which may be pursued in the form of future follow-up work are described in Chapter 10, in fulfilment of Objective VIII.

Part I

Literature review

CHAPTER 2

Road safety modelling approaches

Contents

2.1 Road safety data 13
2.2 Descriptive road safety modelling approaches 14
2.2.1 The fundamental components of a road accident 14
2.2.2 Measuring the risk of road accidents 15
2.2.3 Factors influencing the risk of road accidents 17
2.3 Predictive road safety modelling approaches 18
2.3.1 Data and methodological issues 20
2.3.2 Predictive modelling methods for analysing accident frequency data . . 22
2.4 Reflections on approaches adopted in road safety modelling 24
2.5 Chapter summary 24

The aim in this chapter is to provide a description of the pertinent literature related to the current approaches employed when conducting road safety investigations. The chapter opens in §2.1 with a brief description of the types of data typically relied upon by road safety researchers. Thereafter, the most popular descriptive approaches for describing the current road safety situation in a country, as well as identifying the contributing factors of an RA, are formally described in §2.2. This is followed, in §2.3, by a review of various predictive models that have been applied in the road safety domain, together with associated data and methodological issues which must be considered during a predictive investigation. A reflection on the road safety modelling literature as a whole is then proffered in §2.4 so as to provide direction for the remainder of this dissertation. The chapter finally closes in §2.5 with a brief summary of its contents.

2.1 Road safety data

In order to construct descriptive or predictive road safety models, a researcher must first acquire an appropriate set of data related to road safety. There are typically two sources of data in road safety analyses, namely *accident data* and *exposure data* [169]. The primary source of accident data are captured by official RA reports collected by police officers when responding to the scene of an accident, while secondary sources may be in the form of insurance company claims, hospital records and accident involvement surveys [268].

Although accident data are the basic information required in any road safety analysis, significant findings are often realised only when accident data are combined with exposure data [169]. There

are many different ways to define RA exposure. One may, for example, define exposure according to the distance travelled, time spent in traffic or number of trips taken by a road user. Exposure data are, typically, not commonly collected exclusively for road safety purposes, but are rather captured for road transport and economic development planning [169]. Developed countries typically employ two methods at a national level to estimate various measures of exposure to their road transport network. These may take the form of *travel habit/vehicle usage surveys* and/or employing specialised *traffic counting systems*. The former requires surveying a sample of vehicle owners about their driving patterns. The information thus obtained is then used to estimate exposure measures in person-kilometres, vehicle-kilometres, travel time and/or number of trips. The latter, on the other hand, is typically employed to estimate the *annual average daily traffic* (AADT) on individual road sections which can, in turn, be used to estimate a vehicle-kilometres measure of exposure [169].

2.2 Descriptive road safety modelling approaches

There exists a wide variety of descriptive methods in the literature for analysing and presenting RA data. For the purposes of this study, however, only three of the most popular descriptive approaches are considered, which identify: (1) The primary components of an RA, (2) the measurements employed for quantifying RAR, and (3) the factors which may influence RAR. More specifically, the first approach attempts to gain a better understanding into the fundamental components of the road transport system which are typically the causes of RA occurrences. The second approach attempts to form a more holistic view of the road safety situation in a country by describing RAR according to three principle measurements. And finally, the first and second approach are unified in the third approach which represents an injury prevention framework for classifying the contributing factors of RAs.

2.2.1 The fundamental components of a road accident

The primary function of road transport systems involves the movement of people or goods by means of human-driven motorised vehicles from their current locations to various intended destinations. This is facilitated by the construction of an extensive road network which connects many locations of interest together. A road transport system may be described according to three fundamental components, namely its *human*, *vehicle*, and *environment* components. In this system, a vehicle's driver is responsible for controlling each vehicle's movement on a road so that it avoids a collision with another person, vehicle or stationary object. An undesired consequence of using such a system is that a collision (or an RA) may, in fact, take place, which may be attributed to a failure of the appropriate interaction between or within one, two or all three of the aforementioned components. Consequently, an RA may be classified into one of the seven categories illustrated in the Venn diagram of Figure 2.1.

Over the years, numerous studies have been conducted in order to understand which components (or combination of components) of an RA are typically to blame for their occurrence, as such knowledge is expected to facilitate a strategic focus of RA prevention and mitigation efforts. In these studies, a finite number of RAs are investigated so as to determine their contributing factor(s). The results of five independent studies[1] are presented in Figure 2.2 according to the Venn diagram format of Figure 2.1. As may be seen, in all of these studies, the human

[1] In the case where the components considered in a study did not conform to the categories of Figure 2.1, they were consolidated accordingly. Cases where the results do not total to 100% are presumably due to rounding errors.

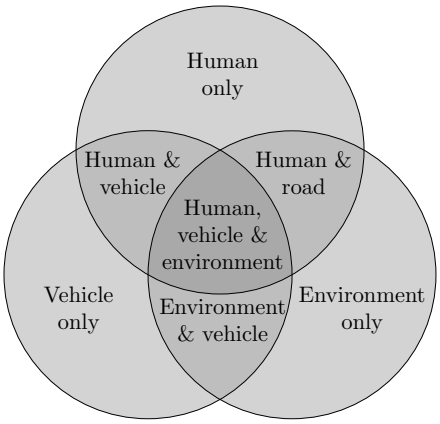
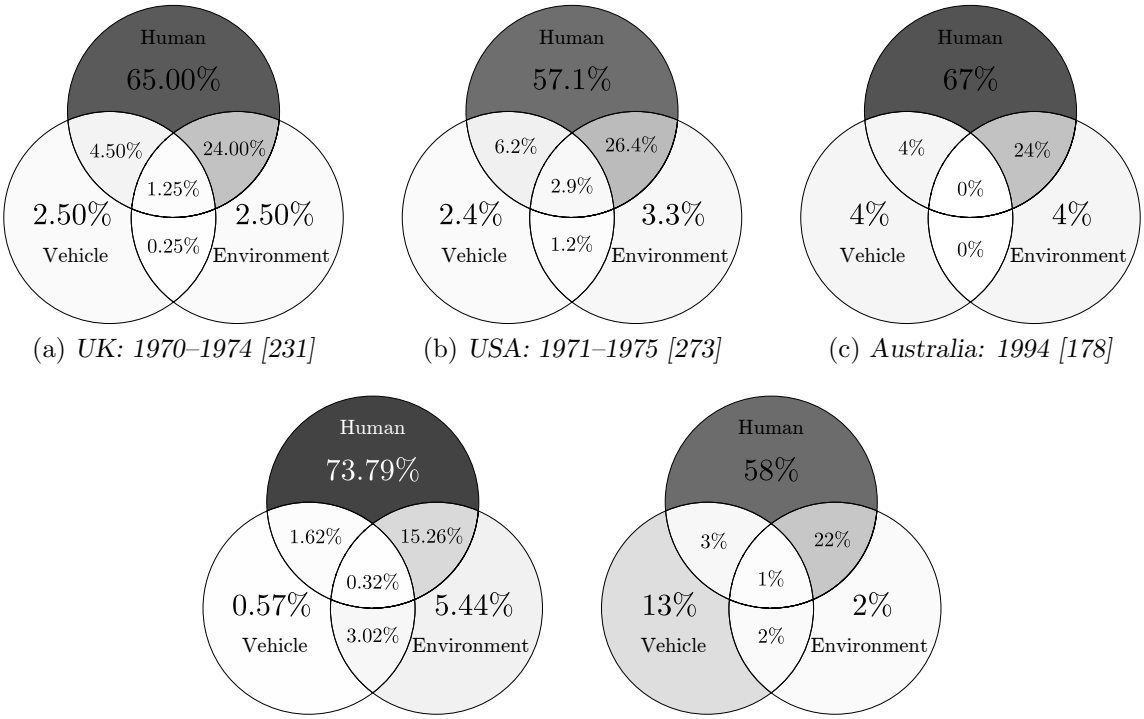


FIGURE 2.1: *A Venn diagram showing the various components of an RA.*



(a) *UK: 1970–1974 [231]* (b) *USA: 1971–1975 [273]* (c) *Australia: 1994 [178]*

(d) *UK: 2007–2010 [220]* (e) *India: 2012–2014 [119]*

FIGURE 2.2: *Results from five independent studies showing the proportion of RAs attributed to one or more of the components of an RA in the format of Figure 2.1.*

component was identified as the leading cause of RAs, and the human-environment component as the second leading cause. These results suggest that RA prevention and mitigation efforts should be focused on reducing the occurrence of human error and improving the interaction between a driver and the road infrastructure design.

2.2.2 Measuring the risk of road accidents

Arguably, the most popular descriptive road safety approach is to describe the road safety situation of a country according to the *three dimensions of road safety.* Rumar [229] claims that the road traffic situation can be described according to three principal dimensions which

attempt to quantify: (1) The *exposure* of a road user to the road transport system, (2) the *risk*[2] of a road user being involved in an RA, and (3) the *consequence* of a road user being involved in an RA. Given these three dimensions, road safety is often described mathematically as

$$\text{Road safety situation} = \text{Exposure} \times \text{Risk} \times \text{Consequence}, \tag{2.1}$$

where the risk of an RA is defined as

$$\text{Risk} = \frac{\text{Number of accidents}}{\text{Exposure}}$$

and the consequence of an RA is defined as

$$\text{Consequence} = \frac{\text{Number of fatalities}}{\text{Number of accidents}}$$

[169]. Consequently, (2.1) can collectively be illustrated graphically by means of a rectangular prism, as shown in Figure 2.3. This representation is often preferred as it simplifies the understanding of the traffic safety situation [186], where a comparatively larger rectangular prism volume indicates a more severe road safety situation.

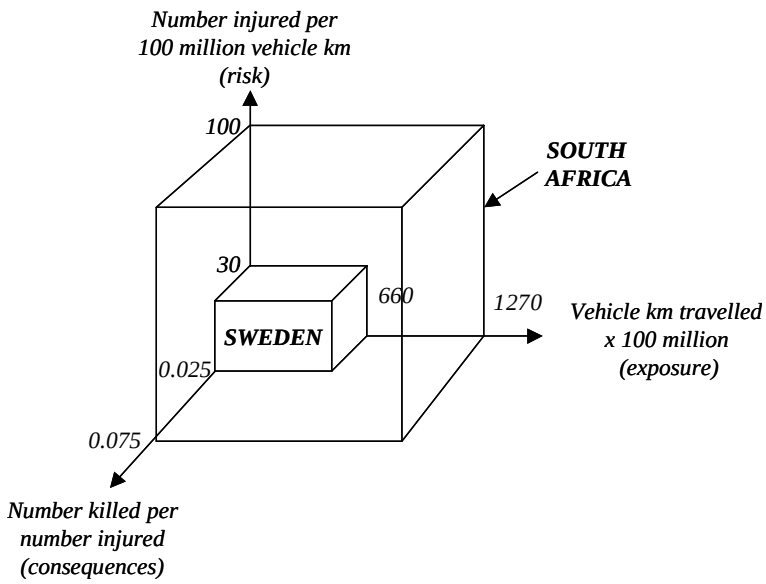


FIGURE 2.3: *Comparison of the road safety problems in Sweden and South Africa according to the three dimensions of road safety [6].*

This theory is further illustrated in Figure 2.4, where the road safety situation is described in a disaggregated manner according to the different modes of transport that existed in Sweden during the period 1990–1992. In this case, the width of a rectangular prism represents exposure (measured as the proportion of total person-kilometres travelled), its height represents risk (measured in number of people injured per million people exposed) and its depth represents the consequence (measured in number of people killed per people injured). As may be seen in the figure, motorcyclists represent a very small proportion of the total person-kilometres travelled by the Swedish public, but exhibit the highest probability of being injured in an RA. It can also be seen that RAs involving pedestrians typically result in comparatively higher fatalities than RAs involving other road users.

[2]Although the term *risk* is widely used in the literature when describing the three dimensions of road safety, in this dissertation, the term *risk* is used explicitly to represent the overall negative effects of RAs. One may, in this section, substitute the term *risk* as the *probability* of experiencing an RA.

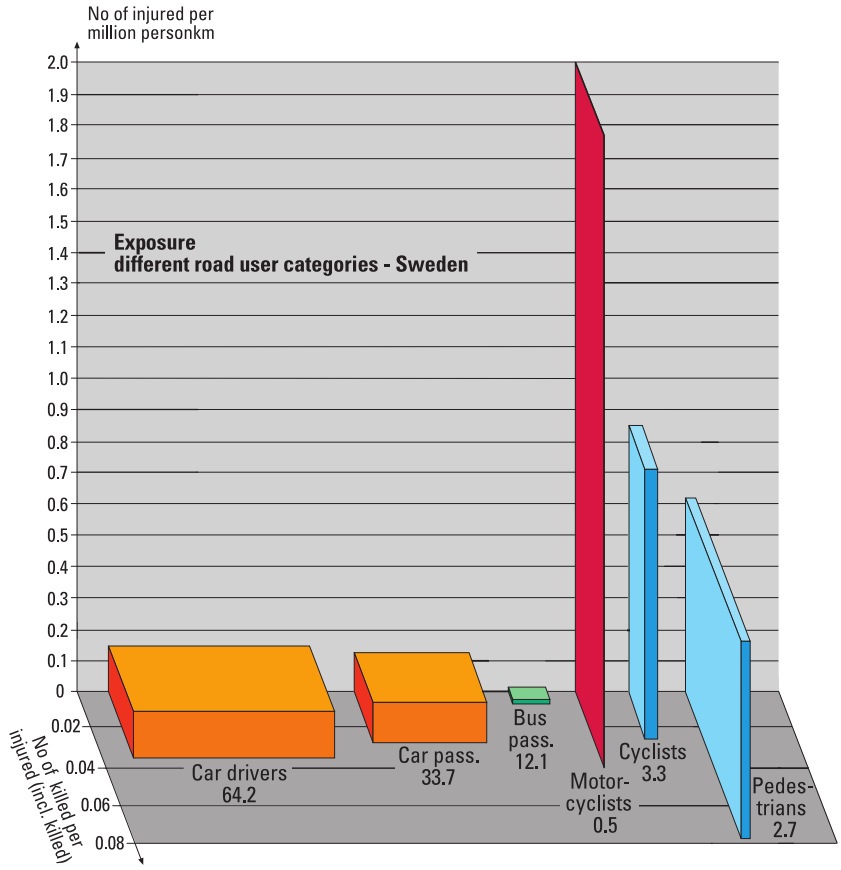
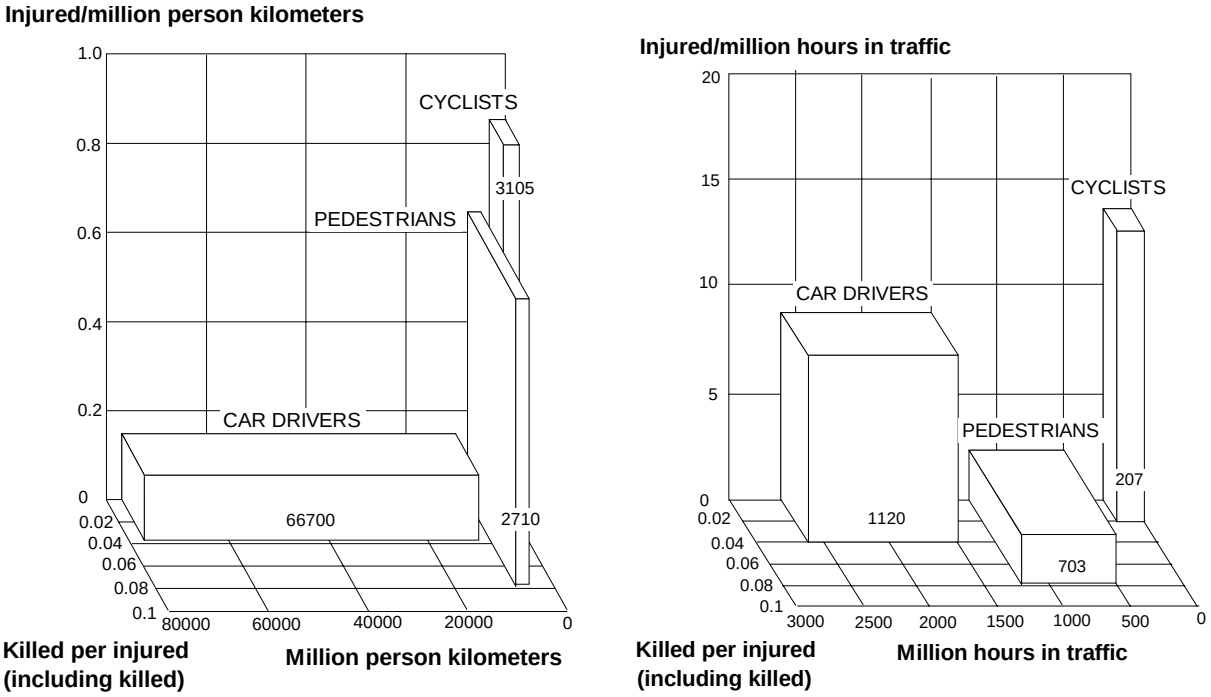


FIGURE 2.4: *The average number of annual fatalities according to the different transportation modes in Sweden during the period 1990–1992 [169].*

Although these results provide an informative description of the road safety situation, investigators comparing categories in this manner should be aware of the unit of exposure employed. Mikkonen and Peltola [169] warned that the choice of exposure unit employed in an investigation may produce different descriptions of the overall RAR. This notion is illustrated in Figures 2.5(a) and 2.5(b), where exposure is measured in person-kilometres (*i.e.* the number of kilometres driven on the road by the Swedish public) and person-hours (*i.e.* the number of hours spent on the road by the Swedish public), respectively, for vehicle drivers, cyclists and pedestrians. As may be seen in Figure 2.5(a), if exposure is measured in person-kilometres, a vehicle driver appears to be less prone to be involved in an RA than a pedestrian or a cyclist. This is because the number of kilometres travelled by a vehicle (*i.e.* its exposure) is typically significantly larger than that of a pedestrian or cyclist. If, on the other hand, exposure is measured in person-hours, as shown in Figure 2.5(b), the probability that a pedestrian is involved in an accident appears significantly lower than those of vehicle drivers and cyclists. In the case where a comparison is drawn between the different transportation modes, however, person-kilometres is often the preferred exposure unit [169]. The consequence (*i.e.* the depths of the cubes), on the other hand, is the same in both Figures 2.5(a) and 2.5(b) since it is not a function of exposure.

2.2.3 Factors influencing the risk of road accidents

The most commonly employed tool in the field of injury prevention is the *Haddon matrix* [181]. Developed in 1970 by William Haddon [97], this evaluation tool is routinely used to identify and evaluate the relative importance of different risk factors [202]. This is achieved by categorising



(a) (b)

FIGURE 2.5: *The three dimensions of road safety for the cases where exposure is measured in (a) person-kilometres and (b) person-hours in Sweden for the year 1992 [169].*

factors which may contribute to an RA in a matrix format, where rows describe the factors occurring before (*pre-accident*), during (*accident*) and after (*post-accident*) an RA, while columns represent the three components of an RA (*i.e. human*, *vehicle* and *environment*), as described in §2.2.1. The respective rows of the Haddon matrix may also be interpreted as the factors affecting the *exposure*, *risk* and *consequence* of an RA [186], as presented in §2.2.2. Once multiple factors have been identified and categorised accordingly, like those shown in Table 2.1, appropriate countermeasures may then be developed and prioritised for implementation over short-, medium- and long-term periods [294].

2.3 Predictive road safety modelling approaches

Lord and Mannering [144] pointed out that the acquisition of detailed driving data (*e.g.* acceleration, braking and steering information) would facilitate the identification of cause and effect relationships in the context of RAs. Unfortunately, this type of data is typically seldom available. Researchers are, therefore, commonly limited to the spatial and temporal data captured by official RA report forms completed by police officers responding to accident scenes, while secondary sources are in the form of insurance company claims, hospital records and accident involvement surveys are sometimes available [268]. Consequently, many road safety researchers have framed their analytic approaches towards discovering the contributing factors that lead to higher numbers of RAs on some road entity (typically along a road segment or at a junction) over some specified time period [144]. This frequency-based view of RA risk is most commonly adopted as it allows for the application of count-data statistical regression models that properly account for the integer nature of the data [144].

Table 2.1: *The categorisation of various risk factors according to the Haddon matrix [45, 97, 230].*

	Human	**Vehicle**	**Environment**
Pre-accident (Exposure)	• Driver impairment – Alcohol – Drugs – Fatigue – Distraction – Stress level • Driver ability – Knowledge – Experience – Vision – Reaction time – Attitude – Habits	• Vehicle roadworthiness – Lights – Breaks – Tires – Handling • Vehicle design – Seat belts – Centre of gravity – Roll protection – Anti-lock brakes – Electronic stability control	• Road design – Intersection type – Road curvature – Shoulder width – Signals and signs – Lighting • Road condition – Marking visibility – Surface grip – Hazards – Construction • Environmental – Weather – Time of day
Accident (Risk)	• Occupant circumstance – Restraint use – Age – Medical condition – Position in vehicle	• Vehicle design – Air bags – Functional restraints – Energy absorption – Interior surface hazards • Vehicle circumstance – Speed of travel – Load weight – Point of impact	• Road infrastructure – Guard rails – Road shoulders – Median dividers • Road hazards – Fixed objects – Roadside ditch
Post-accident (Consequence)	• Victim response ability – Age – Injury severity – Physical ability – Protective clothing – Impairments – Communication ability	• Vehicle design – Fuel system integrity – Vehicle flammability – Ease of vehicle evacuation – Crash alert system	• Emergency response – Emergency communication – Road congestion – Distance to trauma care – Wreckage extrication • Medical capacity – On-site equipment – Paramedic skill – Rehabilitation programme

2.3.1 Data and methodological issues

Over the years, many important data and methodological issues have been identified by researchers working with accident frequency data. It has been shown that these issues may lead to erroneous predictions and incorrect inferences related to the attributes of an RA. A description of the most prominent data and methodological issues identified in the road safety literature are as follows:

Over-dispersion. A common accident frequency data consideration is that of *over-dispersion*, which occurs when the variance of accident counts on a road entity exceeds the count mean significantly. This is typically problematic in the case of the most popular count-data modelling approach, namely the Poisson regression model, which requires equality between the mean and variance. Use of this model in the presence of over-dispersion may result in biased and inconsistent parameter estimates which may, in turn, lead to erroneous inferences of the RA attributes [36, 155, 161, 165, 201].

Under-dispersion. A more uncommon accident frequency data consideration is that of *under-dispersion* which occurs when the accident count mean on road entities exceeds the variance significantly. This typically occurs when the sample-mean value is relatively low which, again, may lead to biased and inconsistent model parameter estimates [189].

Time-varying explanatory variables. Since accident frequency data are typically gathered over a specified time period, there may be situations in which the values of RA attributes change significantly over this time period. Ignoring the potential variation in explanatory variables during a period may therefore result in the loss of potentially important explanatory information [144]. Consider, for example, the case in which one were to model the accident frequency along a road segment per month and precipitation is an associated RA attribute. In this case, the distribution of precipitation over a month is likely to be highly influential in generating RAs, but typically the researcher only has aggregated precipitation data at his or her disposal. Important information is thus lost when employing discrete time intervals (where larger intervals result in more information loss) which, again, may result in biased and inconsistent model parameter estimates [285].

Temporal and spatial correlation. In an attempt to mitigate the information lost due to time-varying attributes, accident data are often considered over smaller time intervals. One may have obtained a years' worth of accident data, for example, and may have partitioned these data into twelve monthly observations in order to determine the number of RAs per month on some road entity. This road entity will, however, generate multiple observations which are correlated over time, since many of the attributes describing inherent properties of the road remain the same. In the same way, there may be a correlation in geographical space when road entities are in close proximity to one another. From a statistical perspective, this is known to affect the precision of a model's parameter estimates adversely [148, 179, 240, 248, 280, 285].

Low sample-mean and small sample size. Due to the rarity of RAs and the significant costs associated with the data collection process, accident data typically only contain a small number of observations. This means that there are often no recorded RAs for a specific road entity. Such data, characterised by a small sample size and a low sample mean, may lead to parameter estimation problems in traditional statistical count-data models [83, 143, 155, 161, 204, 292].

Injury-severity and crash-type correlation. Accident data records are often categorised according to the maximum injury severity sustained by an involved victim (*e.g.* no injury, possible injury, non-incapacitating injury, incapacitating injury or fatal) or the collision type (*e.g.* rear-end, t-bone, head-on, rollover, sideswipe or single-vehicle run-off-the-road). The most common modelling approach is to consider the frequency of RAs (*i.e.* all accident severities and collision types together) [37, 135]. Some road safety researchers have, however, sought to develop separate accident frequency models for the various accident severities and collision types. This leads to high degrees of correlation among these two attribute values and, consequently, means that a more complex model structure is required to account for this cross-model correlation [12, 16, 85, 152, 153, 168, 201, 249].

Under-reporting. Minor RAs which did not lead to any sustained injuries are often not reported to the police and are rather handled by the conflicting parties [5]. This typically leads to a bias towards more serious RA occurrences in an accident data set due to the under-reporting of less severe RAs. It has also been shown that the probability of an RA being reported can vary, not only as a function of accident severity, but also as a function of other factors such as the age of the victim, whether the victim was a driver, passenger or pedestrian, and the number of vehicles involved in the RA [105, 115]. This unavoidable fact may lead to biases in count-data models if under-reporting is not considered in the model estimation process [130, 151].

Omitted-variables bias. It is often tempting to simplify a model by selecting only a subset of the explanatory variables (*e.g.* considering an attribute describing the junction detail as the only explanatory variable). Leaving out important RA attributes may, however, result in biased parameter estimates which could lead to erroneous inferences [285], especially in cases where the omitted variables are highly correlated with the variables(s) included [144].

Endogenous variables. There may be cases in which the explanatory variables depend on the accident frequency, referred to as *endogenous variables*. Consider, for example, the problem of determining the effectiveness of ice-warning signs in reducing the frequency of ice-related accidents. If one were to develop an accident frequency model, a possible explanatory attribute would be the presence or absence of ice-warning signs. Ice-warning signs are, however, more likely to be placed at road entities which have historically high numbers of ice-related RAs, and are, therefore, endogenous in nature [37]. Ignoring this fact will lead to a bias in the parameter estimates of a model [144].

Functional form. A model's functional form is a critical element in the modelling process as it dictates the relationship that may be represented between explanatory and dependent variables. Most count-data models assume a linear relationship between explanatory and dependent variables which may not necessarily be the case in reality. A few road safety researchers have suggested that non-linear functions are able to better characterise the relationships between accident frequencies and explanatory variables [20, 166]. These functions are, however, often fairly complex and may require more advanced parameter estimation procedures [144].

Fixed parameters. In traditional statistical modelling approaches, parameter estimates should not vary across observations. In the case of predicting accident frequency, this implies that the effects of the explanatory variables must be the same for all road entity observations. If, for example, a road entity is a road segment, there may be unobserved variations between these observations which may lead to biased parameter estimates (and possible erroneous inferences) if the model is estimated as if they were fixed [144].

2.3.2 Predictive modelling methods for analysing accident frequency data

There is a wide variety of methods available in the literature for analysing accident frequency data. The most popular of these methods are based on the *generalized linear model* (GLM) framework which has explicit theoretical foundations, and can produce interpretable coefficients for each explanatory variable included in the GLM [297]. There have also been some recent investigations in which supervised learning models have been applied to predict accident frequency. A description of the most prominent predictive models in road safety research are as follows:

The Poisson regression model. Since accident frequency data are, by definition, non-negative integers, the application of *Ordinary Least Squares Regression* (OLSR) is not appropriate as it assumes a continuous dependent variable. For this reason, early RA researchers used the Poisson regression model to predict the number of accidents at a road entity [145, 155, 165, 167, 188, 189], which may be described mathematically as

$$P(y_i) = \frac{\exp(-\lambda_i)\lambda_i^{y_i}}{y_i!}, \tag{2.2}$$

where $P(y_i)$ is the probability of road entity i containing y_i RAs per time period and the Poisson parameter λ_i for road entity i is the expected number of RAs per year, $\mathrm{E}[y_i]$. The Poisson parameter λ_i typically specified by the functional form $\lambda_i = \exp(\boldsymbol{\beta}\boldsymbol{x}_i)$, where $\boldsymbol{x}_i$ is a vector of explanatory variables and $\boldsymbol{\beta}$ is a vector of estimated parameters [144]. Although the Poisson regression model served as a starting point for predicting accident frequency, RA researchers have found that accident data often exhibit characteristics which rendered the application of this simplistic model problematic. More specifically, Poisson regression models cannot account for over- and under-dispersion, and often produce biases when accident data exhibit low sample means and small sample sizes [144].

The negative binomial regression model. The negative binomial regression (or Poisson-gamma regression) model was proposed as an alternative model to overcome the over- and under-dispersion data-handling inabilities of the Poisson regression model [74, 108, 145, 155, 165, 171, 188, 189, 203, 239]. The negative binomial regression model assumes that the Poisson parameter λ_i in (2.2) follows a gamma probability distribution by specifying the functional form of the Poisson parameter as $\lambda_i = \exp(\boldsymbol{\beta}\boldsymbol{x}_i + \epsilon_i)$, where ϵ_i is a gamma distributed error term with a mean of 1 and a variance of α. This facilitates mathematical manipulation of the relationship between the distribution's mean and variance as $\mathrm{Var}(y_i) = \mathrm{E}[y_i](1 + \alpha\mathrm{E}[y_i]) = \mathrm{E}[y_i] + \alpha\mathrm{E}[y_i]^2$. The negative binomial regression model is arguably the most frequently employed model in accident frequency prediction. The model does, however, have some limitations, the most notable being its inability to handle under-dispersed data like the Poisson regression model and that it often produces biases when accident data exhibit low sample means and small sample sizes [146].

The Poisson-lognormal model. Some road safety researchers have also proposed using the Poisson-lognormal model as an alternative to the negative binomial model for predicting accident frequency [2, 147, 166]. The Poisson-lognormal model is almost identical to the negative binomial regression model, the only difference being that the Poisson parameter λ_i is lognormal-distributed (as opposed to being gamma-distributed). The Poisson-lognormal model offers more flexibility than the negative binomial regression model but still has its limitations. In particular, the model parameter estimation process is more complex since the Poisson-lognormal distribution cannot be represented in closed form. Additionally,

a Poisson-lognormal model often produces biases when accident data exhibit low sample means and small sample sizes like the negative binomial regression model, [166].

Zero-inflated Poisson and negative binomial model. Zero-inflated models have been proposed to combat data characterised by many zeros which, as previously described, often cause biases when employing Poisson or negative binomial regression models. Zero-inflated models account for the multiple zero count observations by employing a splitting strategy which models an accident-free versus accident-prone propensity of a road entity. The probability that a road entity is either in a zero or non-zero state is calculated by means of a binary logit (or probit model) [132, 285]. Since its inception, both zero-inflated Poisson and negative binomial models have been popular among RA researchers [37, 131, 135, 241, 242]. Although being widely applicable, especially in the case where accident data exhibit a low sample mean, other road safety researchers have criticised the application of these models. Lord *et al.* [143, 149] argued that, since the accident-free state has a long-term mean of zero, zero-inflated models cannot properly reflect the accident data generating process.

Conway-Maxwell-Poisson model. The Conway-Maxwell-Poisson distribution was first introduced by Conway and Maxwell [49] in 1962 for modelling queues and service rates, and serves as a generalisation of the Poisson distribution. A noteworthy advantage of the Conway-Maxwell-Poisson distribution is that it can handle both under- and over-dispersed data — a fact that has peeked the interests of road safety researchers predicting accident frequency. Lord *et al.* [149] showed that the Conway-Maxwell-Poisson model is comparable to the negative binomial regression model when over-dispersed accident frequency data is present, but superior in the case of under-dispersed accident frequency data. It should, however, be noted that the Conway-Maxwell-Poisson model can be negatively influenced by low sample means and small sample sizes like the negative binomial regression model [144].

Gamma model. Oh *et al.* [189] proposed the gamma model for analysing accident frequency data characterised by either under- and over-dispersion. Although this model exhibits good predictive capabilities, the dual-state nature of the model results in one of the states having a long-term mean of zero, and, like in the case of zero-inflated models, cannot properly reflect the accident data generating process [144]. The gamma model has seen limited usage in road safety research since its introduction.

Support vector machine models. *Support Vector Machine* (SVM) models are based on statistical learning theory and have more recently been applied in accident frequency prediction [138]. SVM models are a set of related supervised learning methods that can be used for both classification and regression tasks, and possess the well-known ability of being able to approximate any multivariate function to a specified degree of accuracy [122]. A general criticism of SVM models is that they operate as "black boxes" in that they do not provide interpretable parameters as in the case of traditional accident frequency models [144]. SVM model parameters are also more complex to estimate and cannot be generalised to other data sets.

Multi-layer perceptron models. *Multi-layer perceptron* (MLP) (or feed-forward neural network) models are function approximators defined by a multi-level network structure consisting of a series of nodes linked together in a hierarchical manner by multiple weight factors [139]. MLP models have also recently been applied in accident frequency prediction primarily as a prediction tool [1, 42, 221, 297]. Overall, these models have been

shown to produce superior non-linear approximation properties when compared to traditional count-data statistical regression models. Again, a major criticism of MLP models are their inability to provide interpretable model results [144].

2.4 Reflections on approaches adopted in road safety modelling

As mentioned in §2.3, when developing predictive road safety models, researchers typically adopt a frequency-based view of RAR as it allows for the application of count-data statistical regression models which properly account for the integer nature of the data. This frequency-based view of RAR commonly aligns with the interests of governments who are often concerned with implementing safety control policies aimed at reducing the total number of RAs in their country. It has, however, been argued that the frequency-based view of RA risk may not be the only view worth considering. The three dimensions of road safety, presented in §2.2.2, provide inspiration for two complementing views, namely: Viewing RAR according to (1) a normalised accident frequency (or accident rate) and (2) as a severity of RAs. If an analyst were to adopt the perspective of a road-user, he or she may be more inclined to subscribe to these two aforementioned views of RAR since it aligns with the two central burdens associated with RAs, namely: (1) The damage or loss of property, and (2) the threat to or loss of human life. It is for this reasons that the author has decided to pursue a rate and severity view of RAR in this dissertation, from both a descriptive modelling perspective and a predictive modelling perspective.

As also stated in §2.3, many road safety researchers frame their analytic approaches towards discovering the contributing factors which lead to higher numbers of RAs on some road entity over some specified time period. This road entity may exist in one of two forms, namely: (1) As a road segment or (2) as road junction. During a predictive modelling investigation, road safety researchers typically resort to manual data aggregation in order to construct a data set that is sufficiently large so that the predictive model employed does not suffer from the effects of small sample sizes, as described in §2.3.2. It is, therefore, proposed that specialised geoprocessing techniques are developed to automate the aggregation of RAs on both road segments and at road junctions.

Despite their inability to provide interpretable model results, both SVM and MLP models have superior predictive performance when compared to traditional count-data statistical regression models, as described in §2.3.2. This fact may be attributed to the non-parametric nature of these models which allows them to approximate both linear and non-linear relationships in accident frequency data. If, however, a predictive model is not constrained by the count-nature formulation of RAR or the strict data distribution assumptions, this opens the door to investigating a host of other supervised learning models. It is for this reason that supervised learning has drawn the attention of the author for further investigation and implementation as an appropriate predictive modelling approach in this dissertation.

2.5 Chapter summary

The purpose of this chapter was to introduce the reader briefly to the most notable traditional descriptive and predictive approaches adopted in the road safety literature when attempting to evaluate the risk posed by RAs. The chapter opened in §2.1 with a description of the type of data typically relied upon by road safety researchers. Thereafter, the most popular descriptive approaches for describing the current road safety situation in a country, as well as identifying

the contributing factors of an RA, were formally described in §2.2. A description of the three main components of an RA was first presented, together with the results of five independent studies — all of which identified the human and human-environment components as the leading causes of RAs. Thereafter, the well-known three dimensions of road safety was introduced and elaborated upon. The Haddon matrix served to unify the three dimensions of road safety and the components of a RA into a single framework for categorising the various RAR factors. The focus then shifted in §2.3 to a review of various predictive models that have been applied in the road safety domain, together with associated data and methodological issues which must be considered during a predictive investigation. The chapter closed in §2.4 with a critical reflection of the road safety modelling approaches adopted in the literature with a view to provide a glimpse of the modelling approach adopted in this dissertation.

CHAPTER 3

Network spatial analysis

Contents

3.1 Constructing a GIS model of the real world 27
3.1.1 Conceptual modelling approaches *28*
3.1.2 Data representation *30*
3.2 Modelling a network space mathematically 31
3.2.1 An object-based model for networks *31*
3.2.2 A field-based model for networks *32*
3.2.3 The binomial point process *32*
3.2.4 The uniform network transformation *34*
3.2.5 Edge biases on bounded networks *35*
3.3 Network point pattern analysis methods 37
3.3.1 Measuring uniformity of point distributions in a network *37*
3.3.2 Estimating point density in a network *41*
3.4 Chapter summary 46

The aim in this chapter is to introduce the reader to various models and techniques used in the spatial analysis of networks later in this dissertation. The chapter opens in §3.1 with an overview of the two geographic information science modelling approaches typically adopted when modelling real-world entities, together with two corresponding data representation techniques typically adopted during implementations of these modelling approaches. With this knowledge, a more in-depth discussion ensues in §3.2 on how networks are modelled mathematically in a GIS context. The focus then shifts in §3.3 to the various statistical network spatial analysis methods that may be employed during an investigation. The chapter finally closes in §3.4 with a brief summary of its contents.

3.1 Constructing a GIS model of the real world

Constructing a tangible geospatial model representation of reality is by no means an easy task. One view is to think of the world as being composed of a number of entities [14, 99, 133]. In this case, the word "entity" includes a broad spectrum of possibilities such as, for instance, the locations of buildings, roadways, and parks, as well as the locations of wildfires, crime, and road accidents. In reality, a variety of questions may arise for which answers are only revealed through information presented in a geospatial sense.

The main concern in the field of *geospatial information science* is to attempt to answer such questions. This is typically achieved by constructing a general framework according to which entities are quantified meaningfully so that a geospatial analysis may be conducted. The general geospatial information science process according to which real-world entities are quantified is depicted in Figure 3.1, and is conducted in two steps. The first step involves conceptualising a collection of entities as an abstract model, called a *conceptual model*. Space may, however, be conceptualised in one of two ways [293]:

1. Space may be viewed as being *relative*, consisting of a set of *objects*, each with spatial properties.
2. Alternatively, space may be viewed as being *absolute*, consisting of a set of *fields* (or places), each with location-specific properties.

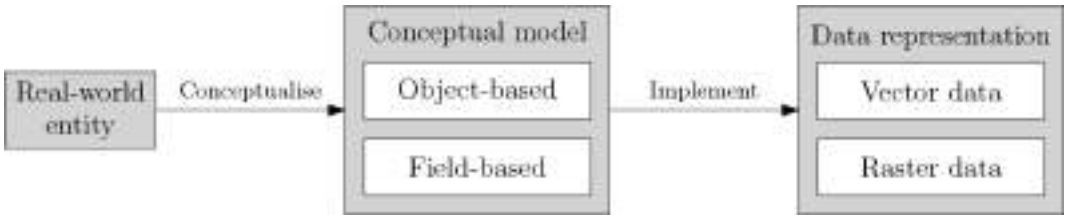


FIGURE 3.1: *Modelling the real world in a geospatial sense [194].*

Consequently, two types of conceptual models are frequently adopted in the geospatial information science literature [35, 87, 293], coinciding with each of these two conceptualisation views. The first is an *object-based* modelling approach, according to which an object may be thought of as a discrete geographic phenomenon (*e.g.* a road or building), while the second is a *field-based* modelling approach, where a field may be thought of as a continuously varying geographic phenomenon (*e.g.* rainfall or temperature). It is also possible to interchange between these two modelling approaches, by either treating groups of objects as fields, or by defining regions of fields as objects. The choice of the most appropriate model is, however, to a large extent, specific to the particular study at hand [14, 197].

The second step is to describe this conceptual model in terms of numerical values so that it may be implemented in a computer-readable form, called a *data representation* [14, 35, 87]. Again, there are two types of data representations that are frequently employed in a GIS, namely *vector* data representation and *raster* data representation. In principle, one may employ either a vector data representation or raster data representation when presented with data collected from an object-based point of view or that of a field-based modelling approach. In practice, however, a vector data representation is almost always associated with an object-based modelling point of view, while a raster data representation is primarily implemented from a field-based modelling point of view [55, 56, 142, 293]. These different approaches and representations are further discussed in the remainder of this section.

3.1.1 Conceptual modelling approaches

In an object-based modelling approach, space is assumed to comprise discrete, identifiable *objects*, each with a geospatial reference [293]. The hierarchical structure of an object-based model is depicted in Figure 3.2. The spatial information associated with these objects are described in terms of *geometric elements*, where a geometric element is simply a point, line, polygon, or polyhedron contained within a Cartesian space. The conceptual form of an object is represented

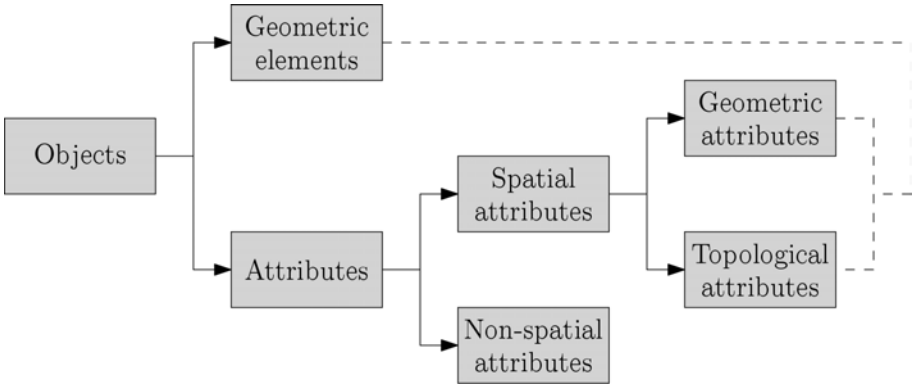


FIGURE 3.2: *The hierarchical structure of an object-based geospatial model framework [194].*

by either a single geometric element or multiple geometric elements. A road network may, for example, be expressed as a series of connected lines, while a building may be represented by a rectangular polyhedron.

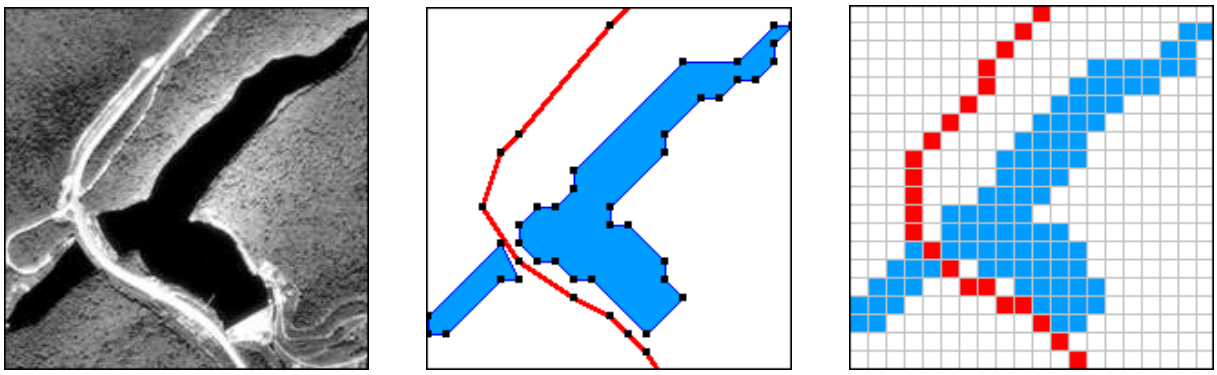
Besides occupying a particular spatial position, objects also posses specific abstract characteristics, commonly referred to as *attributes*. These attributes may be classified into the classes of *spatial attributes* or *non-spatial attributes* [133]. A spatial attribute is an attribute that directly refers to a physical location on the earth's surface, such as a road accident object, which may have specific "latitude" and "longitude" attributes describing where the accident took place. In contrast, an attribute which cannot be related to an exact location is called a non-spatial attribute. A building object may, for example, have an "owner" attribute which does not reveal any spatial information about the position or physical structure of the building.

Spatial attributes may further be categorised into the classes of *geometric attributes* or *topological attributes*, both derived from the geometric elements forming the object [194]. Geometric attributes capture an object's quantitative characteristics, such as the length of a road or the area of a park. Topological attributes, on the other hand, may be seen as the spatial relationships that exist between the geometric elements comprising an object. They represent the qualitative characteristics of an object, which are ultimately preserved upon transformation (stretching or shrinking) of the topological space. Examples of such topological attributes include the intersection of two roads or the inclusion of a lake area within a park's boundaries.

In a field-based modelling approach, however, an entity is represented by a function of one or more attribute values, for every possible point in a Cartesian space [99, 133, 142, 216]. In other words, an entity, in this case, refers to a specific geographic phenomenon which may be represented by a spatial distribution, formalised by a so-called *field function* mapping information from a spatial framework to an attribute domain [293]. One may, for example, wish to model the atmosphere's (an entity) temperature in degrees (an attribute value) on the Earth's surface (a space). Field-based models are typically used to describe attribute values associated with a numerical scale (interval and/or ratio), such as elevation, rainfall, terrain slope, barometric pressure, soil moisture and ocean salinity [194]. Field-based models may also accommodate categorical scales (nominal and/or ordinal). Given a study region, for example, the land usage may be represented by a multi-valued function, defined on every point in the study region, where attributes may assume one of several categorical values, such as, forest, water, grassland, cultivated crops or developed area.

3.1.2 Data representation

As previously mentioned, vector data representation is simply a computer implementation of an object-based modelling point of view. By definition, the object-based model describes attributes as being either spatial or non-spatial, where the former may further be decomposed into data representations describing geometric attributes as well as topological attributes (as per Figure 3.2). The data representations associated with geometric attributes are characterised by means of a finite set of coordinates forming geometric elements [194]. Consequently, a set of geometric attributes implicitly designates the topological attributes of an object. Since geometric elements only consist of discrete points, lines, polygons, or polyhedrons[1], a continuous shape must be approximated by more than one of these elements. In the case of a curved segment, for example, one may represent the segment as a series of connected line segments, while a curved surface may be approximated by an arrangement of connected polygons. Given an aerial photograph, like that in Figure 3.3(a), for example, a vector data representation may be constructed in a manner as shown in Figure 3.3(b). The data representations of non-spatial attributes are typically formulated using a table-based database, known as a *relational database* [63, 69].



(a) (b) (c)

FIGURE 3.3: *(a) An aerial photograph with corresponding two-dimensional (b) vector and (c) raster representations [93].*

There are noteworthy advantages associated with representing data in vector form. First, a specific degree of precision and accuracy may be maintained if a high spatial resolution is employed when constructing objects [93]. Points, lines, simple polygons and simple polyhedra also occupy significantly lower disk storage space when stored in vector format rather than in raster data format. Furthermore, complex topological attributes may also be constructed in a vector data format [235]. In a municipal water network, for example, one may easily model the intricate connections of pipes and valves in order to monitor the direction and flow of water. This is not possible with raster data models.

Naturally, there are also disadvantages associated with vector data models. Complex geometric algorithms associated with distance calculations and polygon overlay often require the knowledge of experienced programmers [235]. Many GIS software packages are, however, equipped with various built-in functions to mitigate these difficulties. Vector data representations may also exhibit slow response times when processing large data sets, especially when using a slow computer. Moreover, if one is presented with continuous data, such as elevation data, it is not normally advisable to represent this in a vector format due to the substantial data generalisation or interpolation typically required [33].

[1]These elements may all be described in terms of vectors [190], hence the name *vector data model*.

Raster data representation stands in contrast to vector data representation. A data representation paradigm typically makes use of a regular matrix (or grid lattice) structure to represent spatial entities, where an individual attribute value is recorded for each grid cell (or *pixel*) location [194]. As a way of comparison, given again the aerial photograph of Figure 3.3(a), a raster representation may be constructed as shown in Figure 3.3(c). This data structure is continuous in the sense that each location in space has an assigned value (even if this value is zero). It is for this reason that a field-based modelling approach, which makes use of a continuous field function, is almost always implemented in conjunction with a raster data representation.

Again, there are some noteworthy advantages associated with the method of raster data representation. Raster processing speed is relatively fast compared to that of a vector data representation due to the simple data structure employed and its relatively simple mathematical procedures. Consequently, new spatial analysis functions typically emerge in raster formats before being transferred into a vector format [235]. Raster data representation is also usually conceptually easier to understand and visualise by an observer.

If a low pixel resolution is adopted, the accuracy of the model typically comes into question, since large pixels appear to sacrifice the details observed in reality. If, on the other hand, a higher pixel resolution is used, the file space required to store the raster data grows polynomially in size. Additionally, it is especially difficult to accurately represent entities with many straight line features, such as road networks [33].

3.2 Modelling a network space mathematically

In the preceding section, two *general* spatial analysis approaches were highlighted, namely object-based and field-based modelling approaches. In this section, these modelling approaches are specified exclusively in the context of *network* spatial analysis. Network spatial analysis differs from conventional spatial analysis, since the assumption is typically made in the former paradigm that entities are constrained to a network space rather than in a continuous Euclidean space [118], as is commonly assumed in the latter paradigm. In this way, one may conceptualise so-called *network entities*, which may represent rivers, roads, railways, pipelines and power lines.

3.2.1 An object-based model for networks

An object-based model assumes, in theory, that a set of objects (or entities) $\mathcal{O} = (\mathcal{A}_O, \mathcal{G}_O)$ populates the world, represented by a set $\{\mathcal{O}, t\}$, where t denotes time. An object $\mathcal{O}$ consists of a set of attributes $\mathcal{A}_O$ together with a set of positional information $\mathcal{G}_O$ comprising a collection of geometric elements defining the object's domain [133].

In a network object-based model, objects are assumed to be an arrangement of connected line segments embedded in a two- or three-dimensional Cartesian space. Okabe and Sugihara [194] proposed the following five geometric conditions that have to be satisfied by line segments:

1. a finite number of line segments may be defined,
2. all line segments are finite in length,
3. all line segments are geometrically straight,
4. line segments are only joined together at their endpoints, and

5. for any pair of points on any of the line segments, there is at least one path connecting these two points.

The set of n_L line segments satisfying the above geometric conditions are referred to as *links* and denoted by $\mathcal{L} = \{\ell_1, \ldots, \ell_{n_L}\}$, where a link ℓ_i is the set of points which satisfies $\{(x_1^{(i)}, y_1^{(i)}) + u(x_2^{(i)}, y_2^{(i)}) \mid u \in [0,1]\}$. The set of endpoints of these links are called *nodes* and contained in the set $\mathcal{E}$. Note that multiple links in $\mathcal{L}$ can share a single node endpoint in $\mathcal{E}$. In this way, a so-called *network* space may be defined as the pair of edge and vertex sets, denoted by $\mathcal{N} = (\mathcal{L}, \mathcal{E})$, or alternatively, by the union of all the edges $\ell_1, \ldots, \ell_{n_L}$ in $\mathcal{L}$, written as $\tilde{\mathcal{L}} = \bigcup_{i=1}^{n_L} \ell_i$, which essentially encapsulates every point comprising an edge (including the vertices) for all edges constituting the network. The former notation $\mathcal{N}$ clarifies that the geometric structure of the network is the point of discussion, while the later notation $\tilde{\mathcal{L}}$ draws the reader's attention to the set of all points contained within the network's domain.

The topological attributes of a network $\mathcal{N}$ are implicitly designated by its geometric structure. If the number and length of all the edges contained in $\mathcal{L}$, together with the number of the vertices contained in $\mathcal{E}$, are finite, $\mathcal{N}$ is said to be a *finite* or *bounded* network. In this dissertation, all networks are assumed to be finite. If no two edges in a network cross, except at their endpoints, $\mathcal{N}$ is referred to as a *planar* network, since the associated network can be embedded in a two-dimensional Cartesian plane. If, however, at least one pair of edges cross at a point other than a vertex, the network is called *non-planar*. A road intersection would, for example, be a form of planar crossing, while a road overpass and underpass would be non-planar in nature. Furthermore, the links in $\mathcal{L}$ may be represented by an ordered pair of vertices, in which case $\mathcal{N}$ is called a *directed* network. If the edges in $\mathcal{L}$ are unordered, the associated network is called *undirected*. A road may, for example, only prohibit the flow of traffic in one direction (*i.e.* a one-way street), in which case the road would be modelled by a directed edge.

3.2.2 A field-based model for networks

In contrast to an object-based model, a field-based model assumes, given a Cartesian space $\mathcal{C}_F$, that the world is represented by a set $\{\mathcal{A}_F, \mathcal{C}_F, t\}$. An entity is defined by a field function $f(\mathcal{A}_F)$, where $\mathcal{A}_F$ denotes a set attributes defined everywhere in $\mathcal{C}_F$, and t is the associated time [133]. It is important to note that the attributes in $\mathcal{A}_F$ are *everywhere* defined in $\mathcal{C}_F$. This was not the case with the attributes in an object-based model, where attributes $\mathcal{A}_O$ were rather restricted to the domain $\mathcal{G}_O \subset \mathcal{C}_F$.

In a network field-based model, a field function is defined for every point within the network space $\tilde{\mathcal{L}}$ [56, 298]. A field function over a network space can be obtained from either the attribute values of the network, or from those entities in the study region which do not form part of the network. In the former case, for example, a field function can be constructed for a road network from a traffic flow attribute which is common to each road segment. In the latter case, suppose that one would like to test the hypothesis that the number of road accidents occurring at night is inversely proportional to the luminosity of the point on the road where accidents occur. In this case, a field function can be constructed by measuring, for each point in the network $\tilde{\mathcal{L}}$, the distance to each of the nearby street lamps.

3.2.3 The binomial point process

In the context of an object-based model, the geometric elements of a network are limited to point and line elements. The attributes of an object (spatial or non-spatial) are, however, not only

limited to deterministic values, but may be stochastic in nature. The stochastic process by which points are spatially distributed over a network space is known as the *stochastic point process*[2]. A fundamental assumption in the field of spatial statistical analysis is that this stochasticity occurs as a result of a binomial distribution of points over the network [194].

Consider a finite network space $\tilde{\mathcal{L}} = \bigcup_{i=1}^{n_L} \ell_i$ embedded in a two-dimensional Cartesian space, as shown in Figure 3.4. For an arbitrary point $p \in \tilde{\mathcal{L}}$, with coordinates (x, y), consider a real-valued function $f(p)$ defined on the domain $\tilde{\mathcal{L}}$. If the conditions

$$0 \leq f(p) < \infty \quad \text{for all } p \in \tilde{\mathcal{L}}$$

and

$$\int_{p \in \tilde{\mathcal{L}}} f(p)\, \mathrm{d}p = 1$$

are satisfied, the function $f(p)$ is called a *probability density function* [194].

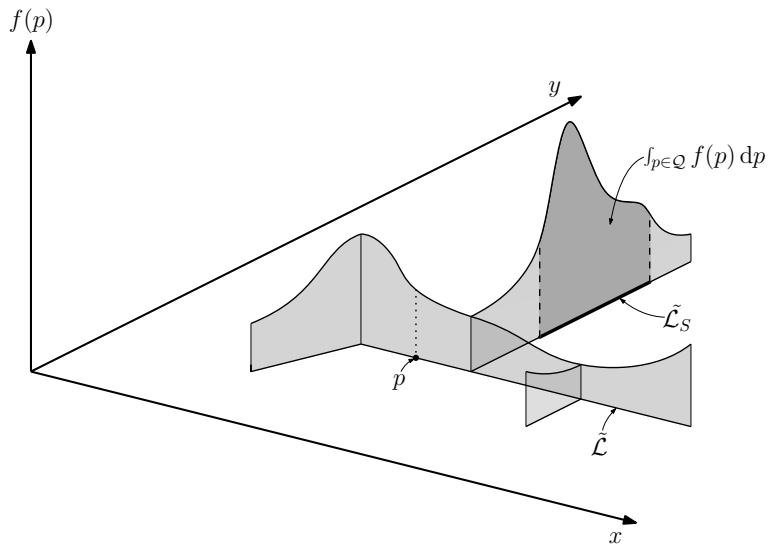


FIGURE 3.4: *A probability density function $f(p)$ on $\tilde{\mathcal{L}}$ [194].*

Suppose a finite set of n_P points $\mathcal{P} = \{p_1, \ldots, p_{n_P}\}$ are *identically and independently* created in $\tilde{\mathcal{L}}$ according to the probability density function $f(p)$, where the phrase "independently and identically" implies that all points in $\mathcal{P}$ are *randomly sampled* from the *same* density function $f(p)$. Let $\tilde{\mathcal{L}}_{\mathcal{S}}$ be an arbitrary subset of $\tilde{\mathcal{L}}$ (*i.e.* $\tilde{\mathcal{L}}_S \subset \tilde{\mathcal{L}}$) and denote the subset of points in $\mathcal{P}$, which lie on $\tilde{\mathcal{L}}_S$, as $\mathcal{P}_S \subset \mathcal{P}$. Denote the number of points in $\mathcal{P}_S$ by $n_S \leq n_P$, where n_S may be described in terms of a random variable following a binomial distribution with parameters n_P and $\int_{p \in \tilde{\mathcal{L}}_S} f(p)\, \mathrm{d}p$. As a result of this property, this stochastic point process is appropriately referred to as the *binomial point process* [112, 257]. The probability that n_S is exactly equal to k is, therefore, given by

$$P(n_S = k) = \binom{n_P}{k} \left(\int_{p \in \tilde{\mathcal{L}}_S} f(p)\, \mathrm{d}p \right)^k \left(1 - \int_{p \in \tilde{\mathcal{L}}_S} f(p)\, \mathrm{d}p \right)^{n_P - k}, \tag{3.1}$$

where the probability that a point in $\mathcal{P}$ is also contained in $\mathcal{P}_S$ is given by $\int_{p \in \mathcal{P}_S} f(p)\, \mathrm{d}p$, $1 - \int_{p \in \mathcal{P}_S} f(p)\, \mathrm{d}p$ is the probability that a point in $\mathcal{P}$ is not contained in $\mathcal{P}_S$, and $\binom{n_P}{k}$ is the

[2]In spatial statistics, considerable research has been devoted to stochastic point processes on a plane [11, 71, 159, 175, 218], while many of these statistical techniques related to a network space were conceived and formalised by Okabe and Sugihara [194].

binomial coefficient indicating the number of combinations in which a subset of k points can be chosen (without replacement) from the n_P points contained in $\mathcal{P}$.

Now consider the case where a binomial point process has a *uniformly distributed* probability density function, taking the form

$$f(p) = \frac{1}{|\tilde{\mathcal{L}}|}, \quad p \in \tilde{\mathcal{L}}, \tag{3.2}$$

where $|\tilde{\mathcal{L}}|$ denotes the length of the network $\tilde{\mathcal{L}}$. When the binomial point process of (3.1) has the probability density function in (3.2), it is called *homogeneous* and reduces to

$$P(n_S = k) = \binom{n_P}{k} \left(\frac{|\tilde{\mathcal{L}}_S|}{|\tilde{\mathcal{L}}|}\right)^k \left(1 - \frac{|\tilde{\mathcal{L}}_S|}{|\tilde{\mathcal{L}}|}\right)^{n-k}.$$

Consequently, when n_P points are generated according to the homogeneous binomial point process on a network space $\tilde{\mathcal{L}}$, the probability of a point being placed on an arbitrary subnetwork $\tilde{\mathcal{L}}_S \subset \tilde{\mathcal{L}}$ depends neither on the shape of $\tilde{\mathcal{L}}_S$ nor on the location of $\tilde{\mathcal{L}}_S$ in $\tilde{\mathcal{L}}$, but rather, only on the length of $\tilde{\mathcal{L}}_S$. According to Diggle [72], the randomness exhibited by the homogeneous binomial point process is referred to as the *complete spatial randomness* (CSR) property[3]. In statistical spatial analysis, a fundamental hypotheses is that points are generated according to the homogeneous binomial point process, commonly referred to as the *CSR hypothesis* [194].

3.2.4 The uniform network transformation

Although the CSR hypothesis is one of the most fundamental assumptions in network spatial analyses, it is often too restrictive when analysing real-world distributions. It is, for example, unrealistic to assume that road accidents occur uniformly within a road network; one would instead expect that they rather occur in proportion to traffic flow volumes [192]. In such a case, the CSR hypothesis would likely be rejected and an *inhomogeneous* binomial point process would be considered instead, specified in terms of a *non-uniform* probability density function. Many of the statistical methods applicable to network spatial analysis are, however, built on the assumption that points are generated according to the CSR property over the network. Fortunately, these methods may still be applied if the non-uniform probability density function is transformed to that of a uniform density function.

Consider a non-uniform probability density function $f(x)$ on the real domain, like that shown in Figure 3.5(a). One may represent this function cumulatively by its *cumulative distribution function* $F(x)$, as shown in Figure 3.5(b), written mathematically as

$$y = F(x) = \int_{-\infty}^{x} f(x)\,\mathrm{d}x, \tag{3.3}$$

where $f(x)$ is the density function of the random variable $x \in \mathbb{R}$. In this case, the following theorem holds.

Theorem 3.1 *(The probability integral transformation [77])*
If a random variable y is generated according to a uniform distribution $f(y)$ on the interval $0 \leq y \leq 1$, then the transformed variable $x = F^{-1}(y)$ has a probability density function given

[3] It is acknowledged that Diggle defines CSR in terms of the Poisson point process, whereas, in this section, it is defined in terms of the binomial point process, due to the fact that a finite network space is always assumed.

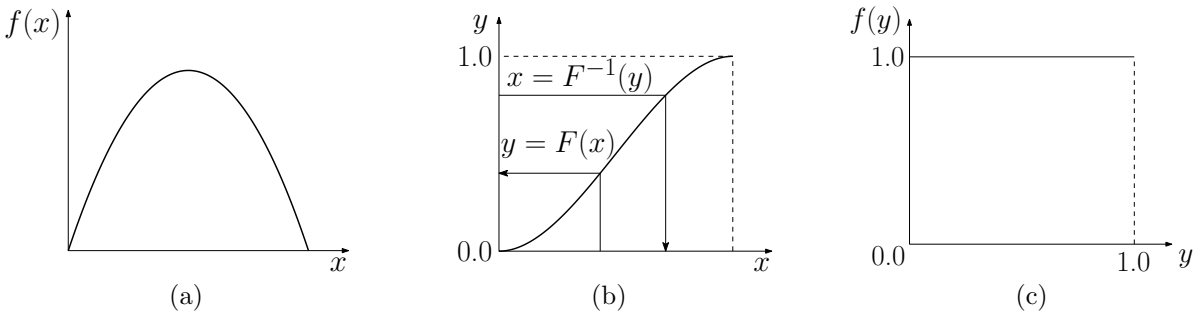


FIGURE 3.5: *The probability integral transformation transforms (a) a probability density function $f(x)$, by means of (b) its cumulative distribution function $F(x)$, into (c) a uniform density function $f(y) = 1$ [194].*

by $f(x)$. Conversely, if x is a continuous random variable generated according to a probability density function $f(x)$, then the transformed variable $y = F(x)$ has a uniform distribution $f(y)$ on the interval $0 \leq y \leq 1$.

This theorem, known as the *probability integral transformation*, was originally conceived by Fisher [77]. It is a powerful tool for transforming a uniformly distributed random number y to a corresponding number x which, in essence, is regarded to be randomly sampled from a probability density function $f(x)$. This is achieved *via* the probability density function's cumulative distribution inverse $x = F^{-1}(y)$. Conversely, if a random number x is generated according to the probability density function $f(x)$, it may be transformed to a corresponding random number $y = F(x)$ which follows a uniform distribution, as shown in Figure 3.5(c).

The network generalisation of the probability integral transformation was first introduced by Okabe and Satoh [192], and is called the *uniform network transformation*. Since all points in $\mathcal{P}$ are assumed to be independently and identically generated according to a binomial point process, a probability density function $f(\boldsymbol{x})$ on a network space $\tilde{\mathcal{L}} = \bigcup_{i=1}^{n_L} \ell_i$ can be decomposed into separate probability density functions $f_1, \ldots, f_{n_L}$ associated with each of the n_L links $\ell_1, \ldots, \ell_{n_L}$. Let $f_i(a)$ be a portion of the probability density function $f(\boldsymbol{x})$, where the domain of $f_i(a)$ is the link ℓ_i, a denotes the distance from an endpoint ($a = 0$) of link ℓ_i to some point along link ℓ_i, and $f_i(a)$ is the density at point a on link ℓ_i. One may then define a transformation for each edge ℓ_i in the same manner as the probability integral transformation in (3.3), written as

$$b = F_i(a) = \int_0^a f_i(u)\,\mathrm{d}u.$$

A network space $\mathcal{N} = (\mathcal{L}, \mathcal{E})$, like that shown in Figure 3.6(a), is thus effectively transformed into a new network space $\mathcal{N}' = (\mathcal{L}', \mathcal{E}')$, illustrated in Figure 3.6(b), such that each link $\ell_i \in \mathcal{L}$ corresponds to a new link $\ell_i' \in \mathcal{L}'$, the node at $a = 0$ in $\mathcal{E}$ corresponds to the node at $b = 0$ in $\mathcal{E}'$, and a point a on the link ℓ_i is transformed to the point $b = F_i(a)$ on the link ℓ_i'. In this way, any non-uniform distribution on $\tilde{\mathcal{L}}$ can be transformed to a uniform distribution $g_i(b) = 1$ on $\tilde{\mathcal{L}}'$.

3.2.5 Edge biases on bounded networks

In network spatial analysis, entities are theoretically assumed only to exist within a bounded network space, although, in practice, this assumption is almost never valid. When defining a study region containing a network, the boundaries of this region may either be defined by *natural* boundaries, in which case the enclosed network is called a *naturally bounded* network,

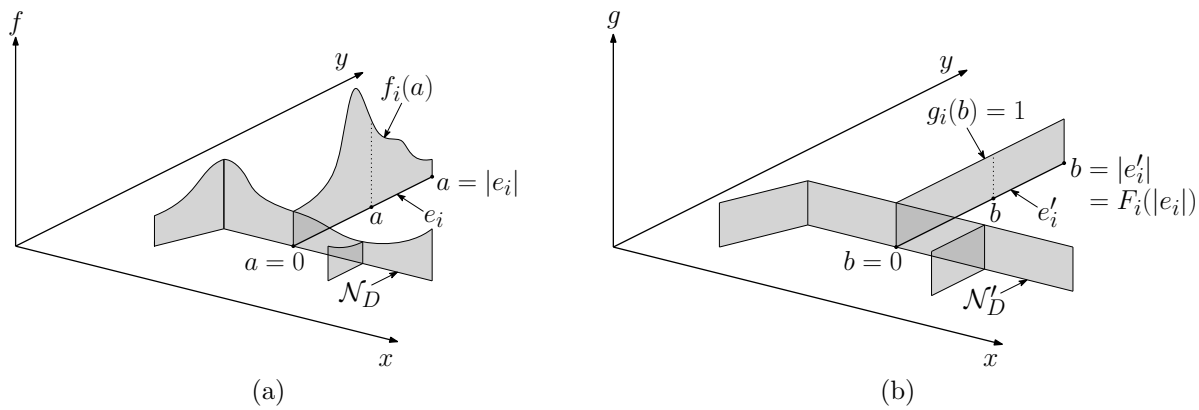


FIGURE 3.6: *The uniform network transformation transforms (a) a non-uniform network density function into (b) a uniform network density function [194].*

or by *artificial* boundaries, in which case the network is called *artificially bounded*. An example of the former situation is depicted in Figure 3.7, where a road network is situated on an island bounded by the island's coastline. In this illustration, the assumption that, for example, no road accidents exist beyond the island's coastline may be deemed valid. If, on the other hand, one were to define the study region as a subset of the original road network (indicated by the bold lines within the dashed rectangle), then assuming that no road accidents exist outside this artificial boundary is almost certainly invalid.

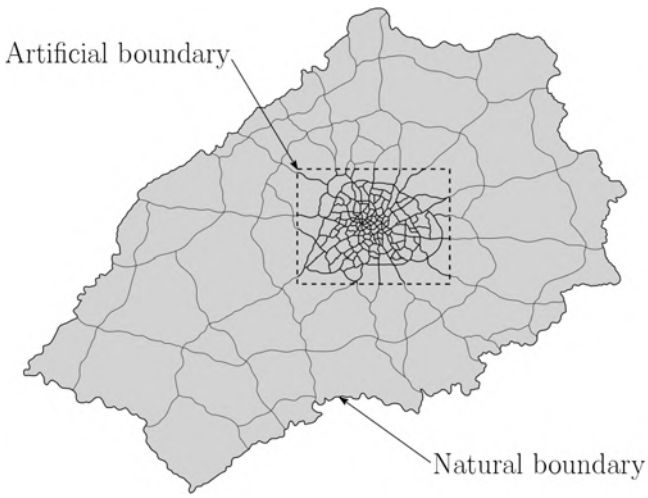


FIGURE 3.7: *A naturally bounded network situated on an island, and an artificially bounded network contained within the dashed rectangle.*

During a spatial analysis investigation, the data associated with entities existing outside a defined study region are typically not available. As a result, the spatial influence of entities which exist outside an artificially bounded network is often simply ignored. This may, however, result in biased calculations when employing various network spatial analysis techniques, since entities that exist outside the study region may ultimately affect the entities within the a study region. This resulting bias is often called *edge biases* or *edge effects*.

An in-depth discussions on the implications of edge effects may be found in the literature [21, 299], but many of the proposed methods for correcting these edge effects, known as *edge correction* [219], are restricted in a planar sense. When dealing with a network space, one can often,

in practice, define an artificially bounded network in such a way that only a small number of artificial edges (or network breaks) are created. Edge effects may therefore not be as severe, when conducting a network spatial analysis, as it may be in the case of planar spatial analysis [194].

3.3 Network point pattern analysis methods

Point pattern analysis (PPA) is the study of the spatial arrangement of points within a defined study region. PPA methods are typically classified into two groups, namely *distance-based* approaches and *density-based* approaches. The former is concerned with how points are distributed relative to one another (*i.e.* the second-order property of a pattern), while the latter characterises the pattern of points in terms of its distribution within a defined study region (*i.e.* the first-order property of a pattern) [86]. The first part of this section is devoted to a discussion on a distance-based method called the network *nearest neighbour distance* (NND) method. Additionally, since researchers studying RAs are typically concerned with identifying so-called accident *hotspots* (or *blackspots*) [17, 79, 154, 255], where the density of RAs is significantly larger than in other areas, a density-based method called the network *kernel density estimation* (KDE) method is also discussed in the final part of this section.

3.3.1 Measuring uniformity of point distributions in a network

Consider a network $\mathcal{N} = (\mathcal{L}, \mathcal{E})$ composed of a set $\mathcal{E} = \{e_1, \ldots, e_{n_E}\}$ of n_E nodes and a set $\mathcal{L} = \{\ell_1, \ldots, \ell_{n_L}\}$ of n_L links. Given a set of points $\mathcal{P} = \{p_1, \ldots, p_{n_P}\}$ on the nodes and links of this network, one would often like to test whether the points in $\mathcal{P}$ are statistically close to one another or dispersed within the network space. A typical question may, for example, be whether or not restaurants tend to cluster together along the streets in a city. In such a point pattern analysis, two of the most common distance-based methods used to answer the above question are the so-called *G-function* method (often also referred to as the NND method) and the *K-function* method [197]. Okabe and Sugihara [194] proposed an extension of the NND method defined on a plane so that it is applicable to a network space, while Okabe and Yamada [196] proposed a similar network extension of the K-function method. Only the derivation of the former is discussed here in detail as only this method is utilised later in this dissertation. The network NND method may be employed to test whether the CSR hypothesis indeed holds by determining whether the shortest-path distance from each point to its nearest point in the network space is statistically short (or long) [194].

Assume that the points in $\mathcal{P}$ are generated according to a homogeneous binomial point process over $\tilde{\mathcal{L}} = \bigcup_{i=1}^{n_L} \ell_i$, thereby implying the CSR hypothesis described in §3.2.3. For a specific point $p_i \in \mathcal{P}$, consider the sub-network $\tilde{\mathcal{L}}(k \mid p_i) \subset \tilde{\mathcal{L}}$ consisting of the set of points in $\tilde{\mathcal{L}}$ that are within a distance k from p_i. This may be written mathematically as $\tilde{\mathcal{L}}(k \mid p_i) = \{p_j \mid d(p_i, p_j) \leq k, p_j \in \tilde{\mathcal{L}}\}$, where $d(p_i, p_j)$ represents the network distance from point p_i to some point $p_j \in \tilde{\mathcal{L}}$. For the point p_i, one may determine the next closest point along the network $\tilde{\mathcal{L}}$ as the point

$$\underset{j \in \mathcal{P},\ j \neq i}{\arg\min}\ d(p_i, p_j),$$

denoted by p_i^*. The distance from p_i to p_i^* is, therefore, the quantity $d(p_i, p_i^*)$, called the *nearest neighbour distance* (hence the method's name).

Given a naturally bounded network space, like that in Figure 3.8, for example, the set of points $\tilde{\mathcal{L}}(2 \mid p_2)$ which are within a distance $k = 2$ from point p_2 are indicated by the thick line

segments. In this case, the nearest neighbouring point to p_2 is the point $p_2^* = p_5$, being a total of $d(p_2, p_2^*) = 4$ units away from p_2, where the shortest network path is attained by travelling along the links through nodes e_7 and e_9.

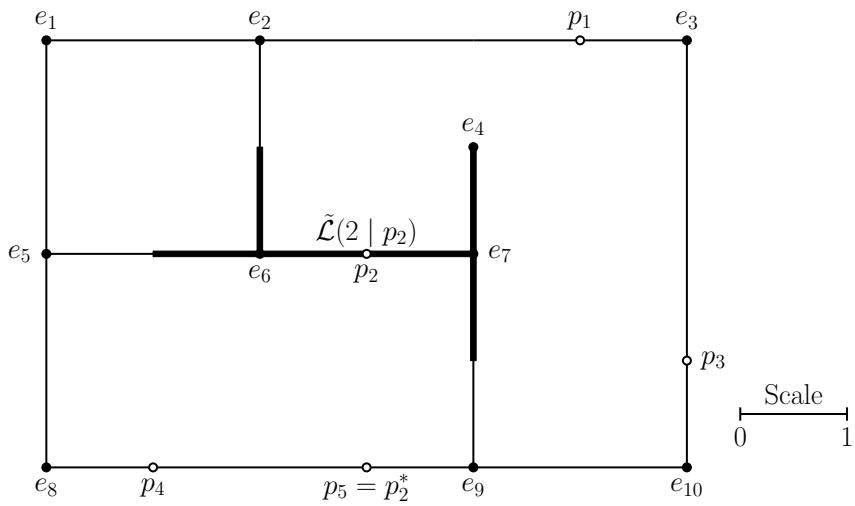


FIGURE 3.8: *The set of points $\tilde{\mathcal{L}}(2 \mid p_2)$ within a distance $k = 2$ from a point p_2 are indicated by the think line segments. The white open points denote the points in $\mathcal{P} = \{p_1, \ldots, p_5\}$ and the solid points denote the nodes in $\mathcal{E} = \{e_1, \ldots, e_{10}\}$, which form a naturally bounded network $\mathcal{N} = (\mathcal{L}, \mathcal{E})$.*

One may, therefore, begin to formulate a statistical test by presupposing that, if the points in $\mathcal{P}$ were *truly* generated according to a binomial point process over $\tilde{\mathcal{L}}$, then the distance between neighbouring points should fall within some confidence band away from one another. In this case, the NND $d(p_i, p_i^*)$ is treated as a random variable under the CSR hypothesis, and the probability density function $f_i(k \mid p_i)$ is given by the probability that the random NND variable $d(p_i, p_i^*)$ is less than or equal to k. This probability may be rewritten as

$$\begin{aligned} \mathrm{P}(d(p_i, p_i^*) \leq k) &= 1 - \mathrm{P}(\text{all points in } \mathcal{P} \text{ excluding } p_i \text{ are further than } k \text{ units away from } p_i) \\ &= 1 - \mathrm{P}(\text{the } n_P - 1 \text{ points } p_1 \ldots, p_{i-1}, p_{i+1}, \ldots, p_{n_P} \text{ lie on } \tilde{\mathcal{L}} \backslash \tilde{\mathcal{L}}(k \mid p_i)), \end{aligned}$$

where $\tilde{\mathcal{L}} \backslash \tilde{\mathcal{L}}(k \mid p_i)$ is the complement of the sub-network $\tilde{\mathcal{L}}(k \mid p_i)$ (*i.e.* the thin lines in Figure 3.8). The probability density function $f_i(k \mid p_i)$ may, therefore, be written as

$$f_i(k \mid p_i) = 1 - \left(\frac{|\tilde{\mathcal{L}}| - |\tilde{\mathcal{L}}(k \mid p_i)|}{|\tilde{\mathcal{L}}|} \right)^{n_P - 1}, \tag{3.4}$$

where $| \bullet |$ denotes the length of $\bullet$. Consequently, one may compute the expected value

$$\mu(p_i) = \mathrm{E}(d(p_i, p_i^*)) = \int_{k=0}^{\infty} k f_i(k \mid p_i) \, \mathrm{d}k$$

and the variance

$$\sigma^2(p_i) = \mathrm{Var}(d(p_i, p_i^*)) = \mathrm{E}[(d(p_i, p_i^*) - \mu(p_i))^2] = \int_{k=0}^{\infty} (k - \mu(p_i))^2 f_i(k \mid p_i) \, \mathrm{d}k$$

of (3.4) for the random NND variable $d(p_i, p_i^*)$. Note that (3.4) also takes into account the bounded nature of a network and, consequently, if this function is applied to a naturally bounded network, edge effects are not exhibited.

With this in mind, a statistical index may be defined for measuring whether the set of distances $\mathcal{D} = \{d(p_i, p_i^*) \mid i \in \{1, \ldots, n_P\}\}$ are statistically short (or long). As an intermediary notion, first consider a statistical index $g(p_i)$, defined for a point p_i, as the ratio of the random NND variable $d(p_i, p_i^*)$ to its expected value $\mu(p_i)$, expressed mathematically as

$$g(p_i) = \frac{d(p_i, p_i^*)}{\mu(p_i)}. \tag{3.5}$$

The expected value of $g(p_i)$ is, by definition, unity (*i.e.* $\mathrm{E}(g(p_i)) = 1$). Since the probability density function of $g(p_i)$ is obtained from $f_i(k \mid p_i)$, a standard statistical approach may be followed to test the CSR hypothesis. Despite the fact that such a test is hardly ever useful (since it only considers a single point p_i), the notion of $g(p_i)$ may be used to formulate a more global statistical index.

When formulating such a statistic, one does not only wish to consider a single point p_i, but also all the other points $p_1, \ldots, p_{i-1}, p_{i+1}, \ldots, p_{n_P}$ in $\mathcal{P}$. The random NND variable across all points, denoted by D, is the average of the variables $d(p_i, p_i^*)$ over all $i \in \{1, \ldots, n_P\}$, expressed mathematically as

$$D = \sum_{i=1}^{n_P} \frac{d(p_i, p_i^*)}{n_P}. \tag{3.6}$$

Let μ be the expected value of the random variable D under the CSR hypothesis (*i.e.* all the points in $\mathcal{P}$ are assumed to be generated according to the binomial point process). The value of $\mu = \mathrm{E}(D)$ is then given by the expected value of $d(p_i, p_i^*) = \mu(p_i)$ computed over all $p_i \in \tilde{\mathcal{L}}$. That is,

$$\mu = \mathrm{E}(D) = \frac{1}{|\tilde{\mathcal{L}}|} \int_{p_i \in \tilde{\mathcal{L}}} \mu(p_i)\, \mathrm{d}p_i,$$

where the integration of $\mu(p_i)$ is performed along the network space $\tilde{\mathcal{L}}$. In practice, the integral in (3.3.1) is hard to compute analytically and so Monte Carlo simulation is typically employed instead in order to estimate its value [194]. Given μ, a statistical index may be defined in a similar manner as in (3.5), as the ratio of the random NND variable D across all points to its expected value μ, expressed mathematically as

$$G = \frac{1}{\mu} \frac{\sum_{i=1}^{n_p} d_S(p_i, p_i^*)}{n_p}. \tag{3.7}$$

This index is an extension of the Clark-Evans index [46], defined in a planar sense, as opposed to that strictly defined in a network space, as proposed by Okabe *et al.* [195]. The expected value of G is again, by definition, unity (*i.e.* $\mathrm{E}(G) = 1$) under the CSR hypothesis.

Much like in the case for a single point p_i, one seeks the probability density function (like that in (3.4)) for the random NND variable D which encompasses *all* the NNDs for the points in $\mathcal{P}$. An explicit probability density function of this kind is, however, difficult to obtain analytically [194]. An alternative approach lies in the property that the random NND variable $D = \sum_{i=1}^{n_P} d(p_i, p_i^*)/n$ approaches the normal distribution as n_P approaches infinity. This property is due to the central limit theorem [287], which applies to the averaging of the random variables $d(p_1, p_1^*), \ldots, d(p_{n_P}, p_{n_P}^*)$. In practice, if n_P is sufficiently large, one may test the CSR hypothesis by means of a probability density function of the normal distribution, where the expected value and variance is discovered through Monte Carlo simulation [194].

Using these results, one may statistically test an observed value G under the CSR hypothesis as follows: Let G^+ and G^- be the upper and lower critical values for a particular significance level α. If $G \in [G^-, G^+]$, there is not sufficient evidence to reject the CSR hypothesis and one

may conclude, at a confidence level of $1-\alpha$, that the points in $\mathcal{P}$ have indeed been generated according to a binomial point process within a network. If, on the other hand, G does not fall within these bounds (*i.e.* $G < G^-$ or $G > G^+$), there is sufficient evidence to reject the CSR hypothesis and one may conclude that the average NND between the points in $\mathcal{P}$ is significantly shorter (or longer) than expected — such a result would imply that the points in $\mathcal{P}$ tend to cluster (or are unnaturally dispersed) within a network.

The preceding formulation of the statistical index G is an aggregated measure of all observed NNDs. One may, however, wish instead to examine the observed index G in a more disaggregated manner. If this is the case, one may define, for a specific distance k, the number of observed points satisfying $d(p_i, p_i^*) \leq k$ as the function $n(k)$. In this way, an *observed* network G-function may be defined as

$$G(k) = \frac{n(k)}{n_P}. \tag{3.8}$$

An *expected* network G-function curve $\overline{G}(k)$ may also be defined to indicate the number of points that are expected to satisfy $d(p_i, p_i^*) \leq k$ at a specified distance k. In the same way, a pair of envelope curves $G^+(k)$ and $G^-(k)$ may also be defined to capture the significance bounds for a specified significance level α. Once again, the expected value and variance of the function $n(k)$, together with the expected upper and lower envelope curves, are difficult to obtain analytically, and hence Monte Carlo simulation is typically employed to determine these curves [194].

Consider, for example, the study conducted by Okabe and Sugihara [194] in which the authors aimed to test whether or not beauty parlours in Shibuya Ward, Tokyo tend to cluster near one another. The spatial representation of the data used in their study is illustrated in Figure 3.9. In the study, the observed statistic index G in (3.7) was calculated for the point distribution within the network, and was reported to have a value $G = 0.62$. The authors statistically tested this result using a normal distribution with $\mathrm{E}(G) = 1.0$ and $\sqrt{\mathrm{Var}(G)} = 0.023$, where these values were obtained *via* Monte Carlo simulation. Under the CSR hypothesis and with a $1-\alpha = 0.95$ confidence level, their finding was that the observed NND between beauty parlours is statistically shorter than expected. This result implies that beauty parlours tend to be purposely located near one another in Shibuya Ward, Tokyo.

FIGURE 3.9: *An example illustrating the locations of various beauty parlours alongside the roads in Shibuya Ward, Tokyo [194].*

The G-function of the study conducted by Okabe and Sugihara [194] is shown in Figure 3.10.

As may be seen in the figure, the observed G-function curve (indicated by the black curve) is above the upper envelope curve (indicated by the grey curve) for $0 \leq k \leq 200$ metres. One may, therefore, extend the previous finding by concluding, at a confidence level of $1 - \alpha = 0.95$, that beauty parlours tend to cluster together significantly within a distance of 0–200m in Shibuya Ward, Tokyo.

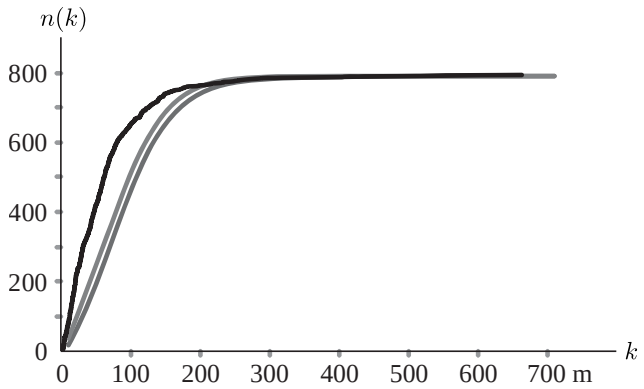


FIGURE 3.10: *The number $n(k)$ of beauty parlours satisfying the NND under the CSR hypothesis at different values of k [194]. The black curve in corresponds to the observed NND G-function while the two grey curves denote the upper and lower envelope curves G^+ and G^- at a significance level of $\alpha = 0.05$.*

3.3.2 Estimating point density in a network

Consider again a network $\mathcal{N} = (\mathcal{L}, \mathcal{E})$ composed of a set $\mathcal{E} = \{e_1, \ldots, e_{n_E}\}$ of n_E nodes and a set $\mathcal{L} = \{\ell_1, \ldots, \ell_{n_L}\}$ of n_L links. Given a set of points $\mathcal{P} = \{p_1, \ldots, p_{n_P}\}$ on the nodes and links of this network, one may require an estimate of the density of points on a network for the purpose of identifying high-density areas within the network. In the GIS literature, this is traditionally achieved by means of *network histograms* which define the point density along a line segment as the number of points on the line segment divided by the line segment's length. Although this approach is simple and easy to implement in practice, it often produces bias [194], especially when a network is approximated at a high level of abstraction (*i.e.* fewer line segments are used to represent the total network structure). To overcome this shortcoming, Okabe *et al.* [193] proposed a method for unbiased density estimation on a network by extending the well-known planar KDE method originally proposed by Fix and Hodges [78].

To provide intuition in respect of the underlying KDE method, first consider a simple example in which five points $p_1, \ldots, p_5$ have been distributed along a one-dimensional line x as illustrated in Figure 3.11. The first step is to construct individual kernels centred at each of the points in $\mathcal{P}$ (indicated by the dashed lines). The effects of each individual kernel are then summed across the length of x to produce a *net* kernel density function $K(\mathcal{P})$ (indicated by the solid line segment). The estimation procedure used to produce this final kernel density function is called the KDE method. In the remainder of this section, the general procedure is described for ensuring that unbiased kernel density functions are defined on a network, as proposed by Okabe and Sugihara [194].

Consider a point p on a non-directed network $\tilde{\mathcal{L}} = \bigcup_{i=1}^{n_L} \ell_i$ embedded in a plane, where all line segments are presented, forming one *connected* network, as illustrated in Figure 3.12. A sub-network $\mathcal{L}_p \subset \tilde{\mathcal{L}}$ is constructed such that the shortest-path distance between p and any other point on $\mathcal{L}_p$ is less than or equal to w (referred to as the *buffer network* of the point p with a

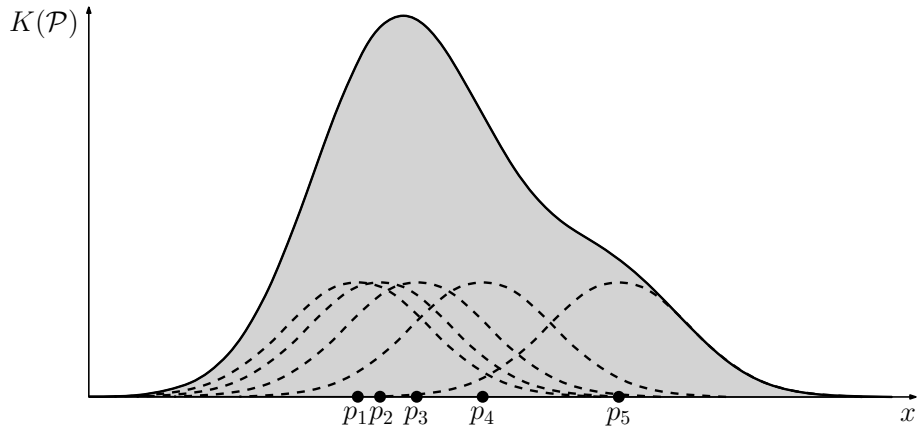


FIGURE 3.11: *An example illustrating five points distributed along a line segment, where the dashed lines represent the individual kernel functions centred at each of the five points and the solid line represents the net kernel density function.*

width of w). For an arbitrary point $q \in \tilde{\mathcal{L}}$, a function $K_p(q)$ may be defined for a given point p which satisfies

$$K_p(q) \begin{cases} \geq 0 & \text{for } q \in \mathcal{L}_p, \\ = 0 & \text{for } q \in \tilde{\mathcal{L}} \backslash \mathcal{L}_p, \end{cases} \tag{3.9}$$

and

$$\int_{q \in \mathcal{L}_p} K_p(q) \, \mathrm{d}q = 1, \tag{3.10}$$

where $\tilde{\mathcal{L}} \backslash \mathcal{L}_p$ in (3.9) is the complement of the sub-network $\mathcal{L}_p$ with respect to the network $\tilde{\mathcal{L}}$, and the integral over $q \in \mathcal{L}_p$ in (3.10) is the one-dimensional integral of the function $K_p(q)$ with respect to q along $\mathcal{L}_p$. In light of the preceding discussion, the following nomenclature is introduced: The function $K_p(q)$ is called the *network kernel density function* for which the point p defines the *kernel centre*. The sub-network $\mathcal{L}_p$ is, in this context, referred to as the *kernel support*, and the value w is called the *bandwidth* of $K_p(q)$.

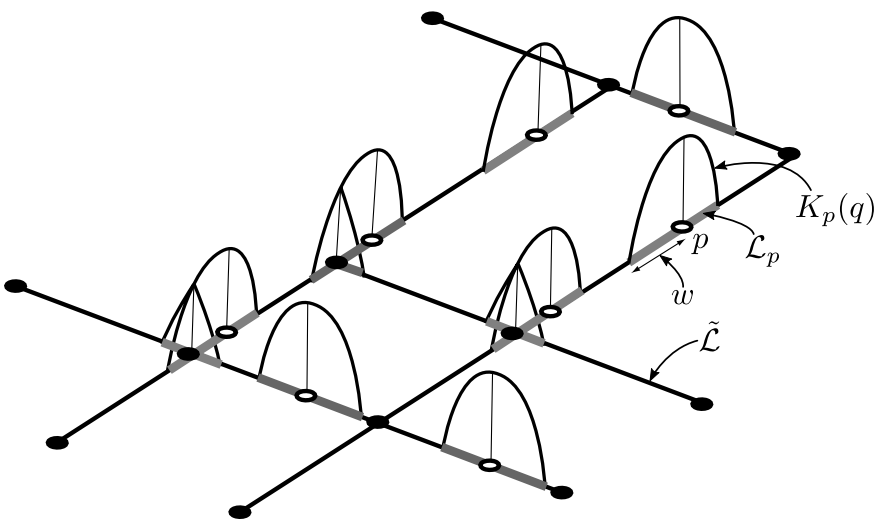


FIGURE 3.12: *Multiple kernel density functions on a network embedded in a plane, where the white open points indicate the kernel centres and the grey bold line segments denote the kernel supports [194].*

For the set of n_P points in $\mathcal{P}$ on the network $\tilde{\mathcal{L}}$, a function $K(q)$, called the *network kernel density estimator*, is defined to be the average density at a point q, expressed mathematically as

$$K(q) = \frac{1}{n_P} \sum_{i=1}^{n_P} K_{p_i}(q). \tag{3.11}$$

The expression in (3.11) is employed to approximate some *true* point distribution function $f(q)$ over the network $\tilde{\mathcal{L}}$, given a finite set of points in $\mathcal{P}$.

With this in mind, the discussion now shifts to ensuring that, given a non-directed connected network $\tilde{\mathcal{L}}$ (like that shown in Figure 3.13), the individual network kernel density functions $K_{p_i}(q)$ for all $p_i \in \mathcal{P}$ do not produce biases within the network $\tilde{\mathcal{L}}$. Let $\mathcal{E}^{(1)} \subset \mathcal{E}$ be the set of nodes of degree one (*i.e.* the nodes incident with only one link, such as the white open points in Figure 3.13) and let $\mathcal{E}^{(\geq 3)} \subset \mathcal{E}$ be the set of nodes of degree three or more (*i.e.* the solid points in Figure 3.13). Assume, for ease of explanation, that the length of each of the n_L links in $\mathcal{L}$ is at least $4w$ (*i.e.* $|\ell_i| \geq 4w$ for all $i \in \{1, \ldots, n_L\}$). Furthermore, let $\mathcal{L}_{e_i^{(1)}}$ be the buffer network for the ith node $e_i{}^{(1)} \in \mathcal{E}^{(1)}$ with a width of $2w$ (indicated by the grey bold line segments incident with white open points in Figure 3.13), where $\mathcal{L}^{(1)} = \{\mathcal{L}_{e_i^{(1)}} |\ e_i{}^{(1)} \in \mathcal{E}^{(1)}\}$ denotes the set of all buffer networks defined by nodes of degree one. Similarly, let $\mathcal{L}_{e_i^{(\geq 3)}}$ be the buffer network for the ith node $e_i{}^{(\geq 3)} \in \mathcal{E}^{(\geq 3)}$, again of width $2w$ (indicated by the grey bold line segments incident with solid points in Figure 3.13), where $\mathcal{L}^{(\geq 3)} = \{\mathcal{L}_{e_i^{(\geq 3)}} |\ e_i{}^{(\geq 3)} \in \mathcal{E}^{(\geq 3)}\}$ denotes the set of all buffer networks defined by nodes of degree three or more. Consider the case when the line segments in $\mathcal{L}^{(1)}$ and $\mathcal{L}^{(\geq 3)}$ are deleted from the network $\tilde{\mathcal{L}}$. In this case, a set of simple line segments $\mathcal{L}^{(S)} = \tilde{\mathcal{L}} \backslash (\mathcal{L}^{(1)} \cup \mathcal{L}^{(\geq 3)})$ is the set of sub-networks with exactly two endpoints (indicated by the thin line segments in Figure 3.13).

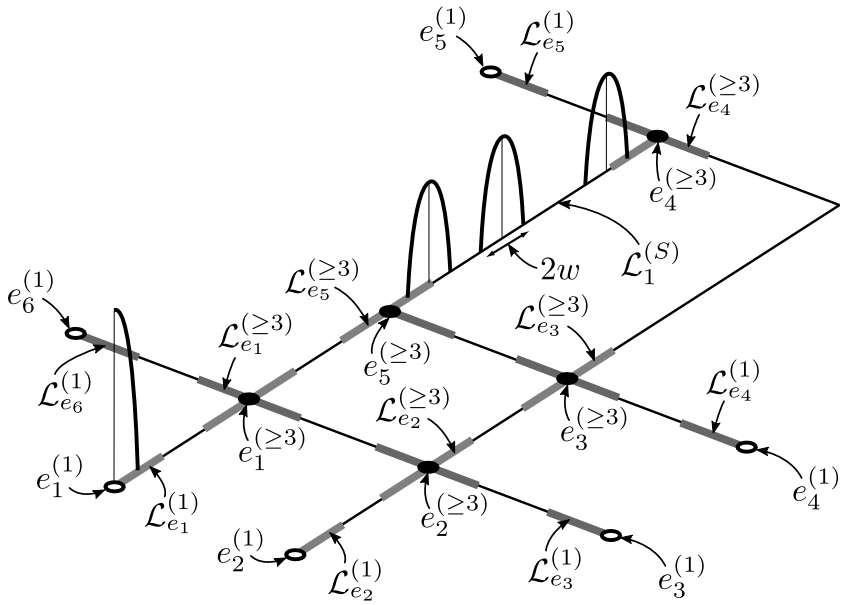


FIGURE 3.13: *A decomposition of a network $\tilde{\mathcal{L}}$ into the sub-networks $\mathcal{L}^{(1)}$ (indicated by the grey bold line segments incident with white open points), $\mathcal{L}^{(\geq 3)}$ (indicated by the grey bold line segments incident with solid points) and $\mathcal{L}^{(S)}$ (indicated by the thin line segments) [194].*

The choice of network KDE methods differ depending on three apparent cases, namely whether a kernel centre lies on a sub-network (1) in $\mathcal{L}^{(S)}$, (2) in $\mathcal{L}^{(1)}$ or (3) in $\mathcal{L}^{(\geq 3)}$. The first case is when a kernel centre lies on a buffer network $\mathcal{L}_i^{(S)} \in \mathcal{L}^{(S)}$. In this case, edge biases (as described in §3.2.5) are not present since a entire kernel support $\mathcal{L}_{p_i}$ for a point p_i is contained within the

network $\tilde{\mathcal{L}}$ (*i.e.* $\mathcal{L}_{p_i} \in \tilde{\mathcal{L}}$), such as in the three kernel density functions on $\mathcal{L}_1^{(S)}$ in Figure 3.13, for example. In contrast, the second case is when a kernel centre lies on the buffer network $\mathcal{L}_{e_i^{(1)}} \in \mathcal{L}^{(1)}$. In this case, edge biases are exhibited because the range of the kernel support extends beyond the defined network $\tilde{\mathcal{L}}$. If this is the case, the kernel density function $K_{p_i}(q)$ is trimmed at the node $e_i{}^{(1)}$, such as in the case of the kernel density function on $\mathcal{L}_{e_1^{(1)}}$ in Figure 3.13, for example. Several methods have been proposed in the literature for correcting these edge biases [60, 92, 113, 217], but these methods are not applicable in the context of KDE defined on a network. The third and final case is when a kernel centre lies on a buffer network $\mathcal{L}_{e_i^{(\geq 3)}}\mathcal{L}^{(\geq 3)}$, in which case deriving and unbiased kernel density estimator is significantly more complex than in the first and second cases. Okabe *et al.* [193], however, provided some computational solutions for unbiased kernel density estimator applicable to the third case, which are discussed in the remainder of this subsection.

The first step, when attempting to specify an unbiased kernel density estimator, is to specify the probability density function $f(q)$. This, however, sounds somewhat contradictory, since the whole purpose of KDE is to estimate the unknown probability density function over the network $\tilde{\mathcal{L}}$. Nevertheless, for the purpose of conducting a theoretical examination in which bias is tested, a density function must be specified. While there exists a large variety of choices for $f(q)$, perhaps the most fundamental probability density function in spatial analysis is the uniform distribution function. This method of examination closely emulates the ordinary Kriging method whereby the expected value and variance of an unknown density function are derived under the assumption that this unknown function is uniform over a defined space [57]. There are also a number of merits associated with choosing the uniform distribution. First, the uniform distribution implies that points are assumed to be generated according to a binomial point process, thus implying the CSR hypothesis (as described in §3.2.3), which is often one of the first tests conducted in a PPA investigation. Secondly, any probability density function defined on a network $\tilde{\mathcal{L}}$ may be transformed into the uniform density function utilising the uniform network transformation (as described in §3.2.4). The bias produced by the kernel function estimator $K(q)$, denoted by $B(q)$, is merely the difference between the expected value of $K(q)$ and the true value of the function $f(q)$ (in this case, the uniform distribution) at the point q, expressed mathematically as

$$B(q) = \mathrm{E}[K(q)] - f(q) = \mathrm{E}[K(q)] - \frac{1}{|\tilde{\mathcal{L}}|}, \quad \text{for } q \in \tilde{\mathcal{L}},$$

where the expected value of $K(q)$ is

$$\mathrm{E}[K(q)] = \int_{p \in \tilde{\mathcal{L}}} K_p(q) f(p) \, \mathrm{d}p.$$

For the purpose of constructing an unbiased kernel density estimator, let the quantity $d(p, q)$ be the distance from a point p to an arbitrary point q. Furthermore, let $k(d(p, q))$ be a function characterised by the following four properties (for all $p, q \in \tilde{\mathcal{L}}$):

1. $\int_{q \in \tilde{\mathcal{L}}} k(d(p, q)) \, \mathrm{d}q = 1$,
2. there exists a real number $w > 0$ such that $k(d(p, q)) > 0$ if $d(p, q) < w$, and $k(d(p, q)) = 0$ if $d(p, q) \geq w$,
3. $k(d(p, q))$ is strictly non-increasing with respect to $d(p, q)$, and
4. $k(d(p, q))$ is continuous with respect to $d(p, q)$.

TABLE 3.1: *The most popular choices of kernel functions [94].*

Name	Function	Bounds
Gaussian	$k(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$	$(-\infty, \infty)$
Epanechnikov	$k(x) = \frac{3}{4}(1 - x^2)$	$[-1, 1]$
Uniform	$k(x) = \frac{1}{2}$	$[-1, 1]$
Triangular	$k(x) = (1 - \lvert x \rvert)$	$[-1, 1]$
Tri-weight	$k(x) = \frac{35}{32}(1 - x^2)^3$	$[-1, 1]$
Tri-cube	$k(x) = \frac{70}{81}(1 - \lvert x \rvert^3)^3$	$[-1, 1]$
Bi-weight	$k(x) = \frac{15}{16}(1 - x^2)^2$	$[-1, 1]$
Cosine	$k(x) = \frac{\pi}{4} \cos\left(\frac{\pi}{2}x\right)$	$[-1, 1]$

Such a unimodal function is called the *base kernel density function.* There exists a wide variety of choices for this function [94], as summarised in Table 3.1. It is worth noting that there is consensus in the literature that the choice of bandwidth of the kernel function has a more significant influence on the results obtained than the choice of base kernel function itself [9, 197, 198, 234, 247]. Okabe *et al.* [193], however, warned that the direct use of a base kernel function in a network space leads to biases at network nodes. They, therefore, proposed two kernel density functions for constructing an *unbiased* kernel density estimator, namely the *equal-split discontinuous kernel density function* and the *equal-split continuous kernel density function.* Only the latter is considered here as a continuous kernel function is often preferred over one that is discontinuous [194] and is, therefore, exclusively considered in this dissertation.

Consider the case where a node e_i is within a distance w from a kernel centre positioned at a point p (as illustrated in Figure 3.14), and let the quantity $n(e_i)$ denote the degree of node e_i. Then the equal-split kernel density function may be expressed in a piecewise manner as a function of the base kernel density function $k(d(p,q))$, which takes the following form:

$$K_p(q) = \begin{cases} k(d(p,q)) & \text{for } 0 \leq d(p,q) < 2d(p,e_i) - w, \\ k(d(p,q)) - \frac{n(e_i)-2}{n(e_i)} k(2d(p,e_i) - d(p,q)) & \text{for } 2d(p,e_i) - w \leq d(p,q) < d(p,e_i), \\ \frac{2}{n(e_i)} k(d(p,q)) & \text{for } d(p,e_i) \leq d(p,q) < w. \end{cases} \tag{3.12}$$

If one were to consider the case where $q = e_i$, the second case in (3.12) would reduce to

$$k(d(p,e_i)) - \frac{n(e_i) - 2}{n(e_i)} k(2d(p,e_i) - d(p,e_i)) = \frac{2}{n(e_i)} k(d(q,e_i)),$$

which is identical to the third case in (3.12) when evaluated at node e_i. This confirms the continuity of the kernel density function $K_p(q)$ at node e_i, thereby satisfying Property 4 of the kernel density function $k(d(p,q))$. Note that Property 1 is also satisfied, since

$$\int_0^{2x-w} k(x)\,\mathrm{d}x + \int_{2d(p,e_i)-w}^{d(p,e_i)} \left[k(x) - \frac{n(e_i) - 2}{n(e_i)} k(2d(p,e_i) - x)\right] \mathrm{d}x + (n(e_i) - 1)\frac{2}{n(e_i)} \int_{d(p,e_i)}^{w} k(x)\,\mathrm{d}x = 1,$$

where $x = d(p,q)$. Okabe *et al.* [193] proved that $K_p(q)$ is indeed an unbiased estimator (a description of this proof is omitted due to its considerable length). It should also be noted that the equal-split kernel density function may include a wide variety of functions as special cases, provided that they also satisfy Properties 2 and 3 above.

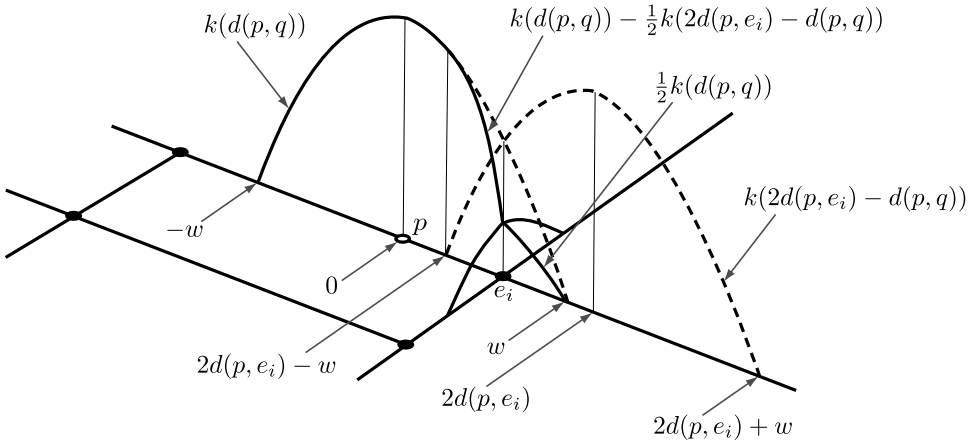


FIGURE 3.14: *An equal-split continuous kernel density function centred at point p, where $k(d(p,q))$ is a base kernel density function [194].*

3.4 Chapter summary

This chapter contained a review of the various models and techniques typically employed in the domain of network spatial analysis. The chapter opened in §3.1 with a description and data representation of the two most popular conceptual GIS modelling approaches, namely the object-based paradigm and the field-based paradigm. This facilitated a discussion in §3.2 on how to model these two conceptual paradigms mathematically. A discussion then ensued in the same section on how the binomial point process may be used to derive the CSR hypothesis for a network space, how the uniform network transformation may be used to transform any non-uniform distribution into one that is uniform, and finally, that cognisance should be taken of edge biases when defining an artificial boundary. Two PPA methods were finally discussed in §3.3, the first of which was the network NND method for testing whether points in a network are statistically close to one another or dispersed within the network space. The second method was the network KDE method for estimating the density of points on a network.

CHAPTER 4

Machine learning

Contents

4.1 An overview of machine learning 48
4.1.1 Machine learning paradigms *48*
4.1.2 The supervised learning procedure *49*
4.2 Data considerations 50
4.2.1 Data taxonomy *51*
4.2.2 Data cleaning *52*
4.2.3 Data partitioning methods *54*
4.3 Model considerations 55
4.3.1 Performance measures *56*
4.3.2 Overfitting vs *underfitting the training data* *57*
4.3.3 Model performance vs *model interpretability* *59*
4.4 Generalised linear models 61
4.4.1 General linear models *61*
4.4.2 The exponential family *62*
4.4.3 Constructing a generalised linear model *63*
4.4.4 Logistic regression *64*
4.5 Decision tree learning 65
4.5.1 Classification and regression trees *66*
4.5.2 Random forests *68*
4.5.3 The C4.5 algorithm *69*
4.6 The naive Bayes classifier 70
4.7 The k-nearest neighbours algorithm 71
4.8 Support vector machines 72
4.8.1 Linear separability *73*
4.8.2 Linear inseparability *76*
4.8.3 The application of kernels *77*
4.9 The multi-layer perceptron model 79
4.10 Chapter summary 80

The aim in this chapter is to introduce the reader to the notion of ML and a selection of algorithms that exist within this realm. The chapter opens in §4.1 with an overview of ML and the various associated paradigms in this field, together with a more in-depth discussion of supervised learning. This is followed in §4.2 by a review of the practical considerations associated with data in general and, thereafter, in §4.3, with additional considerations related to the appropriate application of ML algorithms. The reader is then introduced in §4.4 to the concept of *generalised linear models* (GLM) — a fundamental concept from which many classical statistical modelling techniques may be derived. The focus then shifts in §4.5 to a review of various tree-based ML algorithms, and this is followed, in §4.6, by a description of the Naive Bayes modelling approach. Thereafter, the *k-Nearest Neighbours* (kNN) learning algorithm is reviewed in §4.7 and this is followed in §4.8 by an in-depth description of one of the most popular ML algorithms — *Support Vector Machines* (SVMs). The last ML algorithm, discussed in §4.9, is the *multi-layer perceptron* (MLP) which is fundamental in the field of deep learning. The chapter closes in §4.10 with a brief summary of the contents presented.

4.1 An overview of machine learning

The term "machine learning" was first coined by Arthur Samuel in 1954 upon the publication of his self-learning program which could successfully play the game of Checkers [232]. Although not literally stated in his paper, the definition of ML as "*the field of study that gives computers the ability to learn without being explicitly programmed*" is often attributed to Samuel. Formally stated, ML involves constructing computer-based algorithms that are able to learn generalised rules from data to make predictions or determinations about a real-world phenomenon without relying on rule-based programming [51, 209]. Nowadays, ML algorithms (or learning models) are employed in a wide variety of useful applications such as recommender systems, fraud detection, weather forecasting, medical diagnostics, robot navigation, and image classification [157], to name but a few.

4.1.1 Machine learning paradigms

The majority of ML algorithms reside, for the most part, in one of four broad learning paradigms which may be described as follows:

Supervised learning. In a *supervised learning* paradigm, an algorithm is presented with a set of *labelled* data comprising input examples together with *known* corresponding output values deemed of particular interest. The task of the algorithm is to approximate the underlying function which maps these input examples to their corresponding output values [173]. In the case where output values are categorical, the problem is called a *classification* problem while in the case of continuous output values, it is called a *regression* problem.

Unsupervised learning. In an *unsupervised learning* paradigm, an algorithm is presented with a set of *unlabelled* data for which there are *no* known corresponding output values. The task of the algorithm is to identify some underlying structure (or pattern) between the various input values [158]. This is typically achieved through *dimensionality reduction* techniques (which aim to project the data onto a lower dimensional space), *density estimation* methods (which attempt to construct a functional estimate of the observed data), and/or though the application of various *clustering* approaches (whereby similar input examples are grouped together). This paradigm is the primary task of *exploratory data analysis* (EDA) which

attempts to first make sense of the acquired data so that meaningful questions can be framed and posed [18].

Semi-supervised learning. In a *semi-supervised learning* paradigm, an algorithm is typically presented with both labelled an unlabelled data. In this paradigm, the algorithm utilises both supervised and unsupervised learning techniques to accomplish *inductive* or *transductive* learning tasks. In the case of the former, unlabelled data are used in conjunction with a small amount to labelled data to improve the accuracy of a supervised learning task [68]. In the case of the latter, the aim is to infer the correct labels for each of the unlabelled examples as a way of avoiding the costly and time-consuming task of manual labelling.

Reinforcement learning. Lastly, in a *reinforcement learning* paradigm, an algorithm, commonly referred to as an *agent*, is exposed to a dynamic *environment* which is typically some abstraction of the real world. The task of the agent is to determine, through a trial-and-error process, the correct set of *actions* (collectively called a *policy*) which maximises (or minimises) some long-term cumulative *reward* (or punishment) [158]. Reinforcement learning differs from supervised learning in that, instead of correcting sub-optimal actions, the focus is on striking a balance between the *exploration* of uncharted territory and the *exploitation* of current knowledge [120].

Turning the attention back to the research related to RAs, Lord and Mannering [144] pointed out that the acquisition of detailed driving data (*e.g.* acceleration, braking and steering information) would better enable the identification of cause-and-effect relationships in the context of RAs. Unfortunately, this type of data is typically seldom available. Consequently, researchers have framed their analytic approaches toward studying the contributing factors that lead to dangerous RAs in some geographical location (*e.g.* a road segment or junction) over some specified period of time since measurable spatial and temporal attributes describing historically recorded RAs are more frequently available. These analysis approaches commonly take the form of the application of statistical and supervised learning models, with the goal of accurately predicting some measure of the severity of an accident. As a result, it is supervised learning that has drawn the attention of the author for further investigation and implementation in this dissertation.

4.1.2 The supervised learning procedure

The primary goal of any supervised learning process is to choose an appropriate ML algorithm that can accurately model an underlying relationship between the variables in a data set. The standard process by which this is achieved is illustrated graphically in Figure 4.1, where the process begins by developing a well-formulated research question. This research question may, in the context of an investigation related to RAs, for example, be *"What vehicle manoeuvre combination leads to the most dangerous accidents between two vehicles?"* To answer this research question by means of supervised learning, a researcher must first transform one or more sets of raw data (deemed relevant to the current investigation) by means of a host of data preprocessing techniques into a set of prepared data. The process of ensuring that the data are appropriately prepared for the application of an ML algorithm is often iterative in nature, usually requiring the data to be reshaped in such a way that one may draw the best scientific conclusions from them. Once a set of appropriately prepared data has been established, the researcher may apply a selection of supervised ML algorithms to the data in an attempt at predicting a specific feature that is able to provide insight into the original research question. Again, this is an iterative process, whereby the performance of each candidate model is trained and evaluated in respect

of the prepared data until an appropriate model has been found. Thereafter, the researcher may draw insight from this model by predicting the outcome of new scenarios presented to the model.

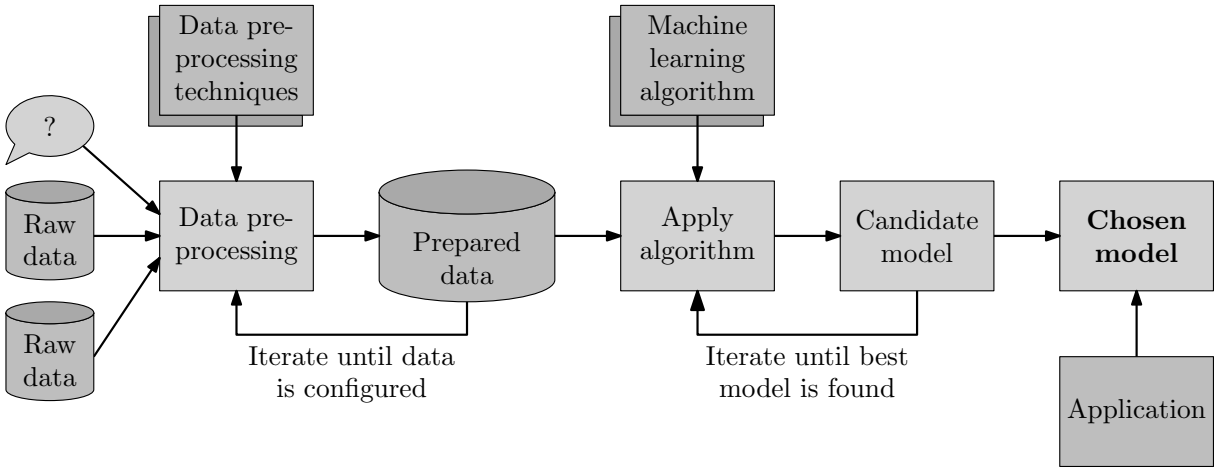


FIGURE 4.1: *The supervised learning process [121].*

As a way of establishing a set of formal notation and naming conventions for use in the remainder of this chapter, the "input" variables, hereafter referred to as *explanatory variables*, are denoted by the set $\mathcal{X} = \{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(m)}\}$ of m vectors, where the ith vector of explanatory variables $\boldsymbol{x}^{(i)} = \langle x_1^{(i)} \cdots x_n^{(i)} \rangle$ has values associated with n *explanatory attributes*[1] of interest. All explanatory variables correspond to a set of "output" variables, hereafter referred to as the *target variables*, which are denoted by $\mathcal{Y} = \{y^{(1)}, \ldots, y^{(m)}\}$, where only one *target attribute* defines this set. A pair of the form $(\boldsymbol{x}^{(i)}, y^{(i)})$, therefore, denotes the ith *observation* recorded for a specific event, where a vector $\boldsymbol{x}^{(i)}$ of explanatory variables is mapped to a single target variable $y^{(i)}$. A data set $\{(\boldsymbol{x}^{(1)}, y^{(1)}), \ldots, (\boldsymbol{x}^{(m)}, y^{(m)})\}$ of m independent observations of a specific event is called a *sample set*. Given such a sample set, the goal of a supervised learning algorithm, hereafter referred to as a *learning model*, is to learn a function $h : \mathcal{X} \mapsto \mathcal{Y}$ such that, for an unseen vector $\boldsymbol{x}$, $h(\boldsymbol{x}) = y$ is a "good" predictor of the actual target variable y. The function $h(\boldsymbol{x})$ is called the *hypothesis* of the learning model and may predict either a categorical or a continuous target variable. Furthermore, if the fundamental working of a learning model is based on a specific assumption of the distribution of observations, it is called a *parametric* model, while learning models that do not make any explicit assumptions about the distribution of observations are termed *non-parametric* models.

4.2 Data considerations

Since the primary input to all ML algorithms is data, one must be aware of the many consideration associated with the appropriate understanding, management and manipulation of data in the context of ML. More specifically, one should be able to: (1) Identify the many formats in which attributes in a data set may be expressed, (2) understand the origin of erroneous and missing values in a data set and how they should be corrected appropriately, and (3) be aware of the various data partitioning methods used during the performance evaluation of a learning model. These topics are discussed individually in this section.

[1]The term *attribute* is used (as opposed to the more common ML term *feature*), in this case, to highlight the analogous meaning of the term to that previously presented in the GIS literature review of §3.1.1.

4.2.1 Data taxonomy

According to Baltzan and Phillips [10], *information* may be seen as *raw data* that have been transformed, through various processes and procedures, into meaningful and understandable content which can ultimately be used to support decision making. Raw data, on the other hand, may be described as data that have not undergone any form of preprocessing [225], and are typically classified according to the inherent structure by which the data were captured, called the *data format*.

The format of data can, in its most primordial form, be classified into two broad categories: Data can either be inherently *structured* or *unstructured*. Structured data, for the most part, exhibit a high degree of organisation so that they can be interpreted easily by a computer, while unstructured data do not follow any conventional data structure [26]. Jiang *et al.* [117] presented a more descriptive classification which groups data formats into to four broad categories, namely: (1) *textual data*, (2) *temporal data*, (3) *transactional data* and (4) *relational data*. The first of these, textual data, are, in essence, a large collection of characters that convey meaning by forming words which, in turn, form sentences. These may, for example, manifest themselves in the form of documents, texts or social media posts — all of which are highly unstructured. Temporal data, on the other hand, are commonly structured and represent information that changes over time (typically referred to as *time-series* data). Stock market graphs and the daily air temperature are examples of inherently temporal data. Transactional data describe one or more objects related to a specific event and, therefore, also contain a time dimension. An example of data in this format is a list of purchased products captured in historical retail store transactions. Finally, and probably the most widely-used data format, is that of a relational data. In this format, data are typically organised in the form of a *tabular* data structure, where each row denotes a unique *record* and each column represents an *attribute* which conforms to a specific *data type*.

The data type of an attribute describes the inherent values a data item can assume and may, in its most basic form, be grouped into one of two categories, namely as a *qualitative* (categorical) attribute or a *quantitative* attribute. The latter describes an attribute which is inherently captured on a numerical scale, such as the velocity or mass of a vehicle [58, 114]. Although all quantitative attributes are numeric, all numeric attributes are not necessarily quantitative. A person's identification number may, for example, be numeric, but not necessarily quantitative. Qualitative attributes, on the other hand, classify all data items into a set of distinct categories which describe the inherent characteristics of the object under consideration. An optometrist record may, for example, describe a client's eye colour as being either blue, green or brown. Both qualitative and quantitative attributes may be further classified according to the hierarchical categories shown in Figure 4.2.

Qualitative attributes may further be categorised into attributes that are either *nominal* or *ordinal* in nature [104], where the former contains two or more categories that are all assumed to be of equal importance (*i.e.* there exists no meaningful sequence in which to order the categories of an attribute). One cannot, for example, say that a brown eye colour is more important than a blue eye colour. Nominal attributes can be further decomposed into attributes that are binary in nature, referred to as *dichotomous*, or contain more than two categories, in which case they are referred to as *multichotomous*. In contrast to nominal attributes, ordinal attributes contain categories that *are* intrinsically ranked [23]. In a survey, for example, one may be required to rate a customer experience as either: "Bad," "Satisfactory" or "Good." It should be noted, however, that, although ordinal attributes express an inherent ordering of the associated categories, the differences between attribute categories is not necessarily equal [104].

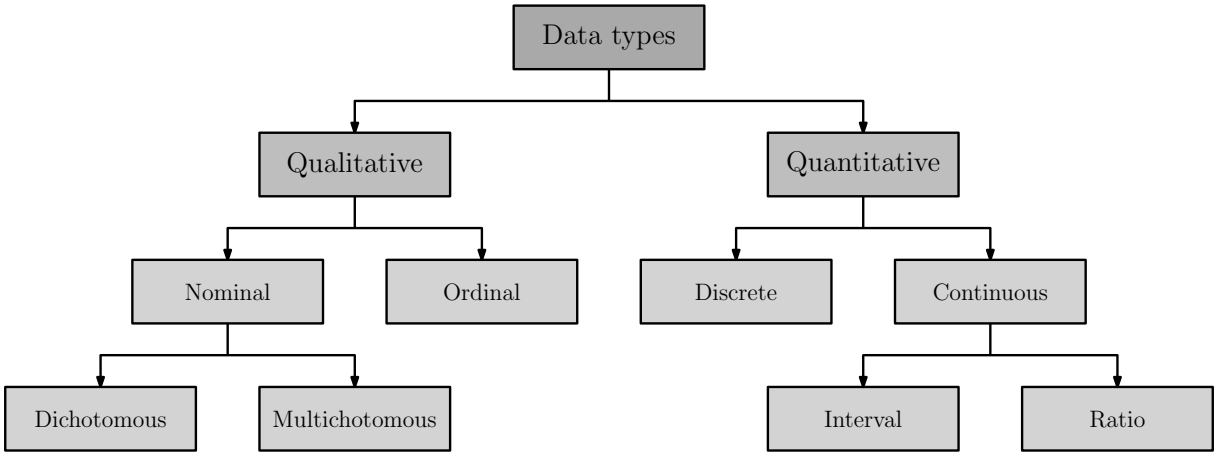


FIGURE 4.2: *A taxonomy of the various data types [256].*

Quantitative attributes may further be categorised into attributes that are either *discrete* or *continuous* in nature. Discrete attributes can assume a single value from a finite set of values, such as the number rolled on a six-face die, for example. All qualitative attributes are inherently discrete since they are countable and can assume only a certain set of values. Continuous attributes, on the other hand, can assume one of an uncountably infinite set of values (which exist along a continuum), such as the temperature in a room, for example. Continuous attributes can further be categorised into attributes which assume values on an *interval* scale or a *ratio* scale. The latter assumes that there exists a so-called *true zero* which reflects the absence of the specific measured feature. An object's weight is, for example, a ratio attribute since a value of zero indicates that the object has no weight. Interval attributes, on the other hand, do not assume this property. Temperature measured in degrees Fahrenheit ($^\circ F$) or Celsius ($^\circ C$) is, for example, an interval attribute since zero $^\circ F$ or $^\circ C$ does not imply no temperature. If, however, temperature is measured in Kelvin, then it may be expressed as a ratio estimate, since zero Kelvin indicates the absence of temperature.

4.2.2 Data cleaning

Probably the most important and, at the same time, mundane tasks associated with the data preparation process is that of *data cleaning* [207]. The purpose of data cleaning is to remove any errors in and/or missing values from a data set so as to ensure that the data being used to formulate a learning model's hypothesis is of a high quality, thereby instilling confidence in the results obtained. This typically involves three types of activities, namely: (1) correcting erroneous values, (2) *imputing* (filling in) missing data entries and (3) detecting outliers [187]. Examples of these occurrences are illustrated in Table 4.1, where the first column denotes the data attribute the recorder attempted to capture, the second column describes the true data that should have been entered by the data recorder, and the last column contains the actual data recorded for an observation. As may be seen, the values for the attributes `Annual income` and `Ethnicity` have been left blank, the attributes `Name`, `Date of birth` and `Place of birth` are incomplete, and the attribute `Sex`, although in the correct format, does not reflect the true value.

According to Hellerstein [106], there exists four main origins of erroneous values in a data set. The first of these is described as *data entry errors* which are primarily attributed to the typographic and misinterpretation errors humans make when attempting to capture the data.

TABLE 4.1: *Examples of erroneous and missing data values [65].*

Attribute	True data	Actual data
Name	Maria Margaret Smith	Maria Smith
Date of birth	2/19/1981	1981
Sex	F	M
Ethnicity	Hispanic and Caucasian	
Place of Birth	Nogales, Somora, Mexico	Nogales
Annual income	$50,000	

The second origin is described as *measurement errors* which occur when measuring a physical property is not conducted according to a proper procedure. The third origin is described as *distillation errors* which occur during the preprocessing and summarising phase before subsequently being added to a system database. The final origin is described as *data integration errors* which occur when multiple data sets are integrated into a single database.

Since the majority of learning models require a complete data set in order to function properly, it is necessary to address any missing data in a data set. Obviously, the best approach would be to manually fill in all incomplete entries to contain the correct information. This is, however, not always practical, especially if there is a significant proportion of missing values in a considerably large data set. The second approach, which may seem more reasonable, would be to delete any observations with missing values entirely, leaving only the observations that contain values for all of the defined attributes. This may, however, result in a substantial loss in the overall set of observations, to the point where there are too few data to conduct a meaningful investigation. Another approach is to identify those attributes in the data set which possess a large proportion of missing values. If the proportion of missing values of an attribute exceed a specific threshold, then the attribute may be disregarded altogether. The final approach would be to employ a specific imputing technique where missing values are filled according to some specified rule. Many imputing methods have been proposed in the literature, namely: *Mean imputation* (replace missing values with the average/most frequent value), *hot deck imputation* (replace missing values with randomly chosen observations of similar values), *cold deck imputation* (replace missing values with systematically chosen observations of similar values), *random imputation* (replace missing values with random values) and *model-based imputation* (construct a learning model to predict missing values based on values of complete observations) [65, 91].

The final activity in the data cleaning process is concerned with the detection of any outliers that may be present in a data set, where an outlier is described as an observation that does not conform to the general pattern observed in a data set [177]. Aguinis *et al.* [3] described a set of best-practice recommendations concerned with *defining*, *identifying*, and *handling* outliers. There exists a wide variety of statistical techniques for the purposes of outlier definition and identification, which are typically described as *univariate* (identifying extreme values of a single attribute), *multivariate* (identifying unusual value combinations for two or more attributes), *parametric* (based on underlying distribution assumptions) and *non-parametric* (not based on any distribution assumptions) methods [288]. The appropriate handling of outliers is, however, problem-specific [187]. If the aim of a study is to identify a general pattern in the data, then outliers are typically removed. If, on the other hand, the goal of a study is to detect unusual observations (*e.g.* a fraudulent bank transactions), then outliers provide essential information.

4.2.3 Data partitioning methods

The initial sample set is typically partitioned into three separate data sets, namely a *training set*, a *validation set* and a *testing set* [31], as shown in Figure 4.3. The training set is used to find optimal parameter settings (or weights) for a learning model so that it fits the set of training examples. The validation set (sometimes called the *hold-out set*) is used to provide an unbiased evaluation of a learning model's fit in respect of the training set while also tuning the hyper-parameters of the model. For this reason, the combined training and validation sets are often referred to as the *learning set*. The evaluation of a model, however, becomes more biased as its predictive ability in respect of the validation set is incorporated into the learning model's configuration [238]. It is, therefore, necessary to evaluate a learning model's performance in respect of an independent test set so as to provide a completely unbiased evaluation of a final model fit in respect of the training set. In order to enhance the performance of learning models, various *data re-sampling* methods may be employed to construct the training and validation sets [114], two of the most popular being the method of *cross-validation* and the *bootstrap re-sampling* method.

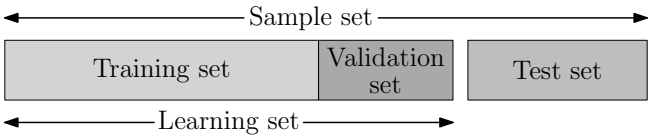


FIGURE 4.3: *Sample data partitioned into the training, validation and testing sets.*

Cross-validation

There are two main types of cross-validation methods, namely: *k-fold* cross-validation and *leave-one-out* cross-validation. In the case of k-fold cross-validation, the set of m observations is partitioned into k equal (or as close to equal as possible) parts (or folds), as illustrated graphically in Figure 4.4. A total of $k-1$ folds of the data are then used for training a learning model, while the remaining fold is used for model validation. This process is repeated a total of k times, each time using a different fold for validation purposes. The final performance of a learning model is then estimated as the average of the misclassification errors seen during the k repetitions.

The leave-one-out method of cross-validation is a special case of the k-fold cross-validation that occurs when $k = m$ (*i.e.* k is equal to the number of observations). As its name suggests, all observations, except one (*i.e.* $m-1$ observations), are used to train the learning model according to this method, and the remaining observation is used for model validation. This procedure is thus repeated a total of m times, where the final performance of a learning model is estimated as the average of the m misclassification errors.

Bootstrap re-sampling

The second type of data sampling method is that of *bootstrap re-sampling* (or just *bootstrapping*), where the aim is to create a *bootstrap sample* of observations for training and validating learning models — the method by which it achieves these subsets is, however, quite different. Bootstrapping involves the repeated selection of observations from the learning set *with* replacement [271] to construct a new set of any cardinality, denoted by L (since observations can be reused). Given

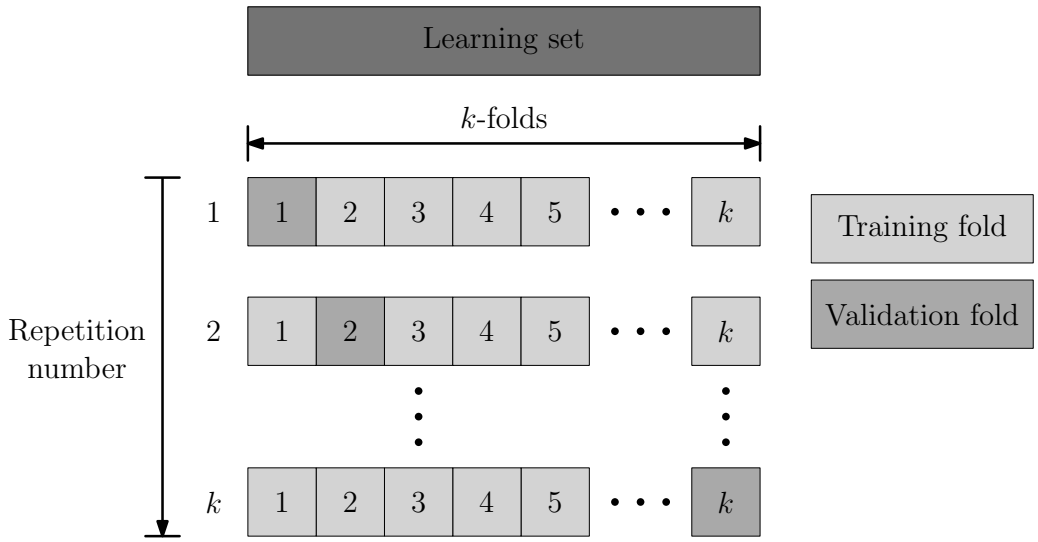


FIGURE 4.4: *The k-fold cross-validation procedure [256].*

a specified validation-training proportion split p, validation and training sets of cardinalities pL and $(1-p)L$, respectively, may be defined a total of b times, as illustrated in Figure 4.5. The final performance of a learning model is then estimated as the average of the misclassification errors achieved over the b validating sets. This method, therefore, stands in contrast to k-fold cross-validation which is limited by the size of the learning set and the number of folds.

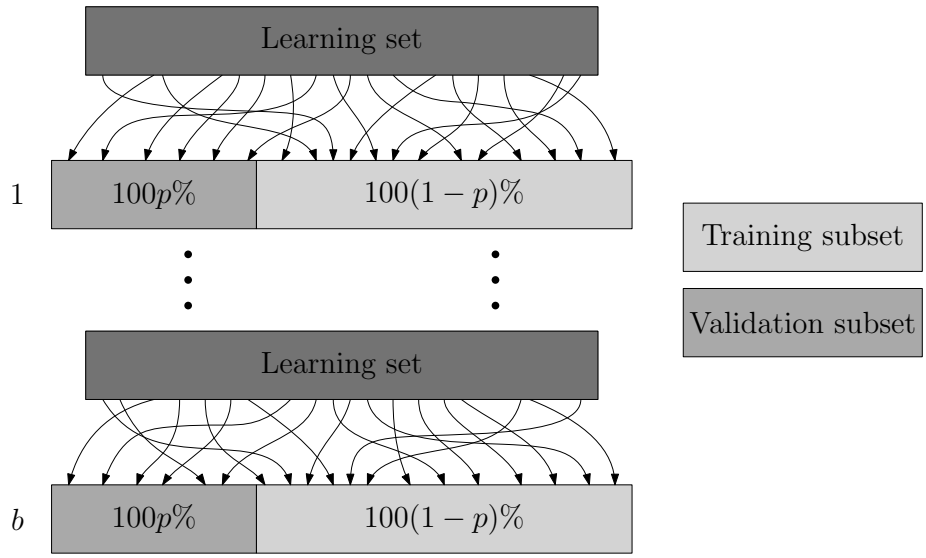


FIGURE 4.5: *The bootstrap re-sampling procedure [256].*

4.3 Model considerations

As previously described in §4.1.2, the primary goal of any supervised learning procedure is choosing an appropriate model. One must, therefore, be aware of various notions that may assist in selecting, configuring and evaluation a model for a specific learning task. More specifically, one should be able to: (1) Evaluate the performance of a learning model according to an appropriate

performance measure, (2) identify whether a learning algorithm is generalising the data, and (3) appreciate the trade-off associated with employing a complex learning model rather than one that is easier to understand. These topics are discussed in this section.

4.3.1 Performance measures

Since the evaluation of a learning model is essential in any ML investigation, one must be able to choose a performance measure that is most appropriate when presented with either a classification or regression task. In the case of the latter, the most popular performance measures are the well-known *mean square error*, *mean absolute error* and *R-squared error* [236]. In the case of a classification task, however, the performance evaluation of a learning model is more involved. The performance of a classifier is typically first described in the form of a confusion matrix, as illustrated in Table 4.2 (for a binary classification task), where each row of the matrix represents the *actual* category of the observations while each column represents the observation category *predicted* by the learning model.

TABLE 4.2: *A confusion matrix showing the possible outcomes of a learning model's predictions.*

		Predicted category	
		Positive	Negative
Actual category	Positive	True positive (TP)	False negative (FN)
	Negative	False positive (FP)	True negative (TN)

Perhaps the best-known performance measure is *classification accuracy* (CA), which describes the total proportion of correct predictions made by a learning model, expressed mathematically as

$$\mathrm{CA} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN} + \mathrm{TN}}. \tag{4.1}$$

Although CA seems to be an intuitive measure by which to assess a classifier, it should never be used when there is a large category imbalance in a data set. Consider, for example, a case in which 950 positive observations and 50 negative observations are contained in a data set. Then a learning model can easily achieve a CA of 95% simply by always predicting the majority category (*i.e.* the positive category).

In order to overcome this flaw, the performance of a learning model can instead be described in terms of two separate measures, namely *precision* and *recall*. Precision describes what proportion of the observations predicated as positive are actually positive observations, expressed mathematically as

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}. \tag{4.2}$$

Recall (also known as *sensitivity*) describes what proportion of the observations that are actually positive were correctly classified as positive, expressed mathematically as

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}. \tag{4.3}$$

Although these two performance measures can be used together to measure the performance of a single learning model, in the case when two learning models are compared, it is not always clear which learning model achieves superior performance. For this reason, an alternative performance

measure, called the F_β-score is defined as

$$F_\beta = \left(\frac{1}{\beta \left(\frac{1}{\text{Precision}} \right) + (1-\beta) \left(\frac{1}{\text{Recall}} \right)} \right). \tag{4.4}$$

This score is a weighted harmonic mean of precision and recall in which β is a weight parameter used to specify the importance of precision over recall. The choice of β is, to a large extent, problem-specific. A value of β close to one essentially reduces (4.4) to (4.2), while a value of β close to zero reduces (4.4) to (4.3). A β-value of $b = 0.5$ weights both precision and recall equally and is known in statistics as the F_1-score.

Arguably the most popular classification performance measure is the *area under curve* (AUC) performance measure [102, 111, 224] as a result of its robustness against significant category imbalances [76]. As its name implies, AUC measures the area under a specific curve, called the *receiver operating characteristic* (ROC) curve. An ROC curve is a graphical plot of the predictive ability of a binary classifier. Consider, for example, a binary classification problem in which a learning model predicts a category probability (*i.e.* a value between zero and one) for a set of observations, where one denotes a strong positive category prediction and zero indicates a strong negative category prediction. The two distributions of the actual positive and actual negative observations are illustrated graphically in Figure 4.6(a) according to the category probabilities predicted by the learning model. An ROC curve (like that shown in Figure 4.6(b)) may, therefore, be constructed by varying the classification threshold (indicated in Figure 4.6(a)) from a category probability of zero to one and plotting the learning model's *true positive rate* (TPR) against its *false positive rate* (FPR) for all possible classification thresholds. The TPR of a learning model may be expressed mathematically as

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} = \text{Recall},$$

while its FPR may be expressed as

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}.$$

The area under the ROC curve (or just the AUC) indicates to what extent a model is able to distinguish between positive and negative observations. The AUC score may also be interpreted as the probability that a learning model ranks a randomly chosen positive observation higher than a random negative observation. An AUC score of one is, therefore, the most ideal scenario as it indicates that a learning model can perfectly distinguish between positive and negative categories. An AUC score of zero indicates that the learning model is, in fact, reciprocating the categories (*i.e.* positive observations are always predicted as negative and *vice versa*). In actuality, this is also an ideal scenario since merely inverting the predictions of the learning model would produce a perfect classifier. An AUC score of 0.5 (indicated by the area under the dashed line in Figure 4.6(b)), on the other hand, is the worst-case scenario as it indicates that a learning model has no discrimination capacity to distinguish between the positive and negative categories (*i.e.* the learning model is no better than random guessing) [261].

4.3.2 Overfitting *vs* underfitting the training data

One of the most fundamental concepts to grasp in ML is the notion of a learning model *underfitting* or *overfitting* a set of training data. Underfitting occurs when a learning model does

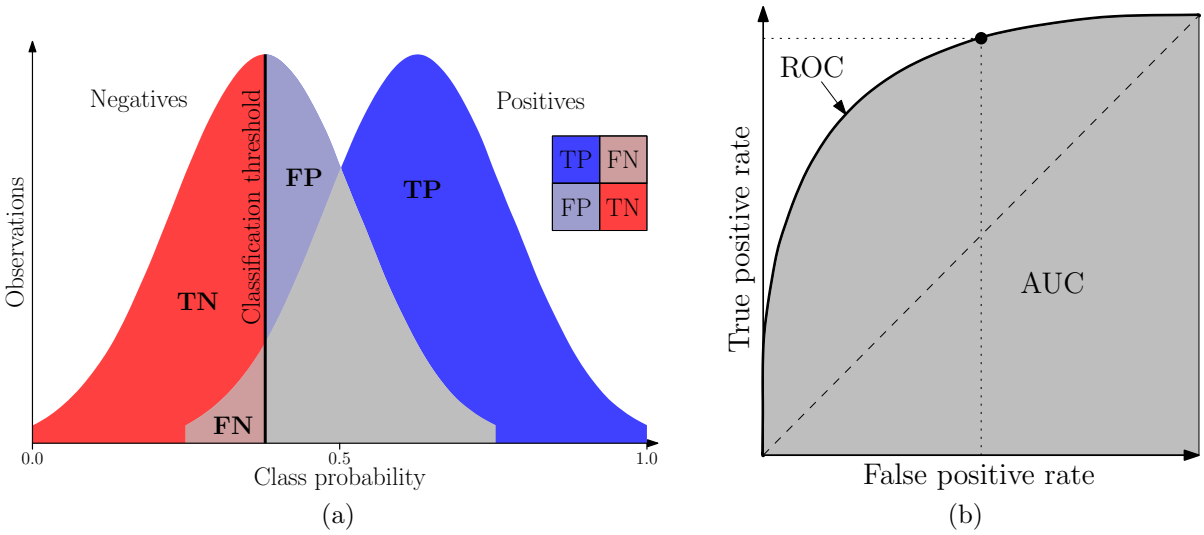


FIGURE 4.6: *(a) The two distributions of the actual positive and negative observations according to the category probabilities predicted by a learning model are used to produce (b) an ROC curve by varying the classification threshold and plotting the TPR of the learning mode against its FPR.*

not perform well in respect of either the training set or the validation set, as illustrated by a learning model with a complexity of "A" in Figure 4.7. In this case, the model is too simple and cannot model the underlying pattern (or trend) in the training set, resulting in a poor

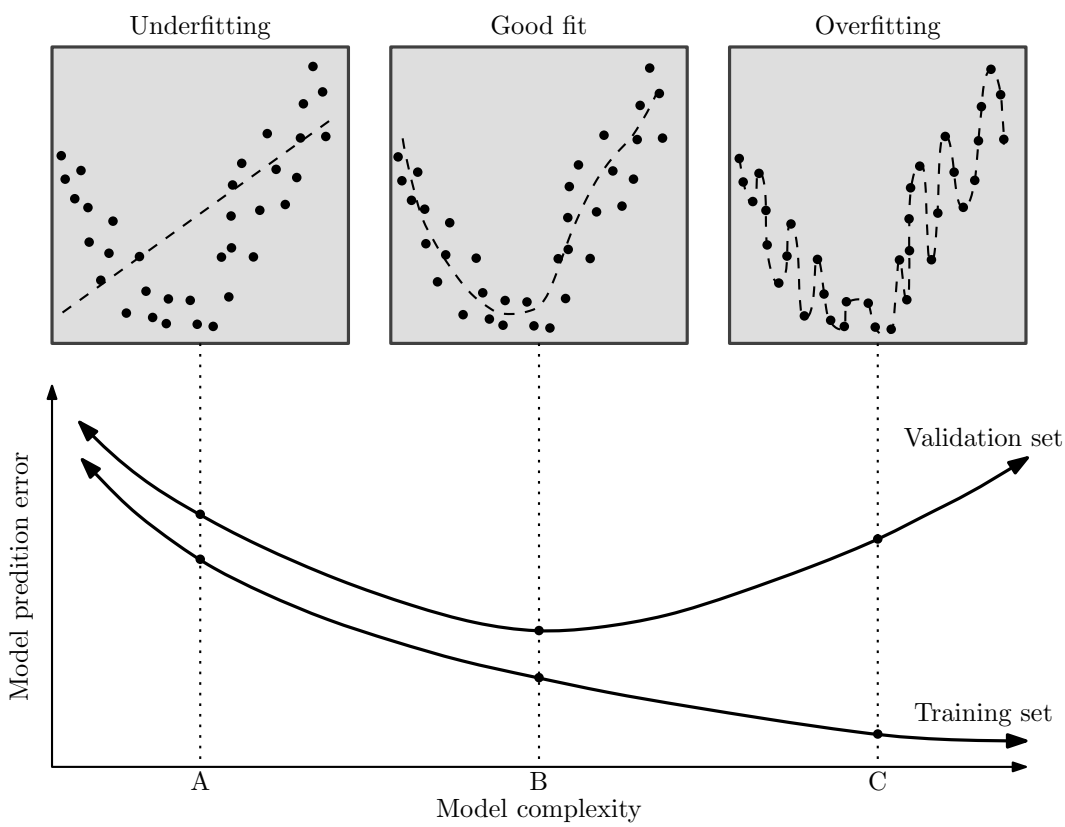


FIGURE 4.7: *An illustration of how the complexity of a learning model results in either underfitting or overfitting the training set, resulting in large prediction errors in respect of the validation set.*

ability to predict unseen observations accurately. Overfitting, on the other hand, occurs when a learning model performs well in respect of the training set, but significantly poor in respect of the validation set, as illustrated by a learning model with a complexity of "C" in Figure 4.7. In this case, the learning model is not learning the underlying trend. It is rather memorising the observations in the training set, again resulting in a poor generalisation ability (and thus a poor prediction performance) in respect of unseen observations. An ideal learning model is, therefore, one that achieves a suitable trade-off between underfitting and overfitting the observations in the training set, as illustrated by a learning model with a complexity of "B" in Figure 4.7. In this case, the learning model is able to generalise the underlying trend displayed appropriately by the observations in the training set which result in a better overall prediction performance in respect of unseen observations.

4.3.3 Model performance *vs* model interpretability

Understanding why an ML model makes a specific prediction may, in many applications, be just as crucial as the accuracy of the prediction itself. A model that is interpretable creates a certain degree of trust in its predictions, provides insights into the process being modelled, and highlights potential biases present in the data under consideration. The lack of interpretability of many ML models (often called *black-box* models) is a frequent barrier to their adoption, since end-users typically prefer solutions that are understandable to a human [176]. This may require researchers to sacrifice models which achieve superior predictive ability in favour of models that yield more understandable results (often called *white-box* models), but possibly achieve a lower overall performance. In the ML literature, this is referred to as the trade-off between the accuracy and interpretability of a model's output.

The ethical considerations associated with autonomous vehicles, for example, are often called into question since the learning models (typically reinforcement learning algorithms [61]) used to control such a vehicle are inherently black-box in nature. Consequently, if an autonomous vehicle is involved in an RA, humans are not able to understand *why* an autonomous vehicle made the manoeuvres it did. This has significant legal implications as the notion of accountability comes into question, especially if an autonomous vehicle RA results in fatalities [19]. Researchers have, therefore, been investing more time into developing techniques that provide insights into the working of specific black-box learning models so that humans may begin to understand the methods by which specific models arrive at their conclusions [136]. A wide variety of methods have recently been proposed for addressing this issue [8, 64, 140, 215, 246, 259]. The understanding of how these methods are related and which method is preferable to another was, however, still unclear until Lundberg and Lee [150] proposed their idea of unifying these methods into a single approach for interpreting model predictions, which they termed *Additive Feature Attribution* (AFA) methods.

An AFA method seeks to locally approximate a complex, non-linear ML model by using a simpler, linear model, called an *explanation model*, whose predictions are much easier to interpret. As a way of explanation, let f be the original learning model to be explained and let g be the associated explanation model. Most explanation models are expressed in terms of so-called *simplified inputs* $\boldsymbol{x}' = \langle x'_1 \cdots x'_N \rangle$ that map to the original inputs through a mapping function $\boldsymbol{x} = \boldsymbol{h}_{\boldsymbol{x}}(\boldsymbol{x}')$. These local approximation methods, therefore, attempt to ensure that $g(\boldsymbol{z}') \approx f(\boldsymbol{h}_{\boldsymbol{x}}(\boldsymbol{z}'))$ whenever $\boldsymbol{z}' \approx \boldsymbol{x}'$. An AFA method is, therefore, one in which its explanation model is a linear function of binary variables which takes the form

$$g(\boldsymbol{z}') = \phi_0 + \sum_{i=1}^{N} \phi_i z'_i, \tag{4.5}$$

where $\phi_0, \ldots, \phi_N \in \mathbb{R}$, $\boldsymbol{z}' \in \{0,1\}^N$ and $\phi_0 = f(\boldsymbol{h}_{\boldsymbol{x}}(\boldsymbol{0}))$ [150]. In essence, an importance ϕ_i is attributed to each of the N features and, by summing the effects of all feature attributions, an approximation is obtained of the original model output $f(\boldsymbol{x})$.

Lundberg and Lee [150] demonstrated that many current methods in the literature conform to the representation in (4.5) but it was still not obvious how to compare one AFA method with another appropriately. They therefore went on to define three desirable properties that any AFA method should exhibit: The first is the so-called *local accuracy property* which requires the output of the explanation model g to match the original model f for the simplified input $\boldsymbol{x}'$ being explained (*i.e.* $f(\boldsymbol{x}) = g(\boldsymbol{x}') = \phi_0 + \sum_{i=1}^{N} \phi_i x_i'$). The second is the *missingness property* which specifies that a missing feature should have no attributed importance (*i.e.* $x_i' = 0 \implies \phi_i = 0$). And finally, the third property is that of *consistency* which states that if feature i always has the same or a greater impact in model f' than in model f, then the importance ϕ_i of feature i should be greater in model f' than in model f, described mathematically as

$$f_x'(\boldsymbol{z}') - f_x'(\boldsymbol{z}' \backslash i) \geq f_x(\boldsymbol{z}') - f_x(\boldsymbol{z}' \backslash i) \text{ for all } \boldsymbol{z}' \in \{0,1\}^N \implies \phi_i(f', \boldsymbol{x}) \geq \phi_i(f, \boldsymbol{x}),$$

where $f_x(\boldsymbol{z}') = f(\boldsymbol{h}_{\boldsymbol{x}}(\boldsymbol{z}'))$ and $\boldsymbol{z}' \backslash i$ denotes setting $z_i' = 0$.

It follows from the theory of cooperative games that there exists only one solution for the values $\phi_0, \ldots, \phi_N$ in (4.5) which always satisfies the three aforementioned properties:

$$\phi_i(f, \boldsymbol{x}) = \sum_{\boldsymbol{z}' \subset_0 \boldsymbol{x}'} \frac{|\boldsymbol{z}'|!(N - |\boldsymbol{z}'| - 1)!}{N!} [f_x(\boldsymbol{z}') - f_x(\boldsymbol{z}' \backslash i)], \tag{4.6}$$

where $|\boldsymbol{z}'|$ is the number of non-zero entries in $\boldsymbol{z}'$, $\boldsymbol{z}' \subset_0 \boldsymbol{x}'$ represents all the vectors for which the non-zero entries are a subset of the non-zero entries in $\boldsymbol{x}'$, and the values ϕ_i are known as Shapley values [244]. In essence, the importance ϕ_i of feature i is determined by comparing a model's prediction with and without the feature, weighted over all possible feature combinations. Due to the large computational cost associated with exploring all possible feature combinations, sampling techniques used to approximate the Shapely values in (4.6) are commonly employed [150].

Using this result, Lundberg and Lee [150] proposed a unified measure of feature importance, called SHAP values, which are defined as the Shapley values of a conditional expectation function of the original model, and are thus solutions to (4.6) for which $f_x(\boldsymbol{z}') = f(\boldsymbol{h}_{\boldsymbol{x}}(\boldsymbol{z}')) = E[f(\boldsymbol{z}) \mid \{z_i = x_i \mid i \in \mathcal{S}\}]$, where $\mathcal{S}$ is the set of non-zero indices in $\boldsymbol{z}'$. By way of example, consider the illustration in Figure 4.8. In this case, $f(\boldsymbol{x})$ is the output predicted by the model, and $E[f(\boldsymbol{z})]$ is the base value (or average model output) which would be predicted if no features were present. If the feature value x_1 is included in the model, then ϕ_1 explains how the base value $E[f(\boldsymbol{z})]$ became the new predicted value $E[f(\boldsymbol{z}) \mid z_1 = x_1]$. Repeating this process for the remaining feature values x_2, x_3 and x_4, estimates of the SHAP values for ϕ_2, ϕ_3 and ϕ_4 are obtained, showing how the model ultimately arrives at the predicted output $f(\boldsymbol{x})$.

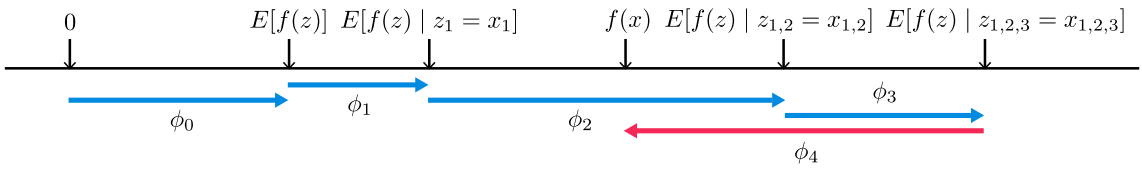


FIGURE 4.8: *The SHAP values ϕ_1, ϕ_2, ϕ_3 and ϕ_4 attributed to the $N = 4$ feature values x_1, x_2, x_3 and x_4, respectively, denote the change in the expected model prediction when conditioning on a feature which explains how the base value $E[f(\boldsymbol{z})]$ becomes the current model output $f(\boldsymbol{x})$ [150].*

4.4 Generalised linear models

The idea of generalising various statistical models into a single, overarching theoretical framework was first conceived in a 1972 paper by Nelder and Wedderburn [183]. In that paper, the authors showed that many of the classical linear statistical models share a number of properties which may be exploited collectively, and that they admit a common method for computing parameter estimates. These commonalities facilitate the study of so-called GLMs as a single class of models, as opposed to considering the various models as an unrelated collection of special topics. In this section, it is shown that the well-known OLSR and logistic regression parametric models are, in fact, special cases of the family of GLMs. In their most basic form, GLMs are an extension of the concepts proposed by its predecessor, *general linear models*[2], which will serve as a starting point for the impending discussion.

4.4.1 General linear models

In a general linear model, the fundamental assumption is that the set of target variables $\mathcal{Y} = \{y^{(1)}, \ldots, y^{(m)}\}$, now denoted by the vector $\boldsymbol{y}$, and the set of explanatory variables $\mathcal{X} = \{\boldsymbol{x}^{(i)}, \ldots, \boldsymbol{x}^{(m)}\}$, now denoted by the *design matrix* $\boldsymbol{X}$, are related linearly by an equation of the form

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{4.7}$$

where $\boldsymbol{\beta}$ is a column vector of unknown parameters, while $\boldsymbol{\epsilon}$ is a column vector containing independently and identically distributed error terms that capture any unmodelled effects or random noise in the data. It is assumed that the expected values $\boldsymbol{\mu} = E(\boldsymbol{y} \mid \boldsymbol{X}; \boldsymbol{\beta})$ of the target variables have an identity relationship with the linear predictor $\boldsymbol{X}\boldsymbol{\beta}$ in the sense that the hypothesis of the general linear model is given by $h(\tilde{\boldsymbol{x}}) = E(y \mid \boldsymbol{x}; \boldsymbol{\beta}) = \boldsymbol{x}\boldsymbol{\beta}$. The strong assumption made by a general linear model is that the error terms are assumed to be sampled from a Gaussian (or normal) distribution with zero mean and a constant variance σ^2, written as $\epsilon^{(i)} \sim \mathcal{N}(0, \sigma^2)$. The density of $\epsilon^{(i)}$ is, therefore, given by

$$p(\epsilon^{(i)}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(\epsilon^{(i)})^2}{2\sigma^2}\right),$$

which implies that

$$p(y^{(i)} \mid \boldsymbol{x}^{(i)}; \boldsymbol{\beta}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y^{(i)} - \boldsymbol{x}^{(i)}\boldsymbol{\beta})^2}{2\sigma^2}\right).$$

The distribution of the observed target variables $\boldsymbol{y}$ are consequently characterised by $\boldsymbol{y} \mid \boldsymbol{X}; \boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{\beta}, \sigma^2)$. Since each observation $(\boldsymbol{x}^{(i)}, y^{(i)})$ is assumed to be independent of all the other observations, the joint density or *likelihood* $L(\boldsymbol{\beta}) = L(\boldsymbol{\beta}; \boldsymbol{X}, \boldsymbol{y}) = p(\boldsymbol{y} \mid \boldsymbol{X}; \boldsymbol{\beta})$ of the data is given by the product of the individual probabilities

$$\begin{aligned} L(\boldsymbol{\beta}) &= \prod_{i=1}^{m} p(y^{(i)} \mid \boldsymbol{x}; \boldsymbol{\beta}) \\ &= \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y^{(i)} - \boldsymbol{x}^{(i)}\boldsymbol{\beta})^2}{2\sigma^2}\right). \end{aligned}$$

[2]The term *generalised linear model* and its abbreviation GLM are often confused with the term *general linear model*. The co-originator, John Nelder, has expressed regret over choosing this terminology [237].

One would like to obtain an estimate of $\boldsymbol{\beta}$ which maximises the value of $L(\boldsymbol{\beta})$, called the *maximum likelihood estimation* (MLE). Instead of maximising $L(\boldsymbol{\beta})$, one may also maximise any strictly increasing function of the likelihood function. As a result, it is often more convenient to maximise a logarithmic form of the likelihood function (due to its relatively simple differentiability), appropriately called the *log likelihood* $\ell(\boldsymbol{\beta})$, where

$$\begin{aligned}
\ell(\boldsymbol{\beta}) &= \log L(\boldsymbol{\beta}) && (4.8)\\
&= \log \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y^{(i)} - \boldsymbol{x}^{(i)}\boldsymbol{\beta})^2}{2\sigma^2}\right)\\
&= \sum_{i=1}^{m} \log \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y^{(i)} - \boldsymbol{x}^{(i)}\boldsymbol{\beta})^2}{2\sigma^2}\right)\\
&= m \log \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{2\sigma^2} \sum_{i=1}^{m} (y^{(i)} - \boldsymbol{x}^{(i)}\boldsymbol{\beta})^2. && (4.9)
\end{aligned}$$

From these results, one may conclude that the final choice of β-values does not depend on the variance σ^2 of the Gaussian distribution (this fact will prove useful later). Hence, for a general linear model, maximising $\ell(\boldsymbol{\beta})$ is equivalent to minimising the cost function

$$\sum_{i=1}^{m} (y^{(i)} - \boldsymbol{x}^{(i)}\boldsymbol{\beta})^2. \tag{4.10}$$

The expression (4.10) may be recognised as that occurring in the well-known OLSR method for estimating the unknown parameters in a linear regression model. The optimal parameter values $\boldsymbol{\beta}^*$ are, therefore, those that minimise the sum of the squared errors between the actual and predicted target variable, for all observations in the training set. For most statistical models, these parameter values are typically computed using a technique called *gradient decent*, the details of which are described by Ng [185]. In the case of OLSR, however, these $\boldsymbol{\beta}$-values may be computed analytically by solving the well-known normal equations [104], to yield

$$\boldsymbol{\beta}^* = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}.$$

4.4.2 The exponential family

The formulation of a GLM is traditionally achieved within the framework of the *exponential family* of distributions. A class of distributions is said to form part of the exponential family if their probability mass/density functions may be written in the form

$$p(y; \boldsymbol{\eta}) = b(y) \exp(\boldsymbol{\eta}^T \boldsymbol{T}(y) - a(\boldsymbol{\eta})), \tag{4.11}$$

where $\boldsymbol{\eta}$ is the so-called *canonical parameter* (or *natural parameter*) of the distribution, $\boldsymbol{T}(y)$ is the sufficient statistic (for the majority of the distributions, it is often the case that $\boldsymbol{T}(y) = y$), $b(y)$ is the *underlying measure*, and $a(\boldsymbol{\eta})$ is the *log partition function*, which ensures that the distribution sums/integrates to one. Hence,

$$a(\boldsymbol{\eta}) = \log \sum\!\!\!\!\!\!\int b(y) \exp(\boldsymbol{\eta}^T \boldsymbol{T}(y))\, \mathrm{d}x.$$

Therefore, for a fixed choice of the functions $b(y)$, $\boldsymbol{T}(y)$, and $a(\boldsymbol{\eta})$, one may define a *family* (or set) of distributions. Furthermore, by varying the canonical parameter $\boldsymbol{\eta}$, one may obtain

different distributions within this family. It may be shown that many well-known distributions, such as the Gaussian, Bernoulli, multinomial, Poisson, gamma, beta, exponential and inverse Gaussian form part of the exponential family of distributions [162].

Using this formulation, one can show that the celebrated Gaussian (or normal) distribution is, in fact, a member of the exponential family of distributions. Recall that, during the derivation of a general linear model in §4.4.1, the value of σ^2 did not depend on the final choice of $\boldsymbol{\beta}$. Consequently, one may choose an arbitrary value for σ^2 without loss of generality. To simplify the subsequent derivation, σ^2 is set to unity (*i.e.* $\sigma^2 = 1$). In this way, the standard Gaussian distribution can be expanded to the form

$$\begin{aligned} p(y;\mu) &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(y-\mu)^2\right) \\ &= \underbrace{\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}y^2\right)}_{b(y)} \cdot \exp\left(\mu y - \underbrace{\frac{1}{2}\mu^2}_{a(\eta)}\right). \end{aligned}$$

Notice that this form of the Gaussian probability density expression is in the exponential family form of (4.11) with $\eta = \mu$, $T(y) = y$, $a(\eta) = \mu^2/2$ and $b(y) = (1/\sqrt{2\pi})\exp(-y^2/2)$.

As in the case of the Gaussian distribution, the well-known Bernoulli distribution can also be shown to reside within the set of exponential family distributions. The Bernoulli distribution specifies the distribution of a variable $y \in \{0, 1\}$ with mean ϕ as

$$\begin{aligned} p(y;\phi) &= \phi^y (1-\phi)^{1-y} \\ &= \exp(y \log\phi + (1-y)\log(1-\phi)) \\ &= \exp\left(\underbrace{\left(\log\left(\frac{\phi}{1-\phi}\right)\right)}_{\eta} y + \underbrace{\log(1-\phi)}_{-a(\eta)}\right). \end{aligned}$$

Again notice that this form of the Bernoulli probability mass expression is in the exponential family of the form (4.11) with $\eta = \log(\phi/(1-\phi))$, $T(y) = y$, $a(\eta) = -\log(1-\phi) = \log(1+e^\eta)$ and $b(y) = 1$. Interestingly, if the definition of η is inverted by solving for ϕ, the relation $\phi = 1/(1+e^{-\eta})$ is obtained. One may notice this as the familiar sigmoid function — a fact that will prove useful during the derivation of logistic regression as a GLM.

4.4.3 Constructing a generalised linear model

Consider a classification or regression problem in which the objective is to predict some target variable y as a function of the explanatory variables $\boldsymbol{x}$. In order to derive a GLM for this problem, McCullagh and Nelder [162] formally proposed the establishment of three model assumptions (or components):

1. A *random component*: Given a matrix of features $\boldsymbol{X}$ parameterised by $\boldsymbol{\beta}$, the target variables $\boldsymbol{y}$ are characterised by a probability distribution that belongs to the exponential family with canonical parameter $\boldsymbol{\eta}$, written as $\boldsymbol{y} \mid \boldsymbol{X}; \boldsymbol{\beta} \sim \text{ExpFamily}(\boldsymbol{\eta})$.

2. A *systematic component*: The linear combination of the input features $\boldsymbol{X}$ produces a linear predictor $\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta}$.

3. A *functional link*: A known monotonic differentiable *link function* $g(\cdot)$ relates the expected values $\boldsymbol{\mu} = E[\boldsymbol{T}(y) \mid \boldsymbol{X}; \boldsymbol{\beta}]$ of the target variable (random component) to the linear

predictor $\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta}$ of features (systematic component), such that the hypothesis of the GLM is given by $h(\tilde{\boldsymbol{x}}) = E[\boldsymbol{T}(y) \mid \tilde{\boldsymbol{x}}; \boldsymbol{\beta}] = g(\tilde{\boldsymbol{x}}\boldsymbol{\beta})^{-1}$.

In the case of OLSR, Assumption 1 was merely a Gaussian distribution and Assumption 3 was and identity link function, implying that $\boldsymbol{\mu} = \boldsymbol{\eta}$. GLMs, therefore, extend the assumption in (4.7) of general linear models to the relation

$$\boldsymbol{y} = g(\boldsymbol{X}\boldsymbol{\beta})^{-1} + \boldsymbol{\epsilon}. \tag{4.12}$$

The distribution of a set of independent target variables $\boldsymbol{y}$ may now be characterised by any distribution in the exponential family while the homogeneity of variance is not satisfied (it rather varies as a function of the covariate mean) [185]. Furthermore, the target variable and the explanatory variables no longer assume a linear relationship but are rather related *via* a link function.

4.4.4 Logistic regression

As in the case of OLSR, one may also show that logistic regression is simply a special case of the family of GLMs. Consider the case in which the set of target variables $\mathcal{Y} = \{y^{(1)}, \ldots, y^{(m)}\}$ are binary in nature (*i.e.* $y \in \{0, 1\}$). In this case, it seems natural to model the conditional distribution of $\mathcal{Y}$ given the set of observations $\mathcal{X}$ as a Bernoulli distribution. Consequently, in the case of logistic regression, the hypothesis $h(\boldsymbol{x})$ is given by

$$\begin{aligned} h(\boldsymbol{x}) &= E[y \mid \boldsymbol{x}; \boldsymbol{\beta}] && (4.13)\\ &= \phi && (4.14)\\ &= \frac{1}{1 + e^{-\eta}} && (4.15)\\ &= \frac{1}{1 + e^{-\boldsymbol{\beta}^T \boldsymbol{x}}}. && (4.16) \end{aligned}$$

Again, (4.13) is the functional link of Assumption 3, (4.14) follows from the fact that $y \mid \boldsymbol{x}; \boldsymbol{b} \sim \text{Bern}(\phi)$, (4.15) follows from Assumption 1 and the earlier derivation in §4.4.2 which showed that $\phi = 1/(1 + e^{-\eta})$ in the formulation of a Bernoulli distribution as an exponential family distribution. Finally, (4.16) follows from Assumption 3.

Notice that $h(\boldsymbol{x})$ is always bounded between zero and one since $h(\boldsymbol{x}) \to 1$ as $\boldsymbol{\beta}^T \boldsymbol{x} \to \infty$ and $h(\boldsymbol{x}) \to 0$ as $\boldsymbol{\beta}^T \boldsymbol{x} \to -\infty$. By convention, $p(y = 1 \mid \boldsymbol{x}; \boldsymbol{\beta})$ can be chosen as

$$p(y = 1 \mid \boldsymbol{x}; \boldsymbol{\beta}) = h(\boldsymbol{x}) \tag{4.17}$$

which means that

$$p(y = 0 \mid \boldsymbol{x}; \boldsymbol{\beta}) = 1 - h(\boldsymbol{x}). \tag{4.18}$$

Given that $p(y = 0 \mid \boldsymbol{x}; \boldsymbol{\beta}) + p(y = 1 \mid \boldsymbol{x}; \boldsymbol{\beta}) = 1$, (4.17) and (4.18) may be combined into the more compact Bernoulli representation

$$p(y \mid \boldsymbol{x}; \boldsymbol{\beta}) = h(\boldsymbol{x})^y (1 - h(\boldsymbol{x}))^{1-y}.$$

Since the m observations in $\mathcal{X}$ are assumed to be generated independently, the likelihood of the

parameters $\boldsymbol{\beta}$ may, therefore, be expressed as

$$\begin{aligned}\mathcal{L}(\boldsymbol{\beta}) &= p(\boldsymbol{y} \mid \boldsymbol{X}; \boldsymbol{\beta}) \\ &= \prod_{i=1}^{m} p(y^{(i)} \mid \boldsymbol{x}^{(i)}; \boldsymbol{\beta}) \\ &= \prod_{i=1}^{m} \left(h(\boldsymbol{x}^{(i)})\right)^{y^{(i)}} \left(1 - h(\boldsymbol{x}^{(i)})\right)^{1-y^{(i)}}, \end{aligned} \tag{4.19}$$

where, again, it is typically easier to maximise (4.19) in terms of its log likelihood

$$\begin{aligned}\ell(\boldsymbol{\beta}) &= \log \mathcal{L}(\beta) \\ &= \sum_{i=1}^{m} y^{(i)} \log h(\boldsymbol{x}^{(i)} + (1 - y^{(i)} \log(1 - h(\boldsymbol{x}^{(i)})). \end{aligned} \tag{4.20}$$

As previously mentioned, optimal values of $\boldsymbol{\beta}^*$ are typically realised by maximising (4.20) using the method of gradient decent [185].

Often, an additional *regularisation* term is added to (4.20) so as to penalise (4.20) for large choices of the value of β in an attempt to avoid overfitting [34]. Although the notion of regularisation may be applied in the case of many different learning models, this concept is described exclusively in the context of the logistic regression model derivation. The two most popular regularisation types are *L1 regularisation* (often termed *lasso regression*) and *L2 regularisation* (often termed *ridge regression*). In the case of the former, the logistic regression log likelihood in (4.20) can be redefined as

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{m} y^{(i)} \log h(\boldsymbol{x}^{(i)}) + (1 - y^{(i)}) \log(1 - h(\boldsymbol{x}^{(i)})) - \lambda \sum_{j=1}^{n} |\beta_j|$$

and, in the case of the latter as

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{m} y^{(i)} \log h(\boldsymbol{x}^{(i)}) + (1 - y^{(i)}) \log(1 - h(\boldsymbol{x}^{(i)})) - \lambda \sum_{j=1}^{n} \beta_j^2,$$

where λ is a parameter used to inflate the regularisation penalty. If $\lambda = 0$, no regularisation is applied while, if a large value of λ is chosen, the β-values will approach zero.

4.5 Decision tree learning

Decision tree learning is one of the most widely used supervised learning approaches — a fact that is primarily attributed to their capability of performing both classification and regression tasks, their relatively simple mathematical formulation and their ability to represent their underlying decision making process visually and explicitly [95]. A decision tree can, in essence, be explained in terms of two entities, namely decision *nodes* and *leaves*. Decision nodes describe the rule by which data are partitioned into smaller sub-parts, while leaves denote the final decisions or outcomes [129]. There exists a wide variety of tree-based learning algorithms in the literature, the most widely used being *classification and regression trees* (CART), *random forests* and the *C4.5* algorithm.

4.5.1 Classification and regression trees

CART is a popular non-parametric algorithm for producing either classification or regression trees, depending on the defined target variable [158]. The explanatory variables $\mathcal{X} = \{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(m)}\}$ can be either numerical or categorical [101]. The stages in a CART analysis are two-fold and employ a binary recursive partitioning method. The first stage is the tree growth stage where so-called *parent* nodes are recursively partitioned into two more homogeneous *child* nodes. The second stage is concerned with limiting the overall tree growth by computing the optimal tree size. These two steps are further elucidated in the remainder of this sub-section in the context of a classification tree since only this paradigm is applied later on in this dissertation.

The process of producing a classification tree begins at an initial node, called the *root node*, in which the entire training set is contained. Tree growth is achieved by partitioning each node in such a way as to maximise a defined splitting criterion. If the jth attribute is defined as the parent node d_p, then there exists an optimal splitting value x_j^* for this node. Once this value has been determined, an *if-then* splitting procedure may follow whereby all observations with a jth attribute value of x_j^* or greater are partitioned into a right-hand child node t_r, as illustrated in Figure 4.9, while all observations with a jth attribute value less than x_j^* are partitioned into a left-hand child node t_l [213]. The question, however, remains as to how to determine this "optimal" splitting value x_j^* for the jth attribute.

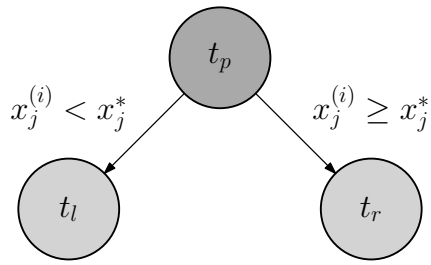


FIGURE 4.9: *Splitting a parent node into corresponding child nodes in a CART [270].*

According to Lewis [137], the primary goal of splitting a parent node into two child constituents is to achieve the largest improvement in predictive accuracy by maximising the *homogeneity* or *purity* of the two subsequent child nodes. The *impurity* of a node t is calculated by means of an *impurity function* $i(t)$ [254], which is defined in terms of the impurity of the parent node, denoted by $i(t_p)$, and the combined impurity of the associated child nodes, denoted by $i(t_c)$. Regardless of the child node split, the impurity of a parent node remains constant. The change in impurity, given a specific child split, may therefore be expressed as

$$\Delta i(t) = i(t_p) - E[i(t_c)], \tag{4.21}$$

where $E[i(t_c)]$ is the expected impurity of the two child nodes, which can also be expressed in terms of the probability of being split into the left- and right-hand nodes, denoted by $p(t_l)i(t_l)$ and $p(t_r)i(t_r)$, respectively. Consequently, in the case of a classification tree, all possible attribute values for all n attributes in $\mathcal{X}$ must be considered as possible "best split," resulting in the maximum value of (4.21), expressed mathematically at a single parent node, as

$$\underset{\substack{x_j < x_j^* \\ j \in \{1,\ldots,m\}}}{\arg\max} \; i(t_p) - p(t_l)i(t_l) - p(t_r)i(t_r). \tag{4.22}$$

Although this process of maximising the change in impurity described in (4.22) is common to both classification and regression trees, the choice of impurity function $i(t)$ is defined in a

different manner [174]. Many different impurity functions have been proposed in the literature. In the case of classification trees, however, the most widely used is the *Gini splitting rule* and the *Twoing splitting rule* [270].

A tree that has been grown to the point where it fits the training data almost perfectly typically results in an overly complex model that performs poorly in respect of the validation data, as illustrated in Figure 4.10. A tree that is too small, on the other hand, is not able to discover the underlying relationships present in the data. Consequently, the ideal tree size is one that achieves a suitable trade-off between overfitting and underfitting the training data. A CART is, therefore, grown to a near optimal size by employing one of two methods: The growth of the tree can be terminated according to (1) a *stopping criterion* or (2) employing a *pruning* procedure.

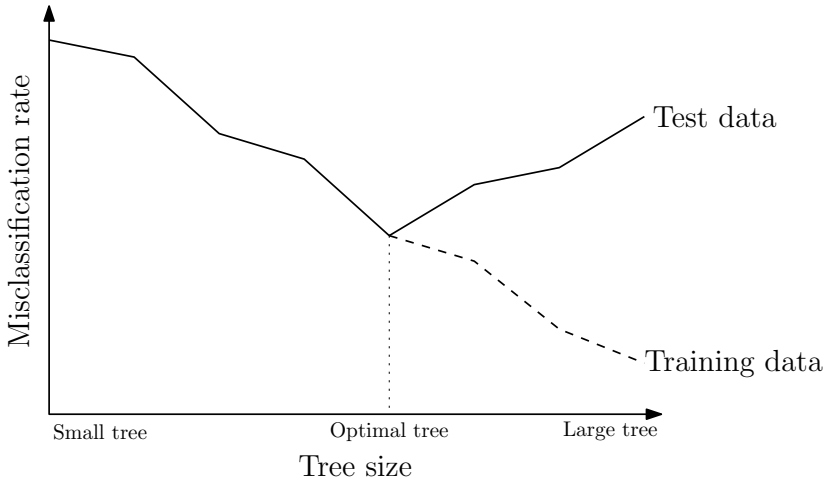


FIGURE 4.10: *The change in misclassification rate according to a tree's size [256].*

As the name implies, the use of the stopping criterion approach forces the recursive tree growing procedure to be terminated once a specified criterion is met. Rashidi *et al.* [213] described two such criteria. The first of these is the *maximum tree depth* criterion which terminates tree growth after a pre-specified number of branching levels have been created. The second is the *minimum node size to split* criterion which terminates tree growth at a specific node if the size of the node (*i.e.* the number of observations contained in the node) is less than a pre-specified value.

Although the stopping criterion provides a simplified method for limiting the growth of a tree, the more popular method is that of pruning [270]. According to this procedure, a tree is first grown to its maximum size and then pruned backwards to a more suitable size [174]. There exists a wide variety of methods for determining the optimal size to which a tree should be pruned backwards, including *reduced-error pruning*, *pessimistic pruning*, *rule post-pruning*, and *cost complexity pruning* [114, 222]. Performance comparisons of the various pruning methods have been conducted in multiple studies, with the majority finding that no single pruning method is superior under all circumstances [75, 172, 212] — a finding that echoes the *No Free Lunch* theorem[3].

Although CART is a powerful predictive model in its original form, there are three attractive methods for enhancing this standard learning model [174]. The first of these relies on a procedure called *bagging* which produces an aggregation of multiple bootstrap samples (hence, bagging is also referred to as *bootstrap aggregating*), as described in §4.2.3. Although the notion of bagging may be applied to any learning model in an attempt to improve its stability and accuracy [114],

[3]The No Free Lunch theorem states that "any two optimization algorithms are equivalent when their performance is averaged across all possible problems" [291].

this procedure is described here exclusively in the context of the CART learning model so as to demonstrate its working. Given a training set $\mathcal{S}$, a total of b separate data subsets $\mathcal{S}_1, \ldots, \mathcal{S}_b$ are randomly drawn from $\mathcal{S}$ and used to construct b CART models. The predictive capability of CART model i is then validated in respect of the data not used to construct the model (*i.e.* the set $\mathcal{S} \backslash \mathcal{S}_i$). In the case of a classification task, the final categorisation is given by the majority vote of the b CART models (for this reason, b should be chosen as odd in the case of a binary classification problem to avoid ties) while, in the case of a regression task, the results of the independent CART predictions are averaged [114]. The predictions achieved by this bagging approach are reported to exhibit less variance than that of a single CART model [104].

The second CART enhancement method is that of *boosting*. Much like bagging, the notion of boosting may be applied to any learning model [114], but, for demonstration purposes, is described here exclusively in the context of the CART learning model. Boosting involves the construction of b CART models *sequentially* from an adaptive training data subset, where each new tree utilises the information discovered by its predecessors. This stands in contrast to bagging, which constructs multiple CART models *independently* from bootstrapped data samples. Initially, all observations are assigned an equal weighting and are randomly sampled according to this weighting to be included in the training sample set $\mathcal{S}$. During the construction and evaluation of each new CART model, the frequently misclassified observations (known as *problematic observations*) are given more weight, resulting in their more frequent inclusion in the new sample $\mathcal{S}'$ for training future CART models. This effectively allows future models to correct the mistakes of their predecessors, thus improving the models' overall prediction ability [174]. It should be noted, however, that boosting may suffer from overfitting if the value of b is chosen too large [114].

The third and, by far, most popular CART enhancement is called *random forests* which is discussed, in its own right, in the next section.

4.5.2 Random forests

Introduced by Breiman [25] and often considered to be a panacea of all data science problems [4], random forests is a versatile and simplistic learning model that has proven effective in a wide range of classification and regression tasks [59]. Other advantages of the method of random forests include its ability to handle categorical and numerical attributes without any need for scaling, its inclination to perform implicit feature selection, and its parallel algorithmic structure which allows for relatively fast training times [67].

Similar to bagging, this method involves the construction of b decision trees produced by CART and combines them together to produce a more accurate, stable prediction. In addition to the randomness induced by employing bootstrap sampling, random forests provide further randomness by imposing the following change to the standard CART model: Starting at the root node, instead of considering each of the m attributes as possible candidates according to which to split the node, only a random subset of $m' < m$ (typically chosen as $m' \approx \sqrt{m}$) attributes are considered. This process of only considering a random attribute subset to split a node is continued until the tree reaches the largest possible size. The process is repeated a total of b times, and the final prediction is typically a majority vote (in the case of classification) or an average (in the case of regression) of all the predictions made by the b trees constructed [174].

The reason why this approach works so well can be described as follows: By only allowing the algorithm to consider a subset of the available attributes at a node split, the algorithm is charged to explain the target variables by not only utilising the "best" attributes present in

the data. In other words, if there is an attribute which is an inherently very strong predictor of the target variables, this attribute will not be available to the algorithm $100(m - m')/m\%$ of the time (on average), thereby forcing the algorithm to explore the use of other attributes which are moderately strong predictors [114]. In the case of bagging, the collection of trees are, to a large extent, similar to one another since they all typically rely on the same strong predictor when splitting the root node. Hence, the members of this collection of trees are highly correlated and an average of highly correlated results typically does not produce a large reduction in variance. For this reason, the method of random forests is typically described as a technique for *decorrelating* the trees produced by CART [114].

4.5.3 The C4.5 algorithm

The C4.5 algorithm was developed by Quinlan [210] and is the successor of his earlier ID3 algorithm [211]. Similar to CART, this algorithm begins with the training set defined at a single root node of a tree which is subsequently partitioned into child nodes based on a specific splitting criterion. The algorithm is, however, only applicable in the case of classification tasks. In the case of the C4.5 algorithm, the *information gain ratio* is used as the default spitting criterion. Just like in CART, if a node is split based on a quantitative attribute, a threshold may be determined as the point at which the training data set is partitioned into two subsequent child nodes. In the case of a qualitative attribute, however, the C4.5 algorithm creates as many child nodes as there are associated categories. For this reason, the C4.5 algorithm is described as a *multi-way split* (or *non-binary split*) tree growing algorithm [126]. The tree growing procedure of the C4.5 algorithm also follows a greedy approach by attempting to find the best local node split which maximises the information gain ratio without employing any form of backtracking [84, 228].

The split criterion employed by the C4.5 algorithm is based on Shannon's well-known concept of entropy from information theory [243]. Entropy is a measure of the disorder in a data subset. Denote the set of C_j categories for attribute j by $\mathcal{C}_j = \{c_1^{(j)}, \ldots, c_{C_j}^{(j)}\}$, and the frequency of observations in a particular subset of data $\mathcal{S} \subset \mathcal{X}$ with a target variable of q by $\text{freq}(q, \mathcal{S})$. Then the entropy of a category $c_k^{(j)}$ of attribute j is given by

$$\text{Entropy}(\mathcal{S}) = -\sum_{q=1}^{Q} \frac{\text{freq}(q, \mathcal{S})}{|\mathcal{S}|} \log_b \left(\frac{\text{freq}(q, \mathcal{S})}{|\mathcal{S}|} \right), \tag{4.23}$$

where Q is the total number of target attribute categories. If a large proportion of observations in a data subset $\mathcal{S}$ have the value of q (*i.e.* $\text{freq}(q, \mathcal{S})/|\mathcal{S}| \approx 1$), then the entropy in (4.23) will be approximately zero (implying a high level of order). Conversely, if there is an equal representation of all attribute categories in the data subset $\mathcal{S}$ (*i.e.* $\text{freq}(q, \mathcal{S})/|\mathcal{S}| \approx 1/C_j$), then the entropy in (4.23) will be close to one (implying a high level of disorder).

With this notion in mind, consider the case in which a data subset $\mathcal{S}$, residing in a parent node t_p, is partitioned based on the categories of attribute j producing C_j child nodes $t_1, \ldots, t_{C_j}$. The *information gain* achieved by this node split can be defined as the total reduction of entropy in the data, expressed mathematically as

$$\text{Gain}(t_p, j) = \text{Entropy}(t_P) - \sum_{i=1}^{C_j} \frac{|t_i|}{|t_p|} \text{Entropy}(t_i). \tag{4.24}$$

This criterion was first proposed for splitting nodes in the ID3 algorithm [211]. A notable drawback associated with this splitting criterion is that it is biased towards attributes with

many categories [264]. This led Quinlan [210] subsequently to build on this notion in order to create the information gain ratio, which is a more robust criterion than information gain and generally produces better classification results. This formulation utilises (4.23) as a basis to define a so-called *split entropy* in terms of the categories of an attribute j (as opposed to the target attribute categories), expressed mathematically as

$$\text{SplitEntropy}(t_p, j) = -\sum_{i=1}^{C_j} \frac{|t_i|}{|t_p|} \log_b \left(\frac{|t_i|}{|t_p|} \right). \tag{4.25}$$

If a large proportion of observations in a data subset $\mathcal{S}$ exhibit the value $c_k^{(j)}$ for attribute j (*i.e.* $|t_i|/|t_p| \approx 1$), then the entropy in (4.23) will be approximately zero, while, conversely, if there is an equal representation of all attribute categories in the data subset $\mathcal{S}$ (*i.e.* $|t_i|/|t_p| \approx 1/C_j$), then the entropy in (4.23) will be close to one. Combining (4.24) and (4.25), it is possible to define the information gain ratio as the proportion of information generated by partitioning a parent node t_p according to attribute j, expressed mathematically as

$$\text{GainRatio}(t_p, j) = \frac{\text{Gain}(t_p, j)}{\text{SplitEntropy}(t_p, j)}. \tag{4.26}$$

Unlike its predecessor, this formulation is more robust since it effectively penalises the choice of splitting attributes with many categories. A drawback of this approach, however, is that (4.26) tends to favour unbalanced splits in which one node partition is much larger than the rest [103].

4.6 The naive Bayes classifier

The naive Bayes classifier is a fast, highly scalable probabilistic classifier based on Bayes' well-known theorem from probability theory with an added independence assumption among all observations. In spite of this apparently over-simplified assumption, the learning model performs surprisingly well in a wide variety of classification tasks, most notably in text categorisation [214]. In addition, the naive Bayes model is relatively simple to construct and requires no complicated iterative parameter estimations, making it particularly useful when confronted with large data sets [233].

Given a set of explanatory variables $\mathcal{X}$ corresponding to a set of target variables $\mathcal{Y}$, models that attempt to learn the mapping directly from $\mathcal{X}$ to $\mathcal{Y}$ or the joint probability $p(y \mid \boldsymbol{x})$ (like in the case of logistic regression) are called *discriminative learning models*. The naive Bayes classifier, on the other hand, is actually part of a much broader class of machine learning algorithms called *generative learning models*, which, in contrast to discriminative learning models, attempt to model the joint probability $p(\boldsymbol{x} \mid y)$, together with the class prior $p(y)$ [185]. The strong assumption made by the naive Bayes algorithm is that the m observations in the set $\mathcal{X} = \{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(m)}\}$ are conditionally independent of one other, given the associated set of target variables $\mathcal{Y}$. In the remainder of this section, the derivation of the naive Bayes learning model is described in the context of a binary classification problem in which all attributes are dichotomous. In addition, the indicator function notation $1\{\bullet\}$ is introduced to aid in this derivation, where $\bullet$ denotes any specified condition. This function takes the value 1 if the condition $\bullet$ is true, or zero otherwise (*e.g.* $1\{3 = 5 - 2\} = 1\{\text{True}\} = 1$, while $1\{2 = 3\} = 1\{\text{False}\} = 0$).

First and foremost, it is necessary formally to state the parameters of the naive Bayes model which are estimates of the probabilities of each of the possible classification outcomes, denoted in terms of a function ϕ as follows: The fraction of observations with a target variable of $y = 1$

and an attribute value of $j = 1$ is given by $\phi_{j|y=1} = p(x_j = 1, y = 1)$, the fraction of observations with a target variable of $y = 0$ and an attribute value $j = 0$ is given by $\phi_{j|y=0} = p(x_j = 1, y = 0)$, and lastly $\phi_y = p(y = 1)$ denotes the total proportion of observations with a target variable of $y = 1$. According to Ng [185], the joint likelihood of the data may, therefore, be defined mathematically as

$$\mathcal{L}(\phi_y, \phi_{j|y=0}, \phi_{j|y=1}) = \prod_{i=1}^{m} p(\boldsymbol{x}^{(i)}, y^{(i)}). \tag{4.27}$$

Maximising (4.27) with respect to $\phi_{j|y=1}$, $\phi_{j|y=0}$ and $\phi_{j|y=0}$ results in the MLE

$$\phi_{j|y=1} = p(x_j = 1, y = 1) = \frac{\sum_{i=1}^{m} 1\{x_j^{(i)} = 1 \cap y^{(i)} = 1\}}{\sum_{i=1}^{m} 1\{y^{(i)} = 1\}},$$

$$\phi_{j|y=0} = p(x_j = 1, y = 0) = \frac{\sum_{i=1}^{m} 1\{x_j^{(i)} = 1 \cap y^{(i)} = 0\}}{\sum_{i=1}^{m} 1\{y^{(i)} = 0\}},$$

and

$$\phi_y = p(y = 1) = \frac{\sum_{i=1}^{m} 1\{y^{(i)} = 1\}}{m}.$$

After fitting the above parameters to the model, the class of a new vector of explanatory variables $\boldsymbol{x}$ may be predicted according to Bayes' theorem [124], as

$$\begin{aligned} p(y = 1 \mid \boldsymbol{x}) &= \frac{p(\boldsymbol{x} \mid y = 1)p(y = 1)}{p(\boldsymbol{x})} \\ &= \frac{(\prod_{i=j}^{n} p(x_j = 1 \mid y = 1))p(y = 1)}{(\prod_{i=j}^{n} p(x_j = 1 \mid y = 1))p(y = 1) + (\prod_{i=j}^{n} p(x_j = 1 \mid y = 0))p(y = 0)} \end{aligned}$$

and selecting the class with the highest posterior probability [185]. The generalisation of this formulation for the case when attributes are multichotomous can be achieved by modelling the conditional probability $p(x_i \mid y)$ as a multinomial distribution (as opposed to a Bernoulli distribution). Furthermore, in the case when the attribute values are continuous, these values are typically discretised into a finite number of bins [185].

4.7 The k-nearest neighbours algorithm

The kNN algorithm is a simple non-parametric classifier that makes no assumption about the relationship between the explanatory variables $\mathcal{X}$ and target variables $\mathcal{Y}$. This algorithm involves computing the distances between a new vector of explanatory variables $\boldsymbol{x}$ to be classified and each of the observations in the training set. The hypothesis $h(\boldsymbol{x})$ of the unclassified explanatory variables $\boldsymbol{x}$ is given by the most common category of the k closest observations to $\boldsymbol{x}$ [114]. Consequently, an odd value of k is typically chosen in the case of a binary classification problem so as to avoid ties. In the kNN algorithm, the distance metric between the independent observations can also be specified, the most popular being the *Euclidean*, *Manhattan* and *Chebyshev* distance metrics [205].

Although the kNN algorithm provides a simple intuitive classification procedure, this algorithm has many associated disadvantages. First, the kNN algorithm is not suitable in the case of categorical attributes (especially nominal attributes) since the distance between categories typically does not make sense. In this case, an appropriate procedure for converting categorical attributes into new attributes, defined on a numerical scale, must be defined. This is typically achieved by

means of a technique called *one-hot encoding* which converts each attribute category into a new binary attribute [281].

In addition, the kNN algorithm is also computationally expensive if the distance to all m observations must computed in order to classify a new observation. A more computationally efficient method can, however, be achieved by pre-computing a Voronoi tessellation of the feature space [134, 283]. A graphical depiction of such a Euclidean distance-based tessellation is shown in Figure 4.11 for a two-dimensional binary classification example with $k = 1$, where blue and red points indicate observations from the respective target categories. As may be seen from this figure, a *Voronoi cell* encapsulates, for each observation $i \in \{1, \ldots, m\}$, the set of points in $\mathbb{R}^2$ for which observation i is the closest, described mathematically as $V_i = \{x \in \mathbb{R}^2 : d(x, x_i) \leq d(\boldsymbol{x}, \boldsymbol{x}_j), j = \{1, \ldots, i-1, i+1, \ldots, m\}\}$, where $d(\boldsymbol{x}, \boldsymbol{x}_i)$ is the distance from an arbitrary point $\boldsymbol{x}$ to the point $\boldsymbol{x}_i$. Consequently, in the case of Figure 4.11, any observation that falls within a red or blue Voronoi cell is classified as being part of the red or blue target category, respectively. A mathematical derivation of the Voronoi tessellation is excluded from this literature review, but may be found in the exposition by Lee [134].

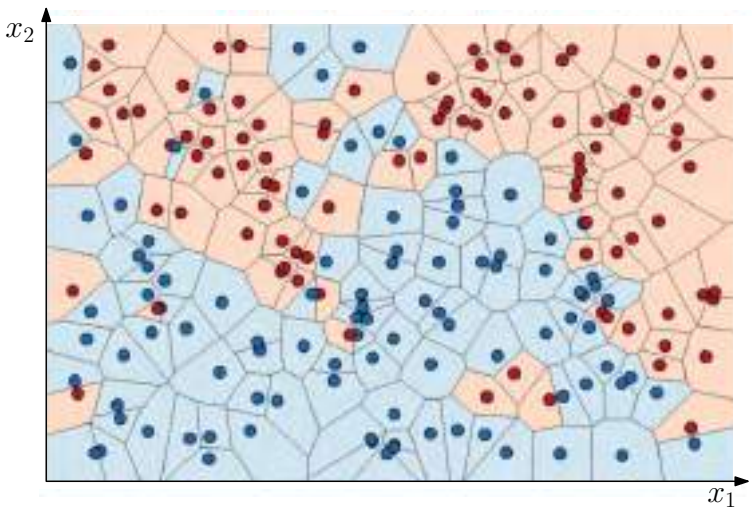


FIGURE 4.11: *A two-dimensional Euclidean distance-based Voronoi tessellation showing Voronoi cells produced by red and blue observations [205].*

4.8 Support vector machines

SVMs were conceived in 1992 by Boser *et al.* [22] as a way of creating non-linear classifiers using the so-called *kernel trick* [96]. SVMs are believed to be among the best "off-the-shelf" learning models, being used widely in industry and academia for both classification and regression tasks [185]. For the purposes of this review, however, SVMs are again described exclusively in the context of a binary classification problem.

In its most simple form, the objective of SVMs is to determine a single decision surface, called a *hyperplane*, which produces the widest margin of separation possible between the observations of two different classes. The set of observations with the shortest perpendicular distance to the hyperplane in each of the two target categories are called the *support vectors*. The category of a new vector of explanatory variables $\boldsymbol{x}$ can, therefore, be predicted by determining on which side of the hyperplane the vector lies [38]. In reality, however, this is rarely achievable since real-life observations are almost never *linearly separable* (*i.e.* the data cannot be separated by

a single linear decision hypersurface) [54]. Consequently, an additional procedure is performed whereby observations are transformed into a higher-dimensional space, using *kernels*, in such a manner that a hyperplane can, in fact, be placed appropriately. The case in which observations are linearly separable is first described.

4.8.1 Linear separability

As a way of graphically describing the working of SVMs, consider the set of two-dimensional observations illustrated in Figure 4.12(a). In this figure, the observations belonging to Class 1 are denoted by black circles while those belonging to Class 2 are denoted by grey squares. As previously described, the aim of SVMs is to determine the widest margin of separation M between the observations of the two different classes. In this example, the observations that lie on the dashed boundary lines B_1 and B_2 denote the support vectors. Notice in Figure 4.12(b) that one may attempt to construct another hyperplane with a margin of separation denoted by M'. This hyperplane, however, achieves a smaller margin of separation than the initial hyperplane H with a margin of separation M, and is thus less favourable.

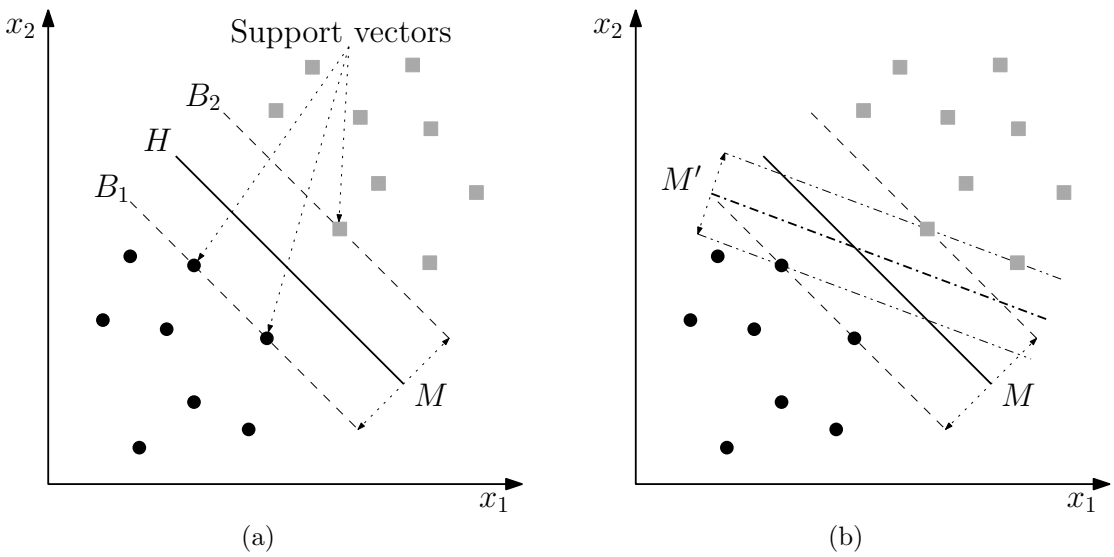


FIGURE 4.12: *Support vectors and the corresponding induced hyperplanes of varying margins that can be created in respect of a training set, depending on the choice of support vectors [256].*

With this notion in mind, the general working of the SVMs learning model is now described mathematically for the case when observations are linearly separable. According to Çaydaş and Ekici [38], the position of the separating hyperplane $H(\boldsymbol{x})$ may be defined by a perpendicular weight vector $\boldsymbol{w}$ and the orthogonal distance b from the origin to the hyperplane, as shown in Figure 4.13. Since $\boldsymbol{w}$ is, by definition, orthogonal to the hyperplane, it follows that an arbitrary point $\boldsymbol{x}_H$ on the hyperplane must satisfy

$$H(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x}_H + b = 0.$$

In the same way, the boundary lines $B_1(\boldsymbol{x})$ and $B_2(\boldsymbol{x})$ parallel to $H(\boldsymbol{x})$ will satisfy

$$B_1(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x}_{B_1} + b = -1 \tag{4.28}$$

and

$$B_2(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x}_{B_2} + b = +1 \tag{4.29}$$

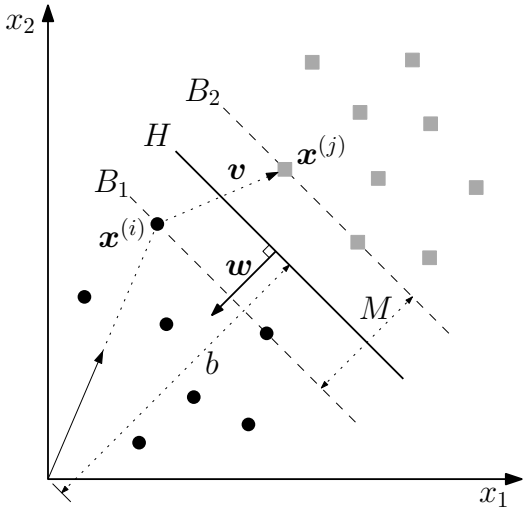


FIGURE 4.13: *Vectors of the SVMs learning model [256].*

for any points $\boldsymbol{x}_{B_1}$ and $\boldsymbol{x}_{B_2}$ lying on the respective boundaries.

Let $\mathcal{X} = \{\boldsymbol{x}^{(1)}, \cdots, \boldsymbol{x}^{(m)}\}$ denote the set of observations which belong to either Class 1 or Class 2. It is customary to assign the class values $y^{(i)} = -1$ and $y^{(i)} = +1$ (as opposed to 0 and +1) to each observation belonging to Class 1 and Class 2, respectively. Given the definitions (4.28) and (4.29) of the boundary planes, any observation i must, therefore, satisfy the constraints

$$\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b \leq -1, \quad \text{if } y^{(i)} = -1 \tag{4.30}$$

and

$$\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b \geq +1, \quad \text{if } y^{(i)} = +1 \tag{4.31}$$

for all $i \in \{1, \ldots, m\}$. With the inclusion of the class labels, (4.30) and (4.31) may be combined into the single constraint

$$y^{(i)}(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b) - 1 \geq 0 \tag{4.32}$$

[286], where a support vector $\boldsymbol{x}^{(i)}$ must, therefore, satisfy

$$y^{(i)}(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b) - 1 = 0. \tag{4.33}$$

Once the support vectors are known, the margin of separation M can be determined as follows. Consider two individual observations $\boldsymbol{x}^{(i)}$ and $\boldsymbol{x}^{(j)}$ lying on the boundary lines $B_1(\boldsymbol{x})$ and $B_2(\boldsymbol{x})$, respectively, as illustrated in Figure 4.13, where the difference between these two observations is denoted by $\boldsymbol{v} = \boldsymbol{x}^{(j)} - \boldsymbol{x}^{(i)}$. The value of M may, thus, be determined by the magnitude of $\boldsymbol{v}$ if it is transformed in such a way that it is orthogonal to $H(\boldsymbol{x})$ (and parallel to $\boldsymbol{w}$). This may be achieved by multiplying $\boldsymbol{v}$ by a unit vector with the same direction as that of $\boldsymbol{w}$, expressed mathematically as

$$M = ((\boldsymbol{x}^{(j)})^T\boldsymbol{x}^{(i)}) \cdot \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|} = \frac{\boldsymbol{w}^T\boldsymbol{x}^{(j)}}{\|\boldsymbol{w}\|} - \frac{\boldsymbol{w}^T\boldsymbol{x}^{(i)}}{\|\boldsymbol{w}\|}, \tag{4.34}$$

where $\|\boldsymbol{w}\|$ denotes the magnitude of $\boldsymbol{w}$.

With this formulation in mind, the focus shifts to deriving an optimal weight vector $\boldsymbol{w}$ which maximises the margin of separation M. Using the fact that observation $\boldsymbol{x}^{(i)}$, lying on the boundary line $B_1(\boldsymbol{x})$, has a class value of $y^{(i)} = -1$ and, similarly, that an observation $\boldsymbol{x}^{(j)}$, lying on the boundary line $B_2(\boldsymbol{x})$, has a class value of $y^{(j)} = +1$, it is possible to rewrite (4.33) as

$$-(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b) - 1 = 0$$

and

$$+(\boldsymbol{w}^T\boldsymbol{x}^{(j)} + b) - 1 = 0,$$

which, when written in terms of $\boldsymbol{w}^T\boldsymbol{x}^{(i)}$ and $\boldsymbol{w}^T\boldsymbol{x}^{(j)}$, become

$$\boldsymbol{w}^T\boldsymbol{x}^{(j)} = -1 - b \tag{4.35}$$

and

$$\boldsymbol{w}^T\boldsymbol{x}^{(i)} = 1 - b, \tag{4.36}$$

respectively. Substituting (4.35) and (4.36) into (4.34) yields

$$M = \frac{1-b}{\|\boldsymbol{w}\|} - \frac{-1-b}{\|\boldsymbol{w}\|} = \frac{2}{\|\boldsymbol{w}\|}. \tag{4.37}$$

The objective is to find the values in the optimal weight vector, denoted by $\boldsymbol{w}^*$, which maximises (4.37). By recognising that maximising $1/\|\boldsymbol{w}\|$ is equivalent to minimising $\|\boldsymbol{w}\|^2$, this formulation is often expressed in terms of a *quadratic programming problem* in which the objective is to

$$\underset{\boldsymbol{w}}{\text{minimise}} \quad \frac{1}{2}\|\boldsymbol{w}\|^2, \tag{4.38}$$

subject to the constraints

$$y^{(i)}(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b) - 1 \geq 0, \;\; i \in \{1, \ldots, m\} \tag{4.39}$$

[286]. This formulation, called the *primal problem*, essentially maximises the separation margin M while maintaining the assertion that all observations of different classes fall on opposite sides of the hyperplane $H(\boldsymbol{x})$. One may, however, go one step further and express the above quadratic programming problem as a *dual optimisation problem*, by writing (4.38) and (4.39) in terms of the Lagrangian

$$\mathcal{L}(\boldsymbol{w}, b, \alpha) = \frac{1}{2}\|\boldsymbol{w}\|^2 - \sum_{i=1}^{m} \alpha_i (y^{(i)}(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b) - 1) \tag{4.40}$$

and maximising with respect to the *Lagrange multipliers* $\alpha_1, \ldots, \alpha_m$ [301]. This can be achieved by first determining the partial derivatives of (4.40) with respect to $\boldsymbol{w}$ and b, and setting each of these equal to zero (*i.e.* $\partial\mathcal{L}(\boldsymbol{w}, b, \alpha)/\partial\boldsymbol{w} = 0$ and $\partial\mathcal{L}(\boldsymbol{w}, b, \alpha)/\partial b = 0$) in order to obtain the requirements

$$\boldsymbol{w} = \sum_{i=1}^{m} \alpha_i y^{(i)} \boldsymbol{x}^{(i)} \tag{4.41}$$

and

$$\sum_{i=1}^{m} \alpha_i y^{(i)} = 0 \tag{4.42}$$

[156]. By substituting (4.41) and (4.42) into (4.40), the dual objective can be expressed as the function

$$\sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y^{(i)} y^{(j)} ((\boldsymbol{x}^{(i)})^T \boldsymbol{x}^{(j)}) \tag{4.43}$$

of α only, which must be maximised subject to the constraints

$$\sum_{i=1}^{m} \alpha_i y^{(i)} = 0, \tag{4.44}$$

$$\alpha_i \geq 0, \;\; i \in \{1, \ldots, m\} \tag{4.45}$$

[258, 286, 301]. This new formulation will prove useful when applying the use of kernels. Consider again the arbitrary points $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ lying on the boundary lines $B_1(\boldsymbol{x})$ and $B_2(\boldsymbol{x})$, respectively. At this point, it is also possible to derive the optimal value of b, denoted by b^*, in terms of the Lagrange multipliers as

$$b^* = -\frac{1}{2}\left(\max_{i,y^{(i)}=-1} \boldsymbol{w}^{*T}\boldsymbol{x}^{(i)} + \min_{i,y^{(i)}=1} \boldsymbol{w}^{*T}\boldsymbol{x}^{(i)}\right) \tag{4.46}$$

[185]. The final hypothesis $h(\boldsymbol{x})$ of SVMs is, therefore, given by

$$h(\boldsymbol{x}) = \sum_{i=1}^{m} \alpha_i y^{(i)} (\boldsymbol{x}^{(i)})^T \boldsymbol{x}^{(j)} + b^*, \tag{4.47}$$

where a new observation $\boldsymbol{x}$ may be classified, according to Figure 4.13, as

$$\boldsymbol{x} = \begin{cases} \text{Class 2 (a square observation)}, & \text{if } h(\boldsymbol{x}) \geq 0, \\ \text{Class 1 (a circle observation)}, & \text{otherwise.} \end{cases}$$

4.8.2 Linear inseparability

The above formulation of SVMs is well suited in the case where two different classes of observations can easily be separated by a linear hyperplane. In reality, however, it is rarely the case that *all* observations from different classes can be separated in this fashion. In Figure 4.14, for example, there is no two-dimensional linear hyperplane that can separate the black circle observations from the grey square observations.

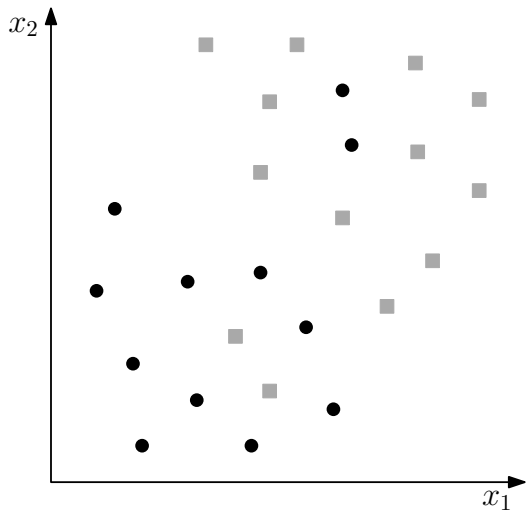


FIGURE 4.14: *A set of observations that cannot be separated by a linear hyperplane [256].*

In such a case, if one were to still to attempt to construct a two-dimensional linear hyperplane, it is necessary to relax the constraint (4.32). This is achieved by introducing additional variables $\xi_i \geq 0$, called *slack variables*, so that (4.32) may be redefined as

$$y^{(i)}(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b) \geq 1 - \xi_i, \tag{4.48}$$

for all $i \in \{1, \ldots, m\}$. This formulation ensures that a point $\boldsymbol{x}^{(i)}$ can always satisfy (4.48) regardless of on which side of the hyperplane it lies. It is, thus, necessary to restrict the extent

to which a point may lie on the 'wrong' side of the hyperplane by penalising the sum of the individual ξ_i-values by a so-called error penalty, denoted by ζ. In this way, the original primal problem (4.38)–(4.39) may be redefined as

$$\underset{\boldsymbol{w}}{\text{minimise}} \quad \frac{1}{2}\|\boldsymbol{w}\|^2 + \zeta \sum_{i=1}^{m} \xi_i, \tag{4.49}$$

subject to the constraints

$$y^{(i)}(\boldsymbol{w}^T \boldsymbol{x}^{(i)} + b) \geq 1 - \xi_i, \quad i \in \{1, \ldots, m\}, \tag{4.50}$$

$$\xi_i \geq 0, \qquad i \in \{1, \ldots, m\}. \tag{4.51}$$

As previously described, the primal problem (4.49)–(4.51) can be converted to a dual problem by introducing Lagrange multipliers, yielding

$$\underset{\alpha}{\text{maximise}} \quad \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y^{(i)} y^{(j)} ((\boldsymbol{x}^{(i)})^T \boldsymbol{x}^{(j)}) \tag{4.52}$$

subject to the constraints

$$\sum_{i=1}^{m} \alpha_i y^{(i)} = 0, \tag{4.53}$$

$$0 \leq \alpha_i \leq C, \quad i \in \{1, \ldots, m\}, \tag{4.54}$$

in which case the final hypothesis is still specified by (4.47). Notice that, in this formulation, the slack variables $\xi_1, \ldots, \xi_m$ are no longer present. Consequently, (4.52)–(4.54) is, for the most part, identical to (4.43)–(4.45), except in the case of (4.54), where ζ acts as a new regularisation term specifying an upper bound on the Lagrange multipliers $\alpha_1, \ldots, \alpha_m$.

It can also be shown that $\xi = 0$ and $0 < \alpha_i < \zeta$ for all support vectors [301]. For all non-support vectors, on the other hand, $\alpha_i = 0$, which indicates that such observations can safely be ignored since they have no effect on the final solution — a fundamental fact that is partly the reason for SVMs being largely computationally efficient [38]. An increase in the value of ζ places more weight on the slack variables $\xi_1, \ldots, \xi_m$, effectively enforcing a stricter separation between classes [160]. Equivalently, by decreasing the value of ζ towards zero, the misclassification rate is deemed less important and so more observations are incorrectly classified. Adjusting the parameter ζ thus assists in striking a balance between overfitting and underfitting the training data.

4.8.3 The application of kernels

As previously alluded to, the dual problem formulation (4.52)–(4.54) allows for the use of the so-called *kernel method* (in addition to the incorporation of slack variables) [158, 301]. This method (commonly known as the *kernel trick* [267]) involves transforming the training data into a higher-dimensional space by means of a *kernel function*, denoted by $k(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)})$. To provide an initial intuition for this trick, consider the data distribution shown in Figure 4.15(a) in which observations of Class 1 (indicated by blue triangles) are surrounded by observations of Class 2 (indicated by red circles). It is clear that, in this form, no two-dimensional linear hyperplane can meaningfully convey an effective separation of these two classes. If, however, the data are transformed according to a polynomial function by making use of a third dimension, as shown

in Figure 4.15(b), then this does allow for the insertion of a linear hyperplane, as illustrated in Figure 4.15(c), which can appropriately separate the two different classes of observations. Translating this result back to a two-dimensional plot effectively creates a non-linear (in this case, circular) hypersurface for the separation of the two classes, as shown in Figure 4.15(d).

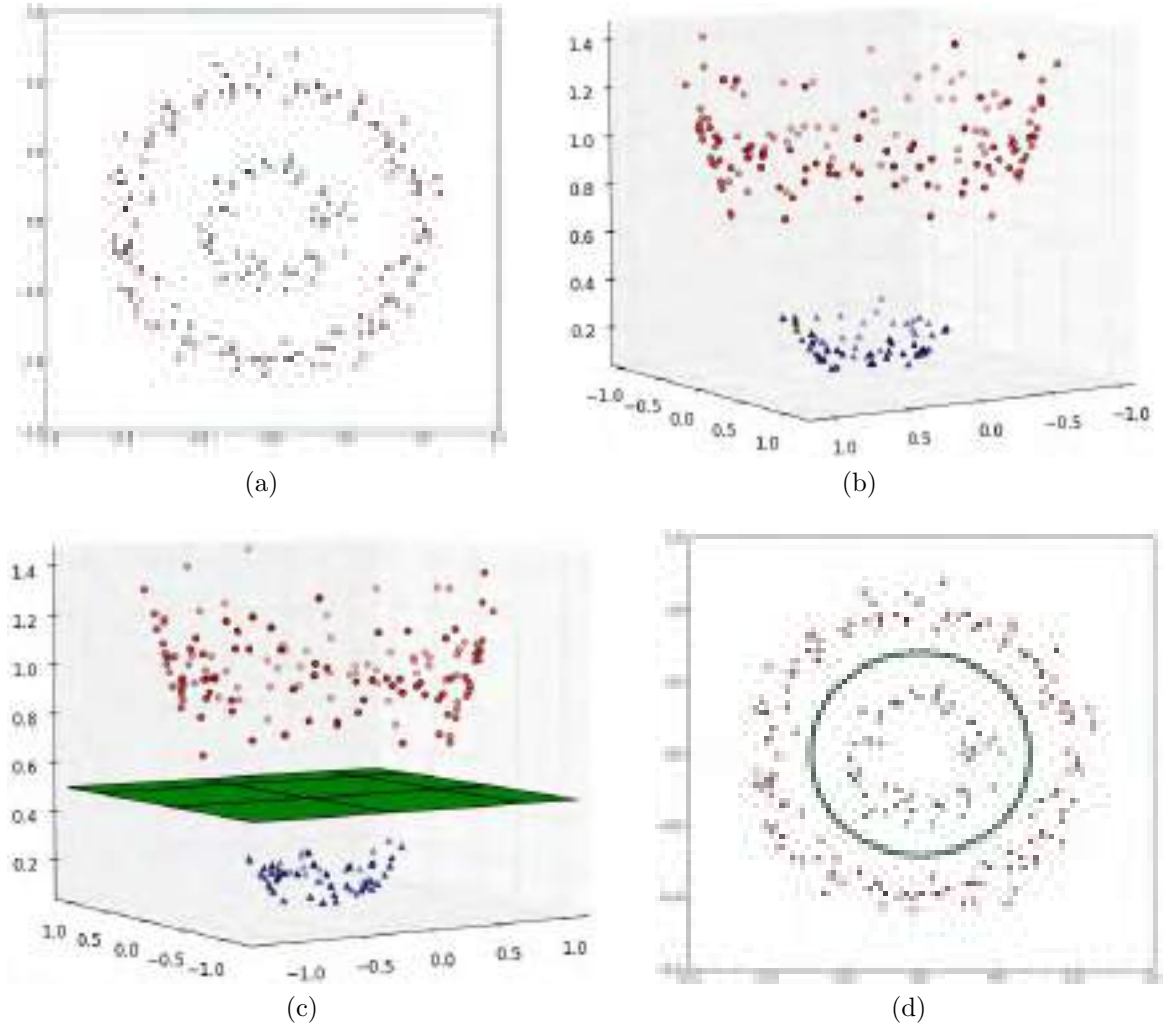


FIGURE 4.15: *A demonstration of the kernel trick whereby (a) a two-dimensional linearly non-separable data set is (b) transformed to three dimensions according to a polynomial kernel function so that (c) a linear hyperplane can be inserted which (d) separates the two different classes of observations [125].*

This method is possible due to the fact that the dual problem of (4.52)–(4.54), together with the hypothesis of (4.47), is based on the inner product $(\boldsymbol{x}^{(i)})^T\boldsymbol{x}^{(j)}$ of two observations. This, therefore, allows each observation $\boldsymbol{x}^{(i)}$ to be replaced by a non-linear hypothesis $\phi(\boldsymbol{x}^{(i)})$, such that $\phi((\boldsymbol{x}^{(i)})^T)\cdot\phi(\boldsymbol{x}^{(j)}) = k(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)})$, according to which the dual problem may be solved. The appropriate choice of kernel function is, to a large extent, problem-specific, although Hill and Lewicki [107] indicated that the *radial basis function* (RBF) kernel is, by far, the most popular choice. Other well-known kernel functions include *polynomial* and *sigmoid* functions [62].

4.9 The multi-layer perceptron model

The MLP is a popular non-parametric supervised learning algorithm that can be used for both classification and regression tasks. This computational model architecture was initiated in 1943 by McCulloch and Pitts [163] who developed a mathematical model which mimicked the functionality of a biological neuron in the human brain. In 1958, Frank Rosenblatt [223], an American psychologist, proposed a more generalized computational model than the McCulloch-Pitts neuron in which weights and thresholds are learnt over time by the model — thus the artificial neuron/perceptron model was born.

Given a set of m observations $\mathcal{X} = \{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(M)}\}$, where the ith observation $\boldsymbol{x}^{(i)} = \langle x_1^{(i)} \cdots x_N^{(i)} \rangle$ has N feature values, and associated continuous target values $\mathcal{Y} = \{y^{(1)}, \ldots, y^{(M)}\}$, an MLP is able to learn a non-linear functional mapping $f : \mathcal{X} \mapsto \mathcal{Y}$ between the values of the features and target variables. This is achieved through multiple layers of interconnected neurons, as illustrated in Figure 4.16, where the values of the features act as the network's *input layer* and the value of the network's *output layer* best corresponds to the defined target variables. Between the input and output layers, the network may contain $h \geq 1$ *hidden layers*. For the ith neuron within the jth hidden layer, a so-called *activation* is calculated as a linear combination of the preceding weights $\boldsymbol{w}_j$ and their corresponding input signals (*i.e.* $\boldsymbol{w}_j^T \boldsymbol{x}$) which is then followed by the application of a specified activation function $g_{ji}(\cdot) : \mathbb{R} \mapsto \mathbb{R}$. A bias b_j is also associated with each neuron within the jth hidden layer which allows for the translation of the activation function. The set of weights $\mathcal{W} = \{\boldsymbol{w}_1, \ldots, \boldsymbol{w}_{h+1}\}$ of the MLP are typically *trained* using the well-known backpropagation method in conjunction with a multi-start gradient-based optimisation approach, usually aimed at minimising the total MSE between the predicted and actual target value [88].

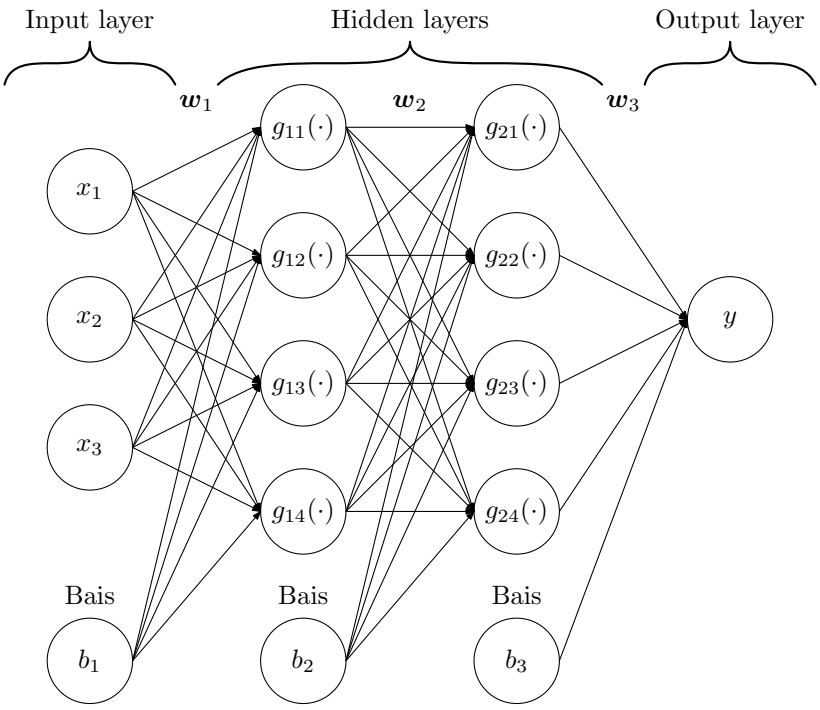


FIGURE 4.16: *A simple two-layer MLP with $N = 3$ input nodes and one output node used to predict the value of a continuous target variable.*

There is little consensus in the ML literature as to specifying the maximum size of an MLP. It is, however, often acknowledged that two hidden layers typically enable the MLP to model

an arbitrary function [13]. Some researchers have also claimed, as a rule-of-thumb, that the maximum number of neurons in a hidden layer should not be more than twice the size of the input layer [15, 260]. In the case of specifying an appropriate activation function, *rectified linear units* (ReLus) typically provide better convergence than sigmoid or hyperbolic tangent functions and are, therefore, the most popular choice [180]. There is also a wide variety of gradient-based optimisation approaches towards training an MLP, the most popular being the traditional stochastic gradient decent method as well as other variants such as RMSProp [269] and Adam [127]. Kingma *et al.* [127] showed that Adam's bias-correction often allows the algorithm to outperform RMSprop slightly towards the end of an optimisation run as gradients become sparser. Ruder [227] concurred that, due to this fact, Adam might be the best overall optimisation choice.

4.10 Chapter summary

The aim in this chapter was to introduce the reader to various important concepts and methods residing in the field of ML. The chapter opened in §4.1 with an overview of ML and the various associated paradigms of the field, namely supervised, unsupervised, semi-supervised and reinforcement learning, together with a more in-depth discussion of the supervised learning procedure. This was followed in §4.2 by a review of the practical considerations associated with the various data formats and how data should be cleaned appropriately, together with a description of the methods of cross-validation and bootstrap re-sampling data splitting often employed during the evaluation of a learning model. Thereafter, additional considerations related to the learning model were described in §4.3, which included an appropriate choice of performance evaluation measure, the notion of a learning model overfitting or underfitting the training data, and the practical considerations associated with a model's interpretability. The reader was then introduced to the concept of generalised linear models in §4.4 which was used to derive the well-known OLSR and logistic regression statistical models. The focus thereafter shifted in §4.5 to a description of various tree-based learning models, namely CART, random forests and the C4.5 algorithm. Thereafter, the working of two relatively simple learning models, namely the naive Bayes classifier and the kNN algorithm, were discussed in §4.6 and §4.7, respectively. This was followed in §4.8 by an in-depth description of how the SVMs learning model works in both the linearly separable and linearly inseparable cases, together with a description of how the kernel trick may be applied in the context of SVMs. The chapter concluded in §4.9 with a description of the popular MLP model.

Part II

System design

CHAPTER 5

Road accident risk evaluation framework

Contents

5.1 The data mining process 83
5.2 A high-level modular data mining framework 86
5.3 The proposed RARE framework 88
5.3.1 Configuring the data appropriately 89
5.3.2 Combining multiple data sets 91
5.3.3 Creating new data attributes 92
5.3.4 Converting data into information 93
5.4 Chapter summary 94

The aim of this chapter is to provide the reader with a top-down description of the RARE framework proposed in this dissertation for the purpose of conducting RAR investigations. The chapter opens in §5.1 with an overview of what data mining entails and the major stages of the data mining process. This facilitates a preliminary discussion in §5.2 during which a generic data mining framework is first proposed. This generic framework is then adopted in §5.3 as the basis for designing the proposed RARE framework. The chapter finally closes in §5.4 with a brief summary of its contents.

5.1 The data mining process

In recent years, the tasks associated with data mining have increasingly been aided by the developments of specialised software [226]. In the field of software engineering, the design and development of immutable software is, in most cases, strongly discouraged. This may be attributed to the underlying assumption that the task for which the software has been designed is static (*i.e.* is not subject to change). The real world is, however, inherently dynamic in nature and, given enough time, results obtained by means of a rigid software system will become obsolete. Many analysts, therefore, gravitate towards a more modular approach when developing a purpose-built data mining solution since it allows for greater system flexibility, re-usability and maintenance [164], while the overall design of the software is usually achieved within the context of a data mining reference model [30].

The *cross-industry standard process for data mining* (CRISP-DM) is a freely available application-neutral reference model that describes the common approaches adopted in the field of data mining [290], and is regarded as the most widely-used reference model [29]. This model was

conceived in 1996 by four industry leaders in the data mining market with input from over 200 data mining users and service providers [245]. The primary drive behind this endeavour was the need to formalise the various stages of the data mining process so that a common reference point could be established amongst practitioners. CRISP-DM decomposes the life cycle of a data mining project into six distinct phases, all of which revolve around project-specific data [43], as illustrated in Figure 5.1. The inner light grey arrows between the six phases suggest that data mining is an iterative process, often requiring an analyst to revisit previous phases as new information is learnt during a project. The outer circle symbolises that data mining itself is also a cyclic process — a deployed solution may often trigger new, more focussed questions about the topic under consideration [245].

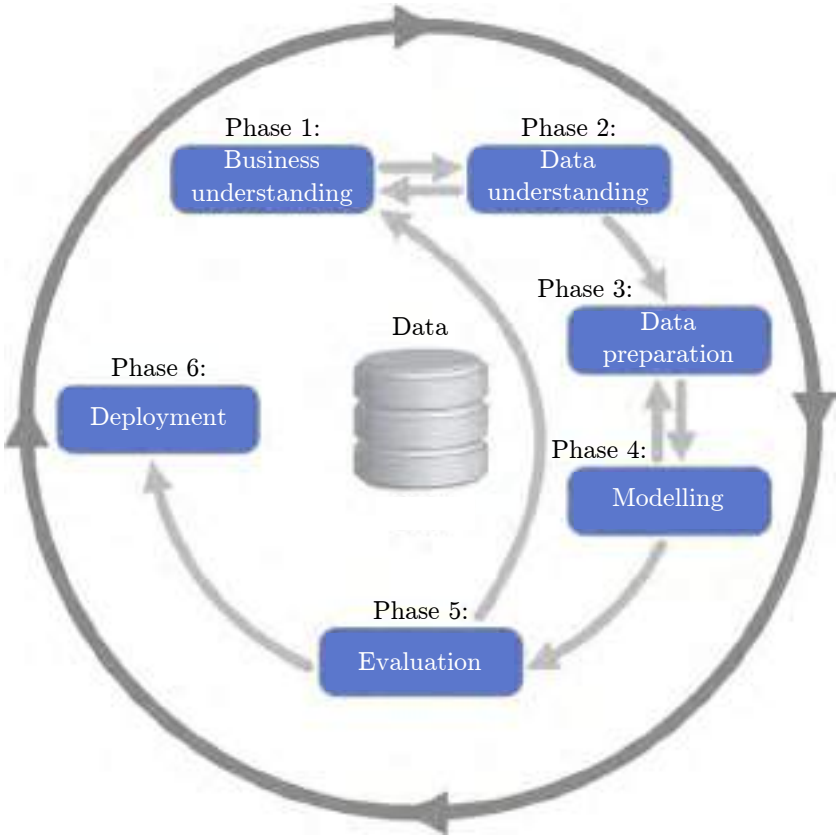


FIGURE 5.1: *The six phases of the CRISP-DM reference model [116].*

Perhaps the most important stage to consider before undertaking any type of organisational project is to develop an initial *business understanding* — a step during which the project objectives are understood by the analyst from a business perspective and translated into a preliminary plan for achieving these objectives. In Phase 1 of the CRISP-DM reference model, business understanding is composed of four tasks, which aim to: (1) *Determine the business objectives* in order to uncover the important factors which may influence the outcome of a project, (2) *assess the situation* of available resources required for the successful completion of a project, (3) *determine the data mining goals* in terms of the business goals, and (4) *produce a project plan* describing an intended plan of action for achieving a set of data mining goals.

Once the business and data mining goals have been aligned, the next stage relies on developing a preparatory *data understanding* — a step during which the analyst collects and becomes familiar with the data so as to establish data suitability for achieving the project goals. During Phase 2 of the CRISP-DM reference model, data understanding is composed of four tasks, which are

concerned with: (1) *Collecting the initial data* to be used in the project, (2) *describing the data* in terms of their format and meta-information, (3) *exploring the data* in order to develop initial hypotheses, and (4) *verifying the data quality* with respect to their completeness and correctness. There is a reciprocative link between the business understanding and data understanding phases in the CRISP-DM reference model since the formulation of the data mining goals often requires at least some understanding of the available data. If the data are deemed not suitable for achieving these defined goals, the analyst must either attempt to acquire additional data streams or return to the business understanding phase with a view to redefine the project plan.

Arguably one of the most time-consuming stages in the data mining process is that of *data preparation* [44] — a step which ensures that the representation and quality of data are sufficient for conducting further analysis [208]. During Phase 3 of the CRISP-DM reference model, data preparation comprises five tasks, which require: (1) *Selecting data* based on specific criteria which are deemed relevant to the specific question under consideration, (2) *cleaning the data* so as to ensure that credible results are obtained from an analysis, (3) *constructing data* by producing entirely new or derived attributes so as to enhance data quality, (4) *integrating data* by combining multiple sets of data into one coherent data set (if applicable), and (5) *formatting the data* so that they are suitable for analytical investigations.

Once the data are in the correct format and are deemed of a sufficiently high quality, the next stage requires the analyst to perform sophisticated data *modelling* — a step involving the selection and application of various modelling techniques and algorithms coupled with the fine-tuning of model parameters in order to produce optimal results. During Phase 4 of the CRISP-DM reference model, modelling comprises four tasks, which are concerned with: (1) *selecting the modelling technique* deemed most appropriate for the task at hand, (2) *generating the test design* of an algorithm with a view to determine the strength of a model empirically, (3) *building the model* based on a prepared data set, and (4) *assessing the model* in terms of specified performance criteria and model interpretability. There is a reciprocative link between the data preparation and modelling phases in the CRISP-DM reference model since each modelling technique has its own specific data format requirements. Often an analyst is required to revisit the data preparation phase in order to reconfigure the data or construct new data attributes according previously unforeseen model specifications.

Before proceeding to the final deployment of a model, the analyst should perform a thorough model *evaluation* — a step during which the model is scrutinised to determine whether all business objectives have been addressed sufficiently. During Phase 5 of the CRISP-DM reference model, evaluation is composed of three tasks, which require the analyst to: (1) *Evaluate the results* produced by a model in terms of its ability to answer the original business objectives, (2) *review the process* followed when constructing the proposed model, and (3) *determine the next steps* of the project which may include either progressing to the deployment phase or returning to the business understanding phase if important business requirements have not been met by the model.

The final stage of the data mining process is concerned with the *deployment* of the project into the business environment. During Phase 6 of the CRISP-DM reference model, deployment comprises four tasks, which require: (1) *Planning the deployment* of a proposed model into the business, (2) *planning the monitoring and maintenance* strategy to ensure that the model is able to function on a day-to-day basis in the business environment, (3) *producing a final report* which documents the model's construction together with necessary model inputs and outputs, and (4) *reviewing the project* with a view to establish the project successes, shortcomings and possible areas of improvement for use in future projects.

5.2 A high-level modular data mining framework

As stated in Chapter 1, the primary research aim in this dissertation is to propose a RARE framework which utilises techniques and methods from the realms of network spatial analysis and machine learning. In view of the preceding section, it is proposed that this framework should be modular in nature. In this way, the modules proposed in this dissertation may not be fully limited to RA analysis but may also be re-used or re-purposed in other problem application domains. Furthermore, if significant advances are made in RA research, the current system may be appended or reconfigured in an appropriate manner so as to facilitate inclusion of these developments. To achieve this objective, it is proposed that all modules should be conceived within the context of a high-level generic framework which encapsulates the major development phases of the data mining process, with a specific focus on the data understanding, data preparation, modelling and evaluation phases defined in the CRISP-DM reference model (*i.e.* Phases 3–5 in Figure 5.1). In other words, a *domain-specific* framework should comprise a number of configurable modules which function within a *high-level* data mining framework.

The framework in Figure 5.2 is proposed as the generic data mining framework in which all modules proposed in this dissertation may be contained. This framework comprises five functional elements, namely a central *database* and four separate classes of modules dedicated to the processes associated with *configuring*, *combining*, *creating*, and *converting* data. The ith class consists of k_i modules (for $i \in \{1, 2, 3, 4\}$) which are assumed to be crafted by a systems designer with a view to accomplish the specific data mining task at hand. This framework, therefore, allows for flexibility in that further modules may be designed and appended to the existing framework if additional functionality is required. Furthermore, the framework also allows for re-usability in the sense that modules may be re-configured to function in other framework designs.

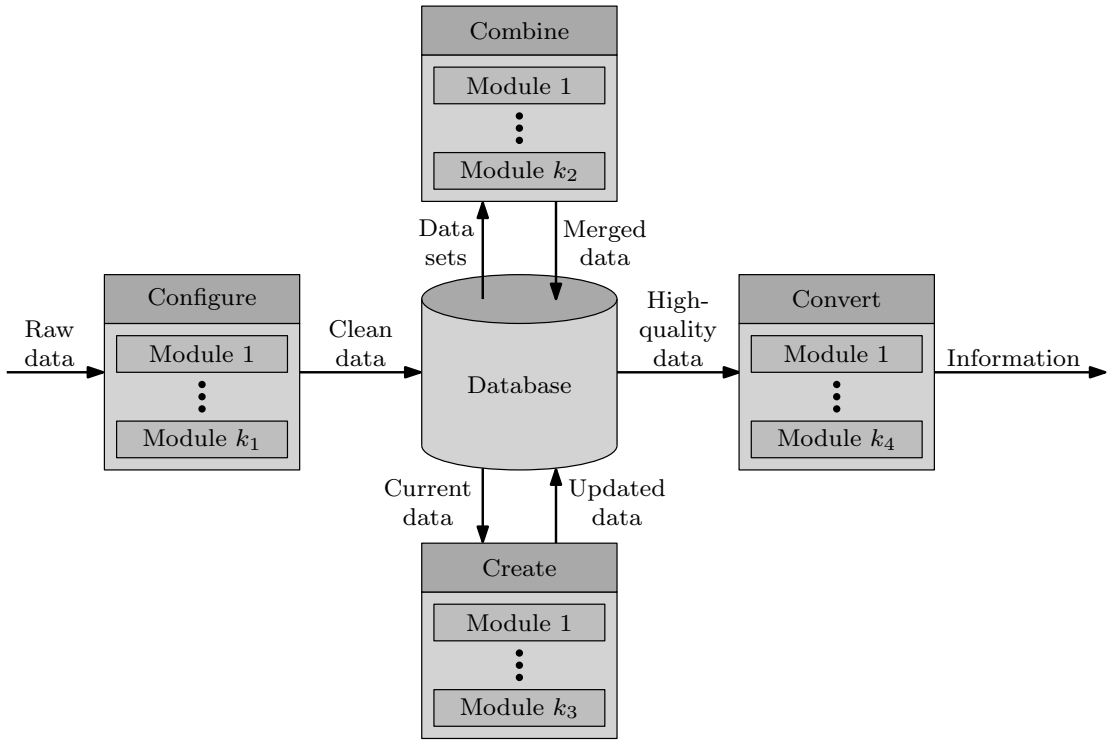


FIGURE 5.2: *Schematic representation of a generic data mining framework showing the flow and transformation of data between functional elements.*

The primary input to this framework is a set of raw data[1]. These data are provided to one of the k_1 modules residing in the Configure class, which may facilitate tasks associated with converting unstructured or semi-structured data into structured data, formatting the data according to a specific framework format, correcting any missing or erroneous values that may exist in the data, and defining the meta-data associated with the various data attributes — all of which would require the discretion of a user. The data sets yielded as output by this class should be considered by the user as appropriately configured, and should accordingly be stored in the database. In the field of data science, these modules accomplish the typical tasks of a so-called *data wrangler*.

The purpose of the database element is to act as a central repository for the data used during the analysis. All data stored in the database are assumed to be logically organised according to some structural type. The database allows modules to perform database queries in order to retrieve the relevant data and provides storage for the various data outputs produced by the modules. The database, therefore, serves as the functional link between the different classes of modules contained in the framework.

In the real world, it may be the case that separate sets of data are actually related in some way. A data record may, for example, be related to multiple records which are stored in tables in a relational database. In the context of data analysis, however, it may be advantageous to represent these two separate tables as a one coherent table so as to maximise the subsequent information extraction. If this is the case, various data integration methods may be devised for this merging procedure. Each of the k_2 modules in the Combine class are, therefore, responsible for combining two sets of data currently stored in the database into a single, coherent data set. This may be achieved, for example, by taking advantage of a unique identifier linking data sets or relating data records based on their separate spatial information (if applicable).

Another important consideration in the field of data mining is the process of data construction (often called *feature engineering* in the machine learning literature) — a process whereby new data attributes are constructed so as to enhance the quality of a data set. Andrew Ng, a well-known machine learning pioneer and co-founder of Google Brain, commented that feature engineering is essential in applied machine learning but is typically difficult, time consuming, and requires domain-specific knowledge of the data under consideration [184]. It is, subsequently, proposed that this high-level framework should incorporate modules which facilitate the engineering of new data attributes with the possible guidance of a domain expert. Each of the k_3 modules in the Create class is responsible for the creation of new information derived from one or more current attribute(s) in a data set. This may be achieved, for example, through procedural, rule-based, or discretisation and/or continuisation methods. All modules in the Combine and Construct classes are, therefore, concerned with enhancing the data currently stored in the database such that they are deemed of a sufficiently high quality.

Once a high-quality data set has been established, it may be passed to one of the k_4 modules residing in the Convert module. The primary focus of these modules is to convert data into useful information and communicate this information in an effective manner to the user. This may be achieved, for example, by the application of various *modelling* techniques such as the application of various machine learning algorithms or through specialised procedures unique to a specific problem domain. Modules in this class should also facilitate the *evaluation* of the various model results. Every class other than the Convert class may, therefore, be seen as a form of *data preparation*, which, as previously mentioned, is essential during any data mining investigation. In this way, the proposed framework encompasses Phases 3–5 of the CRISP-DM reference model.

[1]Raw data inputs, in this case, may refer to structured or unstructured data sets.

5.3 The proposed RARE framework

In this section, the high-level working of the RARE framework proposed in this dissertation is formally documented. The reader is introduced to the various module functions and their dependencies which will prove useful in the following chapter when the working of every individual module in each of the various framework classes is discussed in greater detail. Before presenting the proposed framework, however, it is first necessary to state formally the assumptions related to the input data. Given some bounded region on the earth's surface, this framework assumes four data inputs stored in a tabular format, which describe:

1. the road *accidents* that have occurred within the bounded region during a specified time period,
2. the *vehicles* involved in the aforementioned accidents,
3. the *victims* associated with the aforementioned accidents and vehicles,
4. and the *road* sections present within the bounded region.

The assumed relationship between these data inputs is illustrated in Figure 5.3 by means of an entity-relationship diagram, as described by Kendall and Kendall [123]. In this diagram, a data generating entity is represented by a rectangular box while lines denote a specific type of relationship between entities, where a two-short-marks symbol (||) means "exactly one", a crow's-foot symbol with a short mark (>|) means "one or more" and a crow's-foot symbol with a open circle (>o) means "zero or more." The relationship between the entities in Figure 5.3 may, therefore, be interpreted as follows: An accident is assumed to have been caused by one or more vehicle(s), whereas a vehicle is associated with exactly one accident. In the same way, an accident is assumed to have at least one victim which may be associated with a single vehicle[2]. Lastly, an accident is assumed to take place on exactly one road whereas a road may contain one or more accident(s).

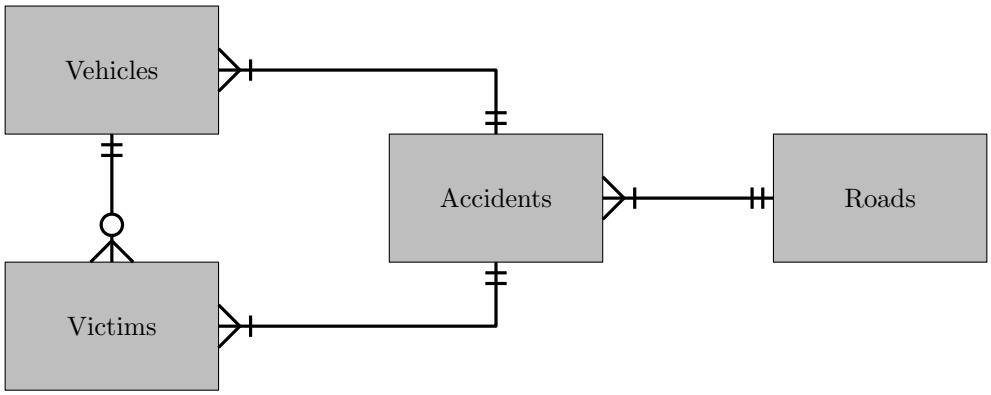


FIGURE 5.3: *An entity-relationship diagram showing an accident entity's relationship with vehicle, victim and road entities.*

Having established these relationships, the framework assumes that the relationship between accidents, vehicles and victims is explicitly defined (*i.e.* there exists a unique identifier linking one or more vehicle and victim record(s) to a corresponding accident record). This is a reasonable

[2]Although one may argue that a specific vehicle or victim may, in their lifetime, be involved in more than one accident, a vehicle and victim record is considered unique in this formulation. In the case of pedestrian victims, their records are associated with the vehicle with which they first made contact.

assumption since, upon investigation of an RA by a data recorder, the attributes of an accident and the vehicle(s) and victim(s) involved are typically recorded and stored together. The framework does not, however, assume any explicit identifier linking accident and road records. In most cases, it is fair to assume that attributes pertaining to a road (*e.g.* traffic flow counts) are recorded and stored separately from RAs as they are not typically managed by the same government department. The framework furthermore assumes that the geographical locations of accidents and roads are defined by a set of spatial attributes. More specifically, it is assumed that all accident and road records are spatially represented by points and straight line segments, respectively, in two-dimensional Cartesian space.

The design of the proposed RARE framework, together with the data movement between modules, is illustrated in Figure 5.4 by means of a *data flow diagram* (DFD), as introduced by Kendall and Kendall [123]. A DFD is a graphical representation of the flow of data between various entities, processes and data storage points which exist in a proposed or existing system, and serves as a form of documentation. A DFD makes use of four symbols: A double rectangle is used to represent an external *entity* which may send and receive data to and from the system, respectively, an arrow is used to indicate the movement of data between various processes in the system, a rectangle with rounded corners is used to represent a *process* of transforming data[3], and an open-ended rectangle denotes a *datastore* which is an unspecified physical data storage facility. The methodology follows a top-down approach by first specifying a Level 0 Diagram which provides a broad overview of the complete structure of the system. Each process may also be decomposed into further sub-processes by means of a *child diagram* if further details of the process's inner-workings is required.

The proposed RARE framework illustrated in Figure 5.4 has, in this case, been superimposed onto the data analysis framework of Figure 5.2 so as to demonstrate the relationship between the two figures. All user parameter inputs to the various modules have, however, been excluded from the DFD as these will be introduced and explained in greater detail in the following chapter. A high-level elucidation of each of the module functions is provided in the remainder of this chapter.

5.3.1 Configuring the data appropriately

The Configure class of the proposed RARE framework in Figure 5.4 is illustrated in greater detail in Figure 5.5, and comprises $k_1 = 2$ modules connected in series. As previously mentioned, the modules in this class are collectively responsible for structuring, formatting and cleaning the raw data inputs such that they are transformed into configured data records. A key assumption of the framework is that the four data inputs are structured in a tabular format. The framework does not account for unstructured data — providing such a functionality would require the design of specific modules that are able to convert data currently stored in unstructured formats (*e.g.* scanned documents) into a formal data structure (*e.g.* a tabular format) which is computer-readable. Such a functionality is considered outside the scope of this dissertation. The primary functions of Module 1.0 are to accept the raw (structured) accident, vehicle, road and victim data inputs entered into the system by the user and format these data sets in an appropriate manner. The user begins by defining the file paths corresponding to the raw data. During the formatting procedure, the user is required to define the data type (as described in Figure 4.2) associated with each of the attributes contained in each respective data set. In addition, this module also facilitates the imputation of missing values in each of the data sets — the majority

[3]Note that the "process" symbol used by Kendall and Kendall [123] is synonymous with modules in the context of the framework proposed in this dissertation.

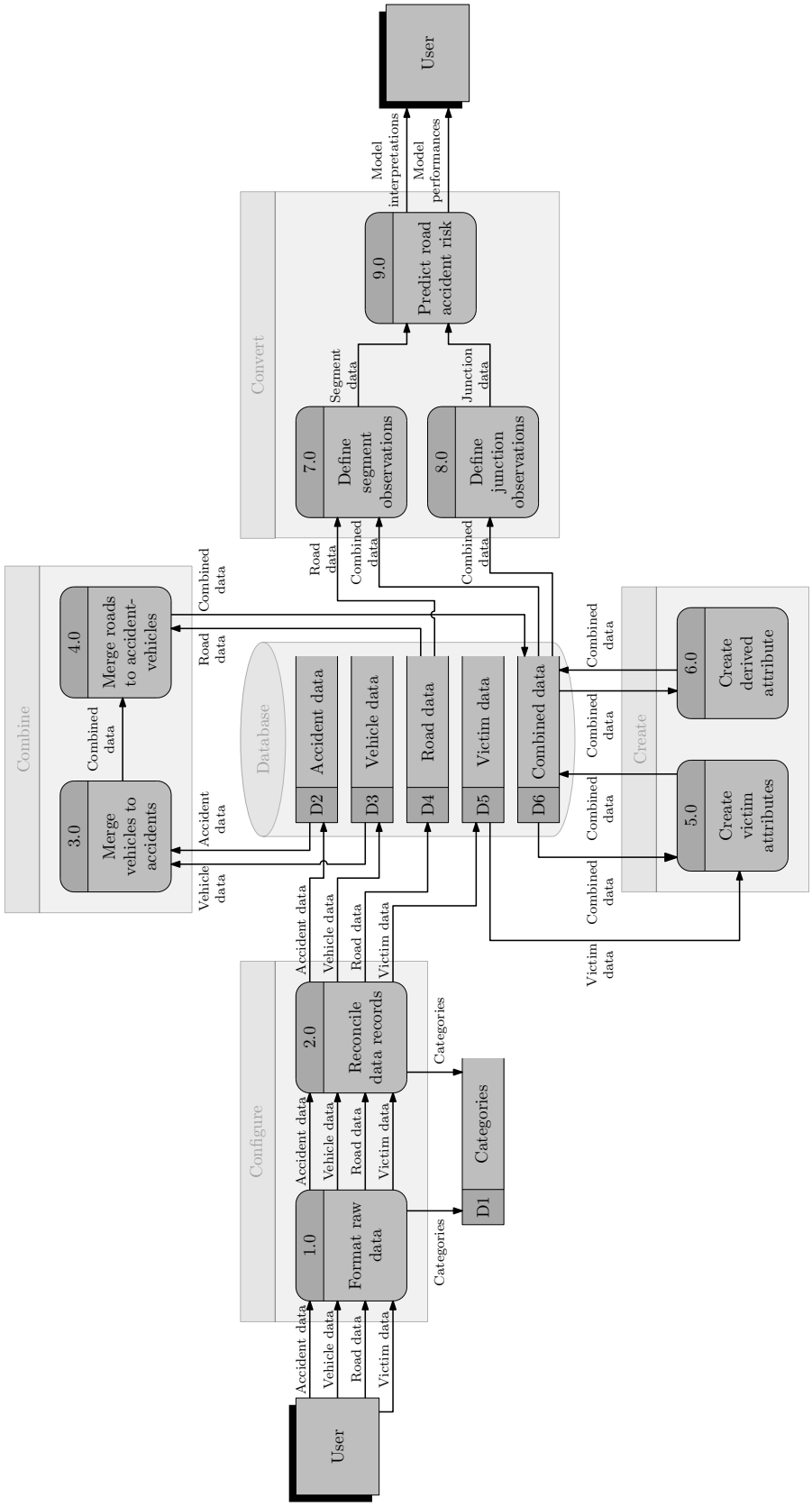


FIGURE 5.4: *The Level 0 Diagram of the proposed RARE framework.*

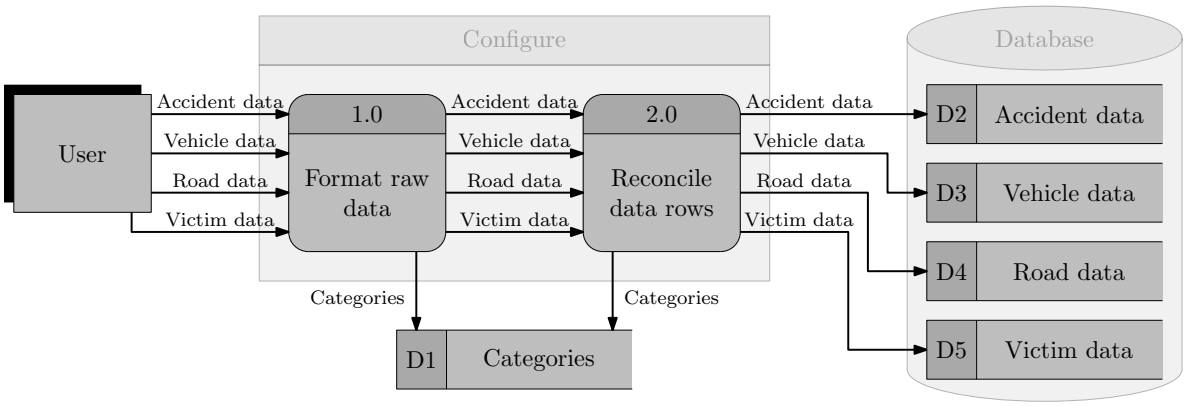


FIGURE 5.5: *A portion of the DFD in Figure 5.4 showing in greater detail the flow of data between modules in the Configure class.*

of algorithms within the realm of ML cannot function in the presence of unknown values. The accident, vehicle, road and victim data sets exiting Module 1.0 should be considered, by the user, to be appropriately formatted and cleaned.

The purpose of Module 2.0 is to ensure that all data records are configured in the specific format required by the RARE framework. In the case of accident, vehicle and victim data, if any records were deleted during the imputation phase of Module 1.0, the corresponding records in the other associated data sets are also deleted. In the case of road data, multiple annual records associated with a specific road may be consolidated in an appropriate manner into a single road record. In this process, it is also necessary to identify attributes in the accident data set that describe inherent properties of the road. The accident, vehicle, road and victim data sets exiting Module 2.0 should be considered, by the user, as appropriately configured and are stored in separate database tables, designated by Datastores D2, D3, D4 and D5, respectively.

5.3.2 Combining multiple data sets

The Combine class of the proposed RARE framework in Figure 5.4 is illustrated in greater detail in Figure 5.6, and also comprises $k_2 = 2$ modules connected in series. As previously mentioned, each module in this class is responsible for combining two sets of data into a single, coherent data set. During this stage, it is important to note the relationship between the data entities, as previously described in Figure 5.3. The accident, vehicle and road data sets are, at present, stored in separate tables in the database. Consequently, if one were, for example, to conduct a standard supervised machine learning analysis, attributes in the vehicle data could not be used to explain a target variable in the accident data since the current structure of the data does not allow for any inference to be drawn between separate database tables. It is, therefore, proposed that two specialised procedures are devised for first merging vehicle records with their corresponding accident records and, secondly, associating each combined record with a corresponding road record, where the former is realised in Module 3.0. This is achieved by first condensing multiple vehicle records associated with a single accident into a single record and then appending this condensed vehicle record to the corresponding accident record. This module functions on the assumption that there is a unique identifier linking one or more vehicle record(s) to a corresponding accident record. In this way, a set of combined accident-vehicle data is constructed and passed to Module 4.0 as input.

As may be expected, the function of Module 4.0 is to merge a combined accident-vehicle record with an associated road record. Unlike Module 3.0, no explicit identifier is assumed to connect

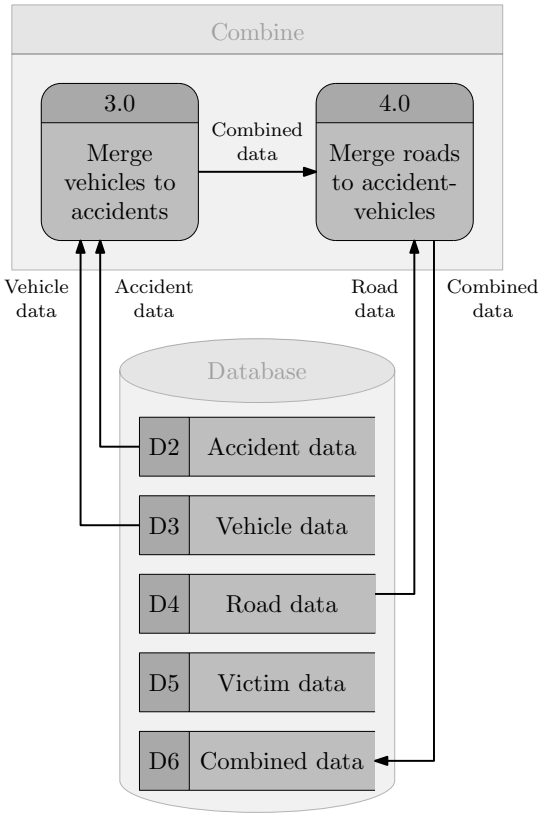


FIGURE 5.6: *A portion of the DFD in Figure 5.4 showing in greater detail the flow of data between modules in the Combine class.*

an accident record with a corresponding road record. An alternative approach is, therefore, adopted by utilising the presumed spatial features of accidents and roads. This approach seeks, for each accident-vehicle record, the shortest perpendicular distance connecting an accident point with some road line segment. In this way, an accident-vehicle record is associated with and concatenated to a single road record. The successive application of Modules 3.0 and 4.0 to the accident, vehicle and road data sets produces a combined accident-vehicle-road data set which is stored in its own distinct database table, designated by Datastore D6.

5.3.3 Creating new data attributes

The Create class of the proposed RARE framework in Figure 5.4 is illustrated greater detail in Figure 5.7, and yet again comprises $k_3 = 2$ modules. As previously mentioned, a module in this class is responsible for the creation of new information derived from one or more attribute(s) currently contained in a data set. In the accidents data set, there is typically a measure of accident severity, which is described as the greatest severity sustained by one (or more) of the victims involved in an accident. If, for example, an accident involved three victims, two of which sustained slight injuries and the other sustained a fatal injury, the overall accident severity would be fatal. Detailed victim data, therefore, provides a finer level of granularity in terms of the overall severity of an accident and are, thus, used in Module 5.0 to construct a continuous severity measure of RA severity which better reflects the overall severity of an accident. This module also facilitates the construction of new accident attributes derived from the associated victim attributes. One may, for example, define a new attribute called `PedestrianAccident` which has the value of `True` if any of the associated victim records have a value of `Pedestrian`

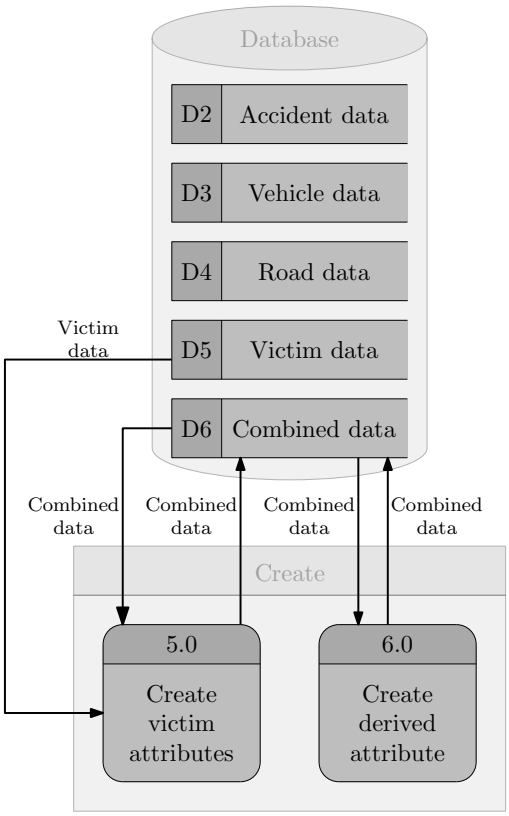


FIGURE 5.7: *A portion of the DFD in Figure 5.4 showing in greater detail the flow of data between modules in the Create class.*

for the `VictimClass` attribute. Once the new victim attributes have been constructed, they are appended to the data currently stored in Datastore D6.

Module 6.0 is used to construct a new attribute derived from an attribute that currently exists in one of the data sets stored in the database. If the new attribute is discrete in nature, the first step involves the user conceiving new attribute categories. The user is then required to define a function of the current combined data attributes which defines a value of the new attribute for each record in the combined data set. In the case of defining a new discrete attribute, this function must assign each combined data record to one of the user-defined categories. If a new continuous attribute is pursued, the defined function is required to associate a numerical value, based on current combined data attribute values, with each combined data record. In this way, a user may define his or her own attributes according to the values of an attribute (or combination thereof) currently contained in the combined data set. This module may, for example, also allow the user to construct an attribute in an RA data set which exists in another unrelated road data set. Module 5.0 may, therefore, allow a user to facilitate a comparison between multiple independent RA data sets (possibly even collected in different countries).

5.3.4 Converting data into information

The Convert class of the proposed RARE framework in Figure 5.4 is illustrated in greater detail in Figure 5.8, and comprises $k_4 = 3$ independent modules. As previously stated in §1.2, the aim of the RARE framework is to quantify and characterise the rate and severity of an RA along both road *segments* and at road *junctions*. Modules 7.0 and 8.0, therefore, facilitate the conversion of accident observations into road segment and junction observations, respectively.

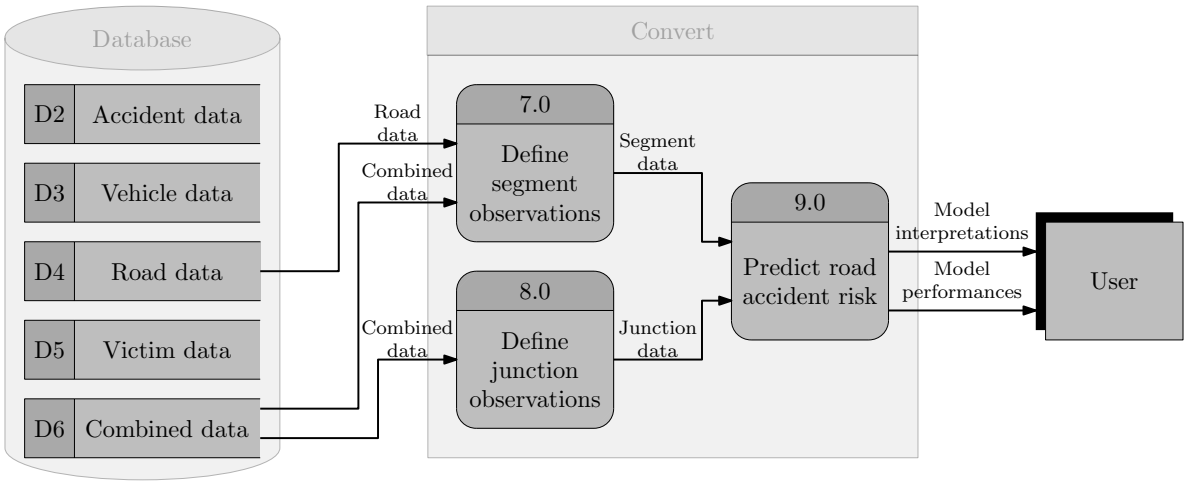


FIGURE 5.8: *A portion of the DFD in Figure 5.4 showing in greater detail the flow of data between modules in the Convert class.*

During both these stages, the rate and severity of RAs are appropriately quantified and may be visualised spatially by the user.

The segment and junction data outputs act as observation inputs to Module 9.0 which attempts to characterise road accident risk through a standard supervised learning procedure, as described in § 4.1.2. This is achieved by incorporating multiple supervised learning models to predict those attributes that are most indicative of road observations with higher RA risk or RA severity. An interpretation of the best performing learning model is then determined by application of the SHAP method described in §4.3.3. Both the learning model performances and selected model interpretations are then presented to the user.

5.4 Chapter summary

This chapter contained an overview of the RARE framework proposed in this dissertation. The chapter opened in §5.1 with a discussion on the various data mining processes within the widely adopted CRISP-DM reference model. With this in mind, a high-level generic framework was thereafter proposed in §5.2 for structuring a modular data mining system. In §5.3, this framework was then used as a basis for the design of the proposed RARE framework. The relationship between the three data inputs was first described by means of an entity-relationship diagram and the input data assumptions were clearly stated. The proposed RARE framework was then introduced in the form of a DFD, and the functions of the various modules were finally explained in more detail.

CHAPTER 6

Framework module components

Contents

6.1 Formatting raw data 95
6.2 Configuring data attributes 97
6.3 Combining vehicle records to accident records 103
6.4 Combining road records to accident-vehicle records 107
6.5 Creating severity attributes 113
6.6 Constructing a derived attribute 115
6.7 Modelling road accident risk along a road segment 116
6.8 Modelling road accident risk at a road junction 118
6.9 Predicting road accident risk 122
6.10 Chapter summary 124

The aim in this chapter is to supplement the previous chapter by providing an in-depth explanation of the components comprising the RARE framework presented there. The chapter opens in §6.1 with a description of how the input data are formatted and cleaned in an appropriate manner. A discussion then follows, in §6.2, on how the proposed RARE facilitates the formatting as well as classification of attributes into informative categories by the user. In §6.3, the method for merging multiple vehicle records to a single record is presented. This is followed, in §6.4, by a further discussion on the method for spatially linking accidents to the roads on which they occurred. Thereafter, the method for creating accident severity attributes is described in §6.5, and this is followed, in §6.6, by a description of how the user may enhance the quality of a data set by constructing new derived attributes. The focus then shifts in §6.7 and §6.8 to a description of a novel approach towards quantifying RAR along road segments and at road junctions, respectively. These results are then provided in §6.9 as input to a supervised learning procedure for identifying the most indicative factors contributing to higher levels of RAR. The chapter finally concludes in §6.10 with a brief summary of its contents.

6.1 Formatting raw data

The primary function of Module 1.0 in the RARE framework of Figure 5.4 is to accept the four raw data inputs entered into the system by the user and to configure these data sets so that they are considered clean and appropriately formatted. This is achieved by means of two sub-modules, as illustrated in Figure 6.1. In Sub-module 1.1, the user is first required to define

the data types corresponding to each of the attributes in the four sets of data. These data sets are then cleaned in Sub-module 1.2 by applying a specified data imputing method. After this procedure, a data set is deemed appropriately formatted. These two sub-modules are now described in greater detail.

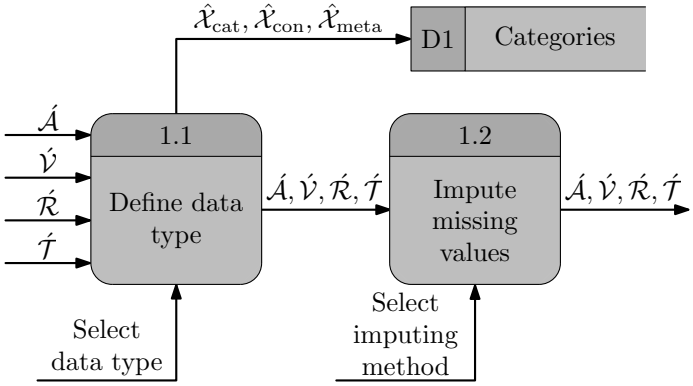


FIGURE 6.1: *A child diagram of Module 1.0 showing in greater detail the flow of data between sub-modules.*

In Sub-module 1.1, the user begins by defining the file paths corresponding to the four aforementioned data sets. The first of these raw data sets is a set of m_A historical accident records gathered over Y years, denoted by $\acute{\mathcal{A}} = \{\acute{\boldsymbol{a}}^{(1)}, \ldots, \acute{\boldsymbol{a}}^{(m_A)}\}$, where $\acute{\boldsymbol{a}}^{(i)} = \langle \acute{a}_1^{(i)} \cdots \acute{a}_{n_A}^{(i)} \rangle$ is a tuple of n_A accident attribute values. The acute notation indicates that the data are still, in the current form, considered raw. In the same way, the user is expected to input a set of m_V vehicle records, denoted by $\acute{\mathcal{V}} = \{\acute{\boldsymbol{v}}^{(1)}, \ldots, \acute{\boldsymbol{v}}^{(m_V)}\}$, where $\acute{\boldsymbol{v}}^{(i)} = \langle \acute{v}_1^{(i)} \cdots \acute{v}_{n_V}^{(i)} \rangle$ is a tuple of n_V vehicle attribute values, as well as a set of m_R road records gathered over Y years, denoted by $\acute{\mathcal{R}} = \{\{\acute{\boldsymbol{r}}_1^{(1)}, \ldots, \acute{\boldsymbol{r}}_Y^{(1)}\}, \ldots, \{\acute{\boldsymbol{r}}_1^{(m_R)}, \ldots, \acute{\boldsymbol{r}}_Y^{(m_R)}\}\}$, where $\acute{\boldsymbol{r}}_y^{(i)} = \langle \acute{r}_{y,1}^{(i)} \cdots \acute{r}_{y,n_R}^{(i)} \rangle$ is a tuple of n_R road attribute values collected during year y, and finally m_T victim records, denoted by $\acute{\mathcal{T}} = \{\acute{\boldsymbol{t}}^{(1)}, \ldots, \acute{\boldsymbol{t}}^{(m_T)}\}$, where $\acute{\boldsymbol{t}}^{(i)} = \langle \acute{t}_1^{(i)} \cdots \acute{t}_{n_T}^{(i)} \rangle$ is a tuple of n_T victim attribute values. It is assumed that each of the attributes in the raw data sets have a column header specifying a unique attribute name. The attribute identifiers contained in the accident, vehicle, road and victim data sets are denoted by the sets $\hat{\mathcal{A}} = \{\hat{A}_1, \ldots, \hat{A}_{n_A}\}$, $\hat{\mathcal{V}} = \{\hat{V}_1, \ldots, \hat{V}_{n_V}\}$, $\hat{\mathcal{R}} = \{\hat{R}_1, \ldots, \hat{R}_{n_R}\}$ and $\hat{\mathcal{T}} = \{\hat{T}_1, \ldots, \hat{T}_{n_T}\}$, respectively. The hat notation is subsequently introduced to signify that the quantity under consideration refers to an attribute *name* (and not an attribute *value*).

At this point, it is necessary to introduce a new shorthand notation used during the course of this chapter. This notation will prove valuable for specifying new data subsets based on defined conditions containing attribute identifiers. For an arbitrary set of data $\mathcal{X}$, the shorthand notation $\mathcal{X}\{\bullet\}$ is introduced to denote the subset of data containing only the records/attributes satisfying the condition denoted by the place holder $\bullet$. The shorthand notation $\mathcal{X}\{\hat{X}_j = a\}$, for example, denotes the subset of data in $\mathcal{X}$ containing only the records for which the value of attribute $\hat{X}_j$ is equal to the value a. In the same way, the shorthand notation $\mathcal{X}\{\hat{\mathcal{X}} \in \hat{\mathcal{Y}}\}$ denotes the subset of data with attributes contained in the set $\hat{\mathcal{Y}}$.

During the first step of the configuration procedure, accomplished in Sub-module 1.1, the user is also required to categorise the data type associated with each of the attributes contained in each of the aforementioned data sets. For an arbitrary data set $\mathcal{X}$ containing m_X records and n_X attributes, an attribute $\hat{X}_j$ can either be categorised as *discrete*, in which case there are $C_j = |\{x_j^{(i)}\}_{i \in \{1,\ldots,m_X\}}|$ unique categorical values, or an attribute can be classified as *continuous* in nature if $x_j^{(i)} \in \mathbb{R}$ for all $i \in \{1, \ldots, m_X\}$, or finally, an attribute can be classified as a *meta-*

attribute, in which case it is not used in any form of algorithmic calculation. This attribute-specific information will prove useful in an ML context where the choice of algorithm entirely depends on the data type of the target attribute. Knowledge of an attribute's data type is also necessary when selecting a specific imputing method, conducted in the next sub-module. Each of the classified categorical, continuous and meta-attributes are denoted by the sets $\hat{\mathcal{X}}_{\text{cat}}$, $\hat{\mathcal{X}}_{\text{cont}}$ or $\hat{\mathcal{X}}_{\text{meta}}$, respectively, which are subsequently stored in Datastore D1 for future reference.

During this stage, it is necessary for the user to define a set of compulsory attributes which are assumed to be contained in $\hat{\mathcal{X}}_{\text{meta}}$. First, it is assumed that the accidents in $\acute{\mathcal{A}}$ each possesses a *unique accident identifier* $\hat{A}_I \in \hat{\mathcal{X}}_{\text{meta}}$, and that each vehicle and victim record in $\acute{\mathcal{V}}$ and $\acute{\mathcal{R}}$ are subsequently linked by means of *link attributes* $\hat{V}_L \in \hat{\mathcal{X}}_{\text{meta}}$ and $\hat{T}_L \in \hat{\mathcal{X}}_{\text{meta}}$, respectively, to a single accident record in $\acute{\mathcal{A}}$. Furthermore, is assumed that a distinction is made between multiple vehicle and victim records associated with a corresponding accident record by means of a *unique vehicle identifier* attribute $\hat{V}_I \in \hat{\mathcal{X}}_{\text{meta}}$ and a *unique victim identifier* attribute $\hat{T}_I \in \hat{\mathcal{X}}_{\text{meta}}$, respectively. It is also assumed that all road records contain an attribute $\hat{R}_T \in \hat{\mathcal{X}}_{\text{meta}}$ describing the AADT along the segment for each of the Y years worth of data — a measurement that will prove imperative for calculating the rate of RAs. In the same way, it is also assumed that all victim records contain an attribute $\hat{T}_S \in \hat{\mathcal{X}}_{\text{cat}}$ describing a categorical measure of the severity sustained by a victim involved in an accident — a measurement that is necessary for calculating the severity of RAs. Lastly, it is assumed that all accident records contain a categorical attribute $\hat{A}_J \in \hat{\mathcal{X}}_{\text{cat}}$ indicating whether the accident occurred at some form of junction or not — this will prove useful in Module 8.0.

Once a user has completed the attribute data type categorisation, the aforementioned data sets are passed to Sub-module 1.2. In this sub-module, the user is expected to select one of the prescribed imputing methods, as described in §4.2.2. The choice of imputing method is somewhat subjective and typically depends on external knowledge about the data. In the case where an attribute contains a sufficiently large number of missing values (according to the user), this module also allows a user to completely discard an attribute column in a data set. Once all missing values in each respective data set have been substituted, the accident, vehicle, road and victim data sets are considered formatted.

6.2 Configuring data attributes

The primary function of Module 2.0 in the RARE framework of Figure 5.4 is to facilitate the structuring of records and attributes so that the data are in the appropriate configuration required by the RARE framework. This is achieved by means of three sub-modules, as illustrated in Figure 6.2. In Sub-module 2.1, multiple annual road records are consolidated into a single representative road record. Sub-module 2.2 is used to reconcile the accident, vehicle and victim records following the possible deletion of records in Sub-process 1.2. Thereafter, the accident data are passed to Sub-process 2.3 so as to determine which accident attributes describe inherent characteristics of the road. After having carried out this procedure, a data set is deemed appropriately configured and stored in a separate table. These three sub-modules are now described in greater detail.

In Sub-module 2.1, each of the m_R road records in $\acute{\mathcal{R}} = \{\{\acute{\boldsymbol{r}}_1^{(1)}, \ldots, \acute{\boldsymbol{r}}_Y^{(1)}\}, \ldots, \{\acute{\boldsymbol{r}}_1^{(m_R)}, \ldots, \acute{\boldsymbol{r}}_Y^{(m_R)}\}\}$ gathered over Y years must be summarised into a single representative record $\acute{\boldsymbol{r}}^{(i)} = \langle \acute{r}_1^{(i)} \cdots \acute{r}_{n_R}^{(i)} \rangle$ for all $i \in \{1, \ldots, m_R\}$. This means that each of the attribute values in $\acute{\boldsymbol{r}}^{(i)}$ must be summarised according to some specified procedure. In the case of the compulsory attribute $\hat{R}_T$, a specialised procedure is performed. Given that the AADT along the kth road is $r_T^{(k)}(y)$ during year y, then

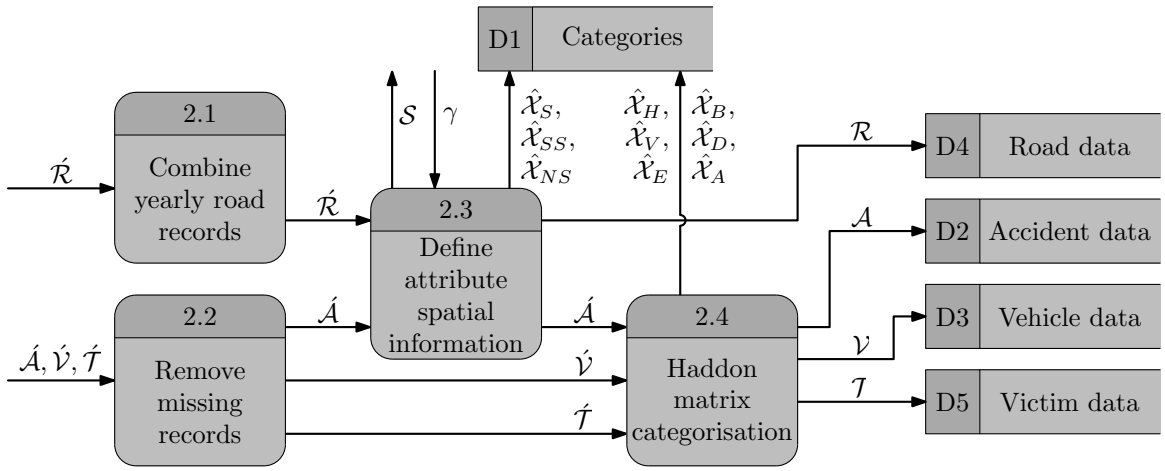


FIGURE 6.2: *A child diagram of Module 2.0 showing in greater detail the flow of data between sub-modules.*

the total number of vehicles that have passed along this road over the Y years may simply be estimated as $r_T^{(k)} = 365 \cdot \sum_{y=1}^{Y} r_T^{(k)}(y)$. For all other attributes, the procedural choice is left to the discretion of the user.

During the imputation phase of Sub-module 1.2, there is a possibility that the user decides to completely remove specific accident, vehicle or victim records containing missing data. If this is the case, then any associated records should also be removed in Sub-module 2.2 according to the following procedure:

1. If an accident record was removed in Sub-module 1.2, remove the corresponding vehicle and victim record(s).

2. If one or more vehicle record(s) was removed in Sub-module 1.2, remove the associated vehicle record(s) and corresponding accident record.

3. If one or more victim record(s) was removed in Sub-module 1.2, remove the associated victim record(s) and corresponding accident record.

4. If an accident record was removed in Step 2 or 3, remove the corresponding vehicle and victim record(s).

These four steps thus ensure that every accident record in $\mathcal{A}$ is associated with all the vehicle and victim records in $\mathcal{V}$ and $\mathcal{T}$, respectively, which formed part of the accident, and that every vehicle and victim record in $\mathcal{V}$ and $\mathcal{T}$, respectively, has a corresponding accident record in $\mathcal{A}$.

In the GIS literature, it is imperative to discern between spatial and non-spatial attributes. As described in §3.1.1, spatial attributes are used to construct one or more geometric elements that define the overall geometric representation of an object. Any attribute that cannot be used to construct an object's geometry is deemed non-spatial. It may, however, be unwise to classify *all* data attributes, which are deemed by the user as *not* spatial, as being exclusively non-spatial attributes. Such a categorisation inevitably asserts that non-spatial attributes possess *no* information in respect of an accident's location which might, in actuality, be false. It is argued that some attributes may in fact contain so-called *spatial information*. Identifying this class of attributes may allow a user to describe similarities in respect of surrounding area and road characteristics similarities which may lead to a certain type of RA. Module 2.3 is, therefore,

proposed to facilitate the categorisation of spatial and non-spatial attributes by the user, together with identifying attributes which contain spatial information, hereafter referred to as *semi-spatial* attributes.

During the first stage of attribute categorisation, accidents and roads are conceptually modelled as distinct objects according to the object-based modelling approach. Both accidents and roads are, therefore, assumed to have a spatial representation and comprise a set of attributes. Accident objects are each assumed to be spatially represented by a single, two-dimensional Cartesian point. Note that the vehicles and victims corresponding to a specific accident object are also encapsulated by this Cartesian point. It is, therefore, assumed that the set of accident spatial attributes $\{\hat{\mathcal{A}} \in \hat{\mathcal{X}}_S\}$ contains attributes defining the exact position of an accident on the earth's surface (*i.e.* a "latitude" and "longitude" attribute). Roads, on the other hand, are each assumed to be represented spatially by a single line segment embedded in a two-dimensional Cartesian plane which collectively makes up a road network. It is, therefore, also assumed that the set of road spatial attributes $\{\hat{\mathcal{R}} \in \hat{\mathcal{X}}_S\}$ contains attributes defining the start and end point of a single road line segment on the earth's surface, denoted by $(\hat{R}_{x_1}, \hat{R}_{y_1})$ and $(\hat{R}_{x_2}, \hat{R}_{y_2})$, respectively. As discussed in §3.1.1, an object's attributes may further be classified as either *spatial* if an attribute directly describes the spatial representation defining the location of the object, or, in the case where the attribute does not represent location-specific information, it is deemed *non-spatial*. In Sub-module 2.1, the user is, therefore, expected to categorise each of the attributes of the accident and road data sets as being spatial or non-spatial, in which case the attribute is appended to the set $\hat{\mathcal{X}}_S$ or the set $\hat{\mathcal{X}}_{NS}$, as appropriate. The attributes in the vehicle and victim data are assumed, by definition, to be non-spatial as they would always describe information related to circumstances inherent to the accident. A summary of all the compulsory attributes which must be defined by the user in Modules 1.0 and 2.0 is shown in Table 6.1.

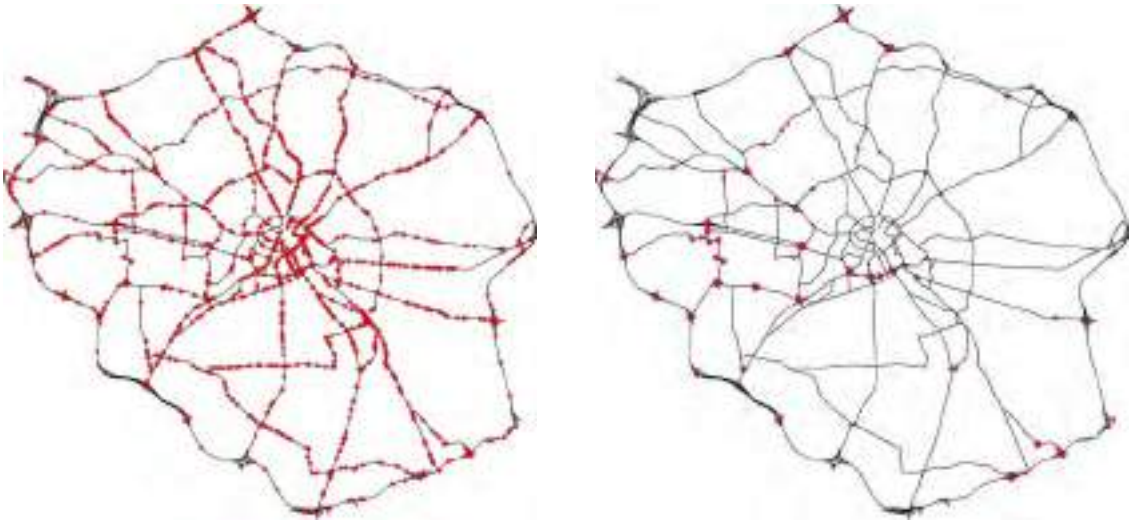
TABLE 6.1: *The set of compulsory attributes to be defined by the user.*

Attribute name	Symbol
Unique identifiers	$\hat{A}_I, \hat{V}_I, \hat{T}_I$
Link attributes	$\hat{V}_L, \hat{T}_L$
Road AADT	$\hat{R}_T$
Severity attribute	$\hat{T}_S$
Junction attribute	$\hat{A}_J$
Accident point	$(\hat{A}_x, \hat{A}_y)$
Road segment	$(\hat{R}_{x_1}, \hat{R}_{y_1}), (\hat{R}_{x_2}, \hat{R}_{y_2})$

The claim that non-spatial attributes may possess spatial-information is now clarified. This notion may best be elucidated by means of an example in which a set of m_A accidents $\mathcal{A} = \{\boldsymbol{a}^{(1)}, \ldots, \boldsymbol{a}^{(m_A)}\}$, the positions of which are assumed to be generated randomly according to a binomial point process over a network $\mathcal{N}$. Accidents are assumed to have two categorical attributes (*i.e.* $\boldsymbol{a}^{(i)} = \langle a_1^{(i)}\, a_2^{(i)} \rangle$) which have initially been labelled as non-spatial by the user. The first of these is an attribute capturing the day of week during which an accident occurred, denoted by $\hat{A}_1$, which has seven distinct categories, namely $a_1^{(i)} \in \{$`Monday`, `Tuesday`, `Wednesday`, `Thursday`, `Friday`, `Saturday`, `Sunday`$\}$ for all $i \in \{1, \ldots, m_A\}$. The second attribute, denoted by $\hat{A}_2 = \hat{A}_J$, describes the type of junction at which an accident took place. This attribute has four distinct categories, namely $a_2^{(i)} \in \{$`Roundabout`, `Crossroads`, `T-Junction`, `NotAtJunction`$\}$ for all $i \in \{1, \ldots, m_A\}$. Note that this notion relies on the fact that the attributes under

consideration are inherently categorical in nature (*i.e.* only the attributes in $\hat{\mathcal{X}}_{\text{cat}}$ are considered in this formulation).

Consider a scenario labelled A in which only the subset of accidents occurring on a Monday are considered, denoted by $\mathcal{A}\{\hat{A}_1 = \texttt{Monday}\}$ and illustrated in Figure 6.3(a). Contrast this with the scenario labelled B, illustrated in Figure 6.3(b), in which only the subset of accidents occurring at a roundabout, denoted by $\mathcal{A}\{\hat{A}_2 = \texttt{Roundabout}\}$, are shown. Obviously, given the subset of accidents in Figure 6.3(b), one would not expect the CSR hypothesis to hold since accidents appear to be significantly clustered at specific points in the road network. It may, on the other hand, be reasonable to assume that the subset of points in Figure 6.3(a) satisfy the CSR hypothesis since accidents appear to be evenly distributed within the road network. One may, as a result, argue that category $\hat{A}_2^{(1)}$ possesses *more* implicit spatial information than category $\hat{A}_1^{(1)}$ since the fact of *knowing* that an accident occurred at a roundabout decreases the possible set of accident locations within the road network. It may, therefore, be justified not only to think of attributes as being exclusively non-spatial, but rather, semi-spatial in nature, where a semi-spatial attribute is somewhere in-between a spatial and a non-spatial attribute.



(a) (b)

FIGURE 6.3: *A road network showing the (a) the locations of accidents occurring on a Monday and (b) the locations of accidents occurring at a roundabout.*

Recall from §3.2.3 that, if points are *truly* identically and independently generated on the network $\mathcal{N}$, then the randomness exhibited by this binomial point process must satisfy CSR hypothesis. Furthermore, recall that one of the methods for testing the CSR hypothesis is the network NND method of §3.3.1 which tests whether the shortest-path distance from each point to its next nearest point in the network space is statistically short (or long). According to this method, a statistical index G is calculated for an observed distribution of points $\mathcal{P}$ together with a pair of upper and lower critical limits G^+ and G^- at a particular significance level of α. As previously stated, if $G \in [G^-, G^+]$, there is not sufficient evidence to reject the CSR hypothesis and one may conclude, at a confidence level of $1 - \alpha$, that the points in $\mathcal{P}$ have indeed been generated according to a binomial point process within the network. If, on the other hand, $G < G^-$ or $G > G^+$, then there is sufficient evidence to reject the CSR hypothesis and one may conclude that the average NND between the points in $\mathcal{P}$ is significantly shorter (or longer) than expected. In particular, if $G < G^-$, one can conclude that points are indeed significantly clustered within the network space. This fact is now used to construct a measure of spatial information.

Consider again the examples of Figure 6.3. If one were to test the CSR hypothesis considering only the subset of accidents occurring at a roundabout, as illustrated in Figure 6.3(b), one would surely expect $G_{(b)} \ll G^{-}_{(b)}$. This notion is indeed correct — the application of the network NND method at a significance level of $\alpha = 0.05$ for Scenario B yields the values $G_{(b)} = 39.48$ and $G^{-}_{(b)} = 255.76$. Contrast this with the case where only the subset of accidents occurring on a Monday are considered — again at a significance level of $\alpha = 0.05$, application of the network NND method for Scenario A yields the values of $G_{(a)} = 104.20$ and $G^{-}_{(a)} = 147.37$[1]. In both these cases, one can comfortably reject the CSR hypothesis at a 95% confidence since $G_{(a)} < G^{-}_{(a)}$ and $G_{(b)} < G^{-}_{(b)}$. One may, however, argue that since $G^{-}_{(b)} - G_{(b)} = 216.28 \gg 43.17 = G^{-}_{(a)} - G_{(a)}$, the category of accidents that occurred at a roundabout holds *more* spatial information than the category of accidents that occurred on a Monday.

As described in §3.3.1, the results of Scenarios A and B can also be viewed in a disaggregated measure in the form of an observed and expected G-function, as shown in Figures 6.4(a) and (b), respectively. These two plots further support the argument that the accidents in Scenario B are significantly more clustered than the accidents in Scenario A. In Scenario B, it is observed that approximately 98% of all accidents occur within a distance of 100 metres from its next nearest accident, illustrated by the blue curve in Figure 6.4(b). This conclusion corroborates what can be observed in Figure 6.3. Contrast this with Scenario A in which approximately only 82% of all accidents occur within a distance of 100m from one another. Notice that if the observed NND between pairs of accidents is significantly small (like that in Scenario B), one would be more inclined to believe that the category under consideration captures information describing location-specific information. Conversely, if the observed NND between pairs of accidents is closer to the expected mean distribution of NNDs (like that in Scenario A), then one would be more inclined to believe that the category under consideration does *not* capture any location-specific information.

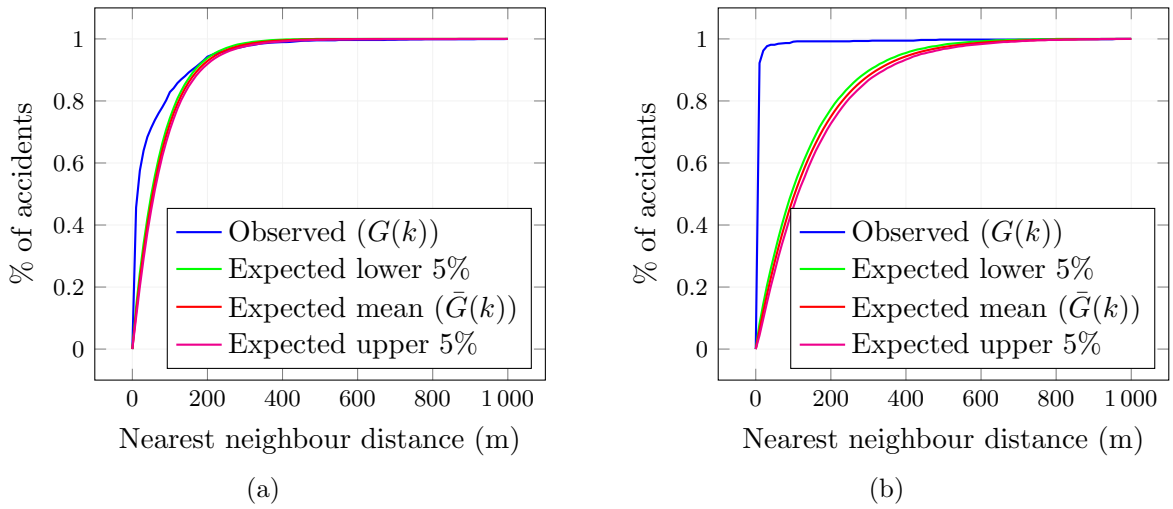


FIGURE 6.4: *The observed and expected network G-function corresponding to (a) the accidents in Figure 6.3(a) occurring on a Monday and (b) the accidents in Figure 6.3(b) occurring at a roundabout.*

Leveraging this fact, it is proposed that, for each categorical attribute, a so-called *spatial score* be constructed to represent a quantitative measure of the total amount of spatial information

[1]These results were obtained using the SANET Standalone toolbox [191].

possessed by an attribute. For an arbitrary data set $\mathcal{X}$ containing m_X records (which may be represented by a set of points $\mathcal{P}$ on a network $\mathcal{N}$) and n_X categorical attributes, a *spatial category score* $s_j^{(c)}$ is first calculated for the cth category in attribute j as

$$s_c^{(j)} = \frac{\int_{k=0}^{\infty} \left(G_j^{(c)}(k) - \overline{G}_j^{(c)}(k) \right) \, \mathrm{d}k}{\int_{k=0}^{\infty} \left(1 - \overline{G}_j^{(c)}(k) \right) \, \mathrm{d}k}. \tag{6.1}$$

The numerator $\int_{k=0}^{\infty} G_j^{(c)}(k) - \overline{G}_j^{(c)}(k) \, \mathrm{d}k$ in (6.1) is the difference between the observed and expected G-functions. Notice that if the observed G-function $G_j^{(c)}(k)$, for the most part, coincides with the expected G-function $\overline{G}_j^{(c)}(k)$, the quantity $\int_{k=0}^{\infty} G_j^{(c)}(k) - \overline{G}_j^{(c)}(k) \, \mathrm{d}k \approx 0$ and hence $s_c^{(j)} \approx 0$. In this case, a value of $s_c^{(j)} \approx 0$ indicates that the cth category possesses almost no spatial information. If, on the other hand, $G_j^{(c)}(k) \approx 1$ for all $k \in \mathbb{R}_{\geq 0}$, then $s_c^{(j)} \approx 1$, indicating that the cth category possesses a significant amount of spatial information. In the case of Scenarios A and B, for example, the categories $A_1^{(1)}$ and $A_2^{(1)}$ result in category spatial scores of $s_1^{(1)} = 0.321$ and $s_1^{(2)} = 0.888$, respectively, where this result should be interpreted as follows: The description that an accident occurred at a roundabout contains *more* spatial information than the description of an accident occurring on a Monday (since $s_1^{(2)} > s_1^{(1)}$). This result, therefore, demonstrates the intended working of the expression in (6.1) — larger values of $s_c^{(j)}$ should indicate larger amounts of spatial information. This spatial category score is calculated for each of the $C_j = |\{x_j^{(i)}\}_{i \in \{1,\ldots,m_X\}}|$ categories of attribute j. An overall *spatial attribute score* S_j may then calculated as

$$S_j = \sum_{c=1}^{C_j} \frac{|\mathcal{X}\{\hat{X}_j = \hat{X}_j^{(c)}\}|}{m_X} s_c^{(j)}, \tag{6.2}$$

where the quantity $|\mathcal{X}\{\hat{X}_j = \hat{X}_j^{(c)}\}|$ is the number of accident records for which the jth attribute $\hat{X}_j$ is equal to its cth category $\hat{X}_j^{(c)}$. The fraction in (6.2), therefore, weights each spatial category score $s_c^{(j)}$ by the proportion of the m_X records containing the value of $\hat{X}_j^{(c)}$ for attribute $\hat{X}_j$. In this way, the spatial category scores of categories containing more records are weighted heavier than those with fewer records. The amount of spatial information can be calculated for all $j \in \{1, \ldots, n_X\}$ categorical attributes in $\hat{\mathcal{X}}$.

The purpose of Sub-module 2.3 is, therefore, also to calculate the spatial attribute scores following the aforementioned procedure for each of the categorical attributes $\{\hat{\mathcal{A}} \in \mathcal{X}_{\text{cat}}\}$ in the set of accident data. The inputs to this sub-module are the road network $\mathcal{N} = (\mathcal{L}, \mathcal{E})$, defined by the set of road spatial attributes $\mathcal{R}\{\hat{\mathcal{R}} \in \hat{\mathcal{X}}_S\}$, together with the set of point locations $\mathcal{P} = \{p^{(1)}, \ldots, p^{(m_A)}\}$ of the m_A accident records in $\tilde{\mathcal{A}}$. The output is the set of spatial attribute scores $\mathcal{S} = \{S_1, \ldots, S_{n_A}\}$ corresponding to the set of categorical attributes.

The onus now shifts to the user to decide what characterises a semi-spatial attribute. This is achieved by presenting the set of spatial attribute scores $\mathcal{S}$ to the user and requiring him or her to define a *spatial threshold* value, denoted by $\gamma \in [0, 1]$, for which the attributes currently residing in $\hat{\mathcal{X}}'_{NS}$ are classified as being either spatial or non-spatial. If the jth attribute $\hat{X}_j$ exceeds this threshold, this attribute is appended to the set of semi-spatial attributes, denoted by $\hat{\mathcal{X}}_{SS}$ (*i.e.* $\hat{\mathcal{X}}_{SS} = \{\hat{A}_j \mid S_j > \gamma, j \in \{1, \ldots, n_A\}\}$). Any attribute that does not exceed this threshold remains in the set of non-spatial attributes (*i.e.* $\hat{\mathcal{X}}_{NS} = \hat{\mathcal{X}}_{NS} \backslash \hat{\mathcal{X}}_{SS}$). The new categories $\hat{\mathcal{X}}_S$, $\hat{\mathcal{X}}_{SS}$ and $\hat{\mathcal{X}}_{NS}$ are then subsequently stored in Datastore D1.

During the next categorisation step, Sub-module 2.4 requires the user to categorise the set of all non-spatial accident, vehicle, victim attributes, currently stored in $\hat{\mathcal{X}}_{NS}$, according to the Haddon matrix, as described in §2.2.3. The user must, therefore, closely examine the meta-data associated with each attribute in these data sets and determine the information context described by an attribute. Given an arbitrary attribute, the first question the user should answer is whether the attribute describes the conditions *before*, *during* or *after* the accident took place, in which case, the attribute is appended to the set $\hat{\mathcal{X}}_B$, $\hat{\mathcal{X}}_D$ or $\hat{\mathcal{X}}_A$, respectively. The second question the user should ask is whether the attribute describes a *human*, *vehicle* or *environmental* factor which may have contributed to this accident, in which case, the attribute is appended to the set $\hat{\mathcal{X}}_H$, $\hat{\mathcal{X}}_V$ or $\hat{\mathcal{X}}_E$, as appropriate. Consequently, an attribute may be further categorised into one of nine mutually exclusive subsets, with the notations of these various subsets denoted as in the light-grey region of Table 6.2. These categories are then subsequently stored in Datastore D1. An accident attribute may, for example, describe the weather condition on the day of the accident, in which case a user would likely conclude that this attribute is an *environmental* factor which existed *before* the accident took place. This attribute would, as a result, reside in the set $\hat{\mathcal{X}}_{BE} = \hat{\mathcal{X}}_B \cap \hat{\mathcal{X}}_E$. This attribute categorisation, therefore, provides a form of attribute-specific information which can act as a mechanism for discernment between data attribute subsets, depending on the time instant and factor they describe. During a risk assessment, for example, the user may only wish to know the *environmental* factors (*i.e.* the attributes residing in the set $\hat{\mathcal{X}}_E$) that contribute to the so-called "riskiness" of an accident. Thereafter, the accident, vehicle, road and victim data sets are each stored in designated tables, denoted by Datastores D2, D3, D4 and D5, respectively.

TABLE 6.2: *The various attribute subsets created when categorising attributes according to the Haddon matrix.*

	Human (H)	Vehicle (V)	Environment (E)	$\hat{\mathcal{X}}_H \cup \hat{\mathcal{X}}_V \cup \hat{\mathcal{X}}_E$
Before accident (B)	$\hat{\mathcal{X}}_{BH}$	$\hat{\mathcal{X}}_{BV}$	$\hat{\mathcal{X}}_{BE}$	$\hat{\mathcal{X}}_B$
During accident (D)	$\hat{\mathcal{X}}_{DH}$	$\hat{\mathcal{X}}_{DV}$	$\hat{\mathcal{X}}_{DH}$	$\hat{\mathcal{X}}_D$
After accident (A)	$\hat{\mathcal{X}}_{AH}$	$\hat{\mathcal{X}}_{AV}$	$\hat{\mathcal{X}}_{AH}$	$\hat{\mathcal{X}}_A$
$\hat{\mathcal{X}}_B \cup \hat{\mathcal{X}}_D \cup \hat{\mathcal{X}}_A$	$\hat{\mathcal{X}}_H$	$\hat{\mathcal{X}}_V$	$\hat{\mathcal{X}}_E$	$\hat{\mathcal{X}}_{NS}$

6.3 Combining vehicle records to accident records

The primary function of Module 3.0 in the RARE framework of Figure 5.4 is to facilitate the merging of a vehicle record in $\mathcal{V}$ with a corresponding accident record in $\mathcal{A}$ so that one may analyse the attributes in these two data sets concurrently. This is achieved by means of two sub-modules, as illustrated in Figure 6.5. In Sub-module 3.1, the vehicle record(s) corresponding to each individual accident are ordered such that the integrity of vehicle records is maintained during the merging procedure. Ordered vehicle records corresponding to a single accident are then condensed in Sub-module 3.2 into one representative record by means of a comma-delimited approach, which is then appended to its corresponding accident record. These two sub-modules are now described in greater detail.

Given a set of accidents $\mathcal{A}$, let $\boldsymbol{A} = \langle \boldsymbol{a}^{(1)}, \ldots, \boldsymbol{a}^{(m_A)} \rangle^T$ denote a matrix of m_A accident records corresponding to the accidents in $\mathcal{A}$, where the ith accident record $\boldsymbol{a}^{(i)} = \langle a_1^{(i)} \cdots a_{n_A}^{(i)} \rangle$ contains

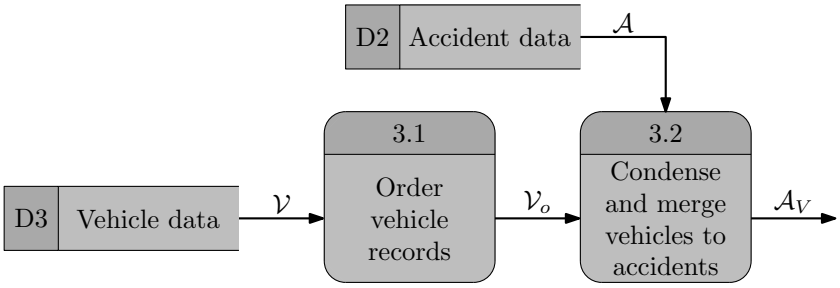


FIGURE 6.5: *A child diagram of Module 3.0 showing in greater detail the flow of data between sub-modules.*

the values of n_A attributes of the accident. The matrix

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{a}^{(1)} \\ \vdots \\ \boldsymbol{a}^{(i)} \\ \vdots \\ \boldsymbol{a}^{(m_A)} \end{bmatrix} = \begin{bmatrix} a_1^{(1)} & \cdots & a_j^{(1)} & \cdots & a_{n_A}^{(1)} \\ \vdots & & \vdots & & \vdots \\ a_1^{(i)} & \cdots & a_j^{(i)} & \cdots & a_{n_A}^{(i)} \\ \vdots & & \vdots & & \vdots \\ a_1^{(m_A)} & \cdots & a_j^{(m_A)} & \cdots & a_{n_A}^{(m_A)} \end{bmatrix}$$

is, therefore, a mathematical representation of a table data structure used in a relational database, as discussed in §3.1.2. During an RA investigation, additional non-spatial attributes are also typically recorded with respect to each of the vehicles involved in the accident. Let $\boldsymbol{V}^{(i)} = \langle \boldsymbol{v}^{(i,1)}, \ldots, \boldsymbol{v}^{(i,K_i)} \rangle$, therefore, denote a matrix of $k_i \in \mathbb{N}$ vehicle records involved in the ith accident, where a vehicle record $\boldsymbol{v}^{(i,j)} = \langle v_1^{(i,j)}, \ldots, v_{n_V}^{(i,j)} \rangle$ contains values for the n_V attributes associated with the jth vehicle involved in the accident. The matrix $\boldsymbol{V}^{(i)}$ may, therefore, be expressed as

$$\boldsymbol{V}^{(i)} = \begin{bmatrix} \boldsymbol{v}^{(i,1)} \\ \vdots \\ \boldsymbol{v}^{(i,j)} \\ \vdots \\ \boldsymbol{v}^{(i,m_V)} \end{bmatrix} = \begin{bmatrix} v_1^{(i,1)} & \cdots & v_j^{(i,1)} & \cdots & v_{n_V}^{(i,1)} \\ \vdots & & \vdots & & \vdots \\ v_1^{(i,k)} & \cdots & v_j^{(i,k)} & \cdots & v_{n_V}^{(i,k)} \\ \vdots & & \vdots & & \vdots \\ v_1^{(i,K_i)} & \cdots & v_j^{(i,K_i)} & \cdots & v_{n_V}^{(i,K_i)} \end{bmatrix}.$$

An accident record $\boldsymbol{a}^{(i)}$ thus exhibits a one-to-many relationship in that it may be linked to at least one vehicle record in the set $\{\boldsymbol{v}^{(i,1)}, \ldots, \boldsymbol{v}^{(i,m_V)}\}$.

During a statistical or ML investigation of RA data, it would be beneficial if one could concurrently analyse the accident attributes $\boldsymbol{A}$ and corresponding vehicle attributes $\boldsymbol{V}^{(i)}$ for all $i \in \{1, \ldots, m_A\}$ (as opposed to analysing them separately). Such an endeavour would most certainly increase the information gained during an investigation since the attributes of an accident may be analysed in conjunction with the attributes of the vehicles involved in said accident. The problem is, however, that there is often more than one vehicle record associated with a single accident record. It is, therefore, necessary to devise a procedure whereby the $m_A \times n_A$ matrix of accident records $\boldsymbol{A}$, and all the corresponding $k_i \times n_V$ matrices of vehicle records $\boldsymbol{V}^{(i)}$ for all

$i \in \{1, \ldots, m_A\}$ may be combined into a single $m_A \times (n_A + n_V)$ accident-vehicle matrix

$$\boldsymbol{A}_V = \begin{bmatrix} \boldsymbol{a}_V^{(1)} \\ \vdots \\ \boldsymbol{a}_V^{(i)} \\ \vdots \\ \boldsymbol{a}_V^{(m_A)} \end{bmatrix},$$

where $\boldsymbol{A}_V^{(i)}$ is a vector containing all information associated with the ith accident record $\boldsymbol{a}^{(i)}$ and corresponding vehicle records in $\boldsymbol{V}^{(i)}$.

The question still remains as to how to construct this vector $\boldsymbol{a}_V^{(i)}$. If only a single vehicle record corresponds to the ith accident record $\boldsymbol{a}^{(i)}$ (*i.e.* $k_i = |\boldsymbol{V}^{(i)}| = 1$), $\boldsymbol{a}_V^{(i)}$ would simply be the concatenation of the vectors $\boldsymbol{a}^{(i)}$ and $\boldsymbol{v}^{(i,1)}$, expressed mathematically as $\boldsymbol{a}^{(i)\frown}\boldsymbol{v}^{(i,1)}$. If, however, $k_i > 1$ vehicles are involved in the ith accident, an alternative procedure is required. First, the set of k_i vehicle records must be transformed into a single representative vehicle record $\boldsymbol{v}^{(i)}$ in order to perform this concatenation. This may be achieved by adopting a comma-delimited approach whereby, for a specific vehicle attribute $t \in \{1, \ldots, n_V\}$, the values associated with each of the k_i vehicles in $\boldsymbol{V}^{(i)}$ are combined into a single representative value $v_t^{(i)} = (v_t^{(i,1)}, \ldots, v_t^{(i,k_i)})$.

An implication of following a comma-delimited approach is that, during an investigation of an RA, the data recorder decides which of the k_i vehicle records in $\boldsymbol{V}^{(i)}$ to label $1, \ldots, k_i - 1$ and k_i, respectively. In reality, there is typically no convention when making this decision and this choice is purely at the discretion of the data recorder. Consequently, the rows of $\boldsymbol{V}^{(i)}$ are inherently unordered and may lead to situations where two k-vehicle accidents, labelled a and b, with identically corresponding values for vehicle attribute V_t (*i.e.* $v_t^{(a,i)}$ can be matched to an equivalent $v_t^{(b,j)}$ and *vice versa* for $i, j \in \{1, \ldots, k\}$), do not both share the same values of $v_t^{(i)}$. As a means of illustration, consider the case where the accidents a and b contain $k_a = k_b = 2$ vehicles each, and vehicle attribute V_t describes the type of vehicle involved in the accident. In this case, there may be a situation in which $v_t^{(a,i)} = (\text{car, truck})$ and $v_t^{(b,i)} = (\text{truck, car}) \neq v_t^{(a,i)}$. The consequence of this is that, during any statistical or ML investigation, no inference will be drawn between accidents a and b with respect to vehicle attribute V_t, as this information will simply be lost. It is therefore necessary to devise a procedure in Sub-module 3.1 whereby the rows of $\boldsymbol{V}^{(i)}$ are ordered in such a way that this type of situation is avoided while still preserving the integrity of a vehicle record.

A pseudocode description of such a procedure may be found in Algorithm 6.1 and makes use of recursion. The notations $\boldsymbol{V}_o^{(i)}$ and $\boldsymbol{V}_u^{(i)}$ are introduced to distinguish between *ordered* and *unordered* vehicle records, respectively. The first input to this algorithm is an unordered $k_i \times n_V$ matrix $\boldsymbol{V}_u^{(i)}$ of vehicle records involved in the ith accident record $\boldsymbol{a}^{(i)}$. The second input is an array $\boldsymbol{\rho}$ stipulating the order (or precedence) in which the n_V vehicle attributes should be used to order the k_i rows in $\boldsymbol{V}_u^{(i)}$, where $|\boldsymbol{\rho}| = n_V$. The construction of this array is somewhat subjective and must be supplied by the platform user. For illustration purposes, consider a $k_i \times 3$ matrix $\boldsymbol{V}_u^{(i)}$ of unordered vehicle records. A vector $\boldsymbol{\rho} = \langle 3\,1\,2 \rangle$ may be specified by the user, requiring the following: The 3rd column in $\boldsymbol{V}_u^{(i)}$ must be considered first, the 1st column in $\boldsymbol{V}_u^{(i)}$ must be considered second, and the 2nd column in $\boldsymbol{V}_u^{(i)}$ must be considered last during the ordering process. The last input is a step counter s which is used to successively step through the vector $\boldsymbol{\rho}$ during the recursion process. The output of the algorithm is an ordered $k_i \times n_V$ matrix $\boldsymbol{V}_o^{(i)}$ of vehicle records.

Algorithm 6.1: **Order vehicles**

Input : An unordered $k \times n_v$ matrix $\boldsymbol{V}_u^{(i)}$ of vehicle records, a vector $\boldsymbol{\rho}$ stipulating the precedence of the n_v vehicle attributes, and the current step s in the order $\boldsymbol{\rho}$.

Output: An ordered matrix $\boldsymbol{V}_o^{(i)}$ of vehicle records.

1 **if** $s > n_V$ **then**
2 $\quad \boldsymbol{V}_o^{(i)} \leftarrow \boldsymbol{V}_u^{(i)}$;
3 $\quad$ **return** $\boldsymbol{V}_o^{(i)}$;
4 **else if** all the values in the $\boldsymbol{\rho}[s]^{\text{th}}$ column of $\boldsymbol{V}_u^{(i)}$ are different **then**
5 $\quad \boldsymbol{V}_o^{(i)} \leftarrow$ sort the rows of $\boldsymbol{V}_u^{(i)}$ according to the $\boldsymbol{\rho}[s]^{\text{th}}$ column;
6 $\quad$ **return** $\boldsymbol{V}_o^{(i)}$;
7 **else**
8 $\quad \boldsymbol{V}_u^{(i)} \leftarrow$ sort the rows of $\boldsymbol{V}_u^{(i)}$ according to the values in the $\boldsymbol{\rho}[s]^{\text{th}}$ column;
9 $\quad j \leftarrow 1$;
10 $\quad$ **while** $j < k$ **do**
11 $\quad\quad$ **if** $c \geq 2$ consecutive values, starting from the jth value in the $\boldsymbol{\rho}[s]^{\text{th}}$ column of $\boldsymbol{V}_u^{(i)}$, are equal **then**
12 $\quad\quad\quad \boldsymbol{V}_t^{(i)} \leftarrow$ extract the c consecutive rows from $\boldsymbol{V}_u^{(i)}$;
13 $\quad\quad\quad \boldsymbol{V}_t^{(i)} \leftarrow$ `OrderVehicles`($\boldsymbol{V}_t^{(i)}$, $\boldsymbol{\rho}$, $s+1$);
14 $\quad\quad\quad$ Insert $\boldsymbol{V}_t^{(i)}$ back into $\boldsymbol{V}_u^{(i)}$ at the same position from whence it was removed;
15 $\quad\quad\quad j \leftarrow j + c$;
16 $\quad\quad$ **else**
17 $\quad\quad\quad j \leftarrow j + 1$;
18 $\quad \boldsymbol{V}_o^{(i)} \leftarrow \boldsymbol{V}_u^{(i)}$;
19 $\quad$ **return** $\boldsymbol{V}_o^{(i)}$;

In Step 1, the first test in the function is whether the step counter s exceeds the number of vehicle attributes n_V. This condition only occurs when all rows in $\boldsymbol{V}_u^{(i)}$ are identical. If this is the case, no row ordering is required and the matrix $\boldsymbol{V}_o^{(i)}$ is simply set equal to the matrix $\boldsymbol{V}_u^{(i)}$. If, however, this if-statement condition is not satisfied, the next if-condition in Step 4 tests whether all the values in the $\boldsymbol{\rho}[s]^{\text{th}}$ column of $\boldsymbol{V}_u^{(i)}$ are different. If this condition is true, the rows of $\boldsymbol{V}_u^{(i)}$ are sorted according to the $\boldsymbol{\rho}[s]^{\text{th}}$ column.

If neither of the subsequent if-statement conditions are true, it means that at least two values in the $\boldsymbol{\rho}[s]^{\text{th}}$ column of $\boldsymbol{V}_u^{(i)}$ are equal. If this is the case, the rows of $\boldsymbol{V}_u^{(i)}$ are initially ordered according to the values in the $\boldsymbol{\rho}[s]^{\text{th}}$ column. The next step is to iteratively step through the k_i vehicle records in $\boldsymbol{V}_u^{(i)}$. This is achieved by means of a while-statement condition in Step 6 which tests whether a counter j, which is initialised to 1, has reached the last vehicle record. If $c \geq 2$ consecutive values, starting from the jth value in the $\boldsymbol{\rho}[s]^{\text{th}}$ column of $\boldsymbol{V}_u^{(i)}$, are found, the rows containing these values are extracted from $\boldsymbol{V}_u^{(i)}$ and stored in a temporary matrix $\boldsymbol{V}_t^{(i)}$. Since the $\boldsymbol{\rho}[s]^{\text{th}}$ column cannot be used to order the rows of $\boldsymbol{V}_t^{(i)}$, the algorithm instead considers the $\boldsymbol{\rho}[s+1]^{\text{th}}$ column. This is achieved by passing the matrix $\boldsymbol{V}_t^{(i)}$, together with the precedence vector $\boldsymbol{\rho}$ and incrementing step counter $s+1$, to the same `OrderVehicles` function. The temporary matrix $\boldsymbol{V}_t^{(i)}$ is then overwritten with the results of the function and inserted back

into $\boldsymbol{V}_u^{(i)}$ at the same position from whence it was removed. The j-counter is then incremented by the number of consecutively equal values found to skip over these values. If the if-statement condition in Step 11 is not satisfied, the j-counter is simply incremented by one in order to consider the next value in the $\boldsymbol{\rho}[s]^{\text{th}}$ column. Finally, the ordered matrix $\boldsymbol{V}_o^{(i)}$ of vehicle records is returned. The successive application of this algorithm can therefore be used to transform the set of vehicle records currently stored in the database into a set of ordered vehicle records, denoted by $\mathcal{V}_o$.

The purpose of Sub-module 3.2 is to condense the vehicle record(s) corresponding to an accident into a single representative vehicle record, and append this record to the corresponding accident record. Given the set $\mathcal{V}_o$ of ordered vehicle records, this single representative vehicle record $\boldsymbol{v}^{(i)}$ may be defined for the ith accident as

$$\boldsymbol{v}^{(i)} = [(v_1^{(i,1)}, \ldots, v_1^{(i,k_i)}), \ldots, (v_{n_V}^{(i,1)}, \ldots, v_{n_V}^{(i,k_i)})].$$

The merged row in the accident-vehicle matrix $\boldsymbol{A}_V$ would therefore be the concatenation of the vectors $\boldsymbol{a}^{(i)}$ and $\boldsymbol{v}^{(i)}$, expressed mathematically as

$$\boldsymbol{a}_V^{(i)} = \boldsymbol{a}^{(i)\frown}\boldsymbol{v}^{(i)} = [a_1^{(i)}, \ldots, a_{n_A}^{(i)}, (v_1^{(i,1)}, \ldots, v_1^{(i,k_i)}), \ldots, (v_{n_V}^{(i,1)}, \ldots, v_{n_V}^{(i,k_i)})].$$

In this way, the accident-vehicle matrix $\boldsymbol{A}_V$ may be constructed by successively merging the vectors $\boldsymbol{a}^{(i)}$ and $\boldsymbol{v}^{(i)}$ for all $i \in \{1, \ldots, m_A\}$. This intermediate result is then passed to Module 4.0.

6.4 Combining road records to accident-vehicle records

The primary function of Module 4.0 in the RARE framework of Figure 5.4 is to associate an accident-vehicle record in $\mathcal{A}_V$ with a corresponding road in $\mathcal{R}$. This is achieved by means of two sub-modules, as illustrated in Figure 6.6. The purpose of Sub-module 4.1 is to spatially associate each of the accident-vehicle records with a corresponding road record. The merging of accident record(s) with an associated road record is then accomplished in Sub-module 4.2. These two sub-modules are now described in greater detail.

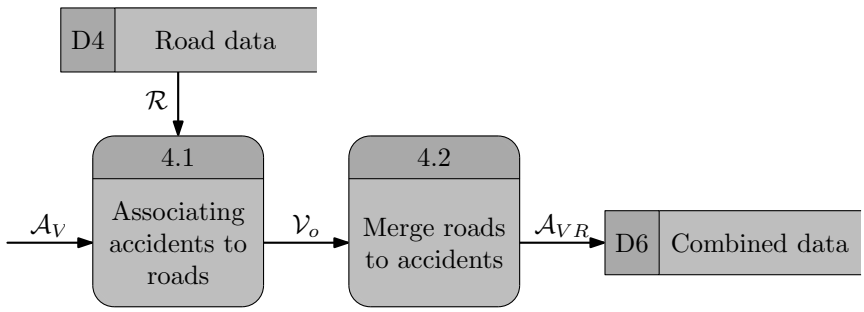


FIGURE 6.6: *A child diagram of Module 4.0 showing in greater detail the flow of data between sub-modules.*

For the set of road attributes $\mathcal{R}\{\hat{\mathcal{R}} \in \hat{\mathcal{X}}_{NS}\}$, let $\boldsymbol{R} = \langle \boldsymbol{r}^{(1)}, \ldots, \boldsymbol{r}^{(m_R)} \rangle^T$ be a matrix containing the m_R road records corresponding to the roads in $\mathcal{R}$, where the kth road record $\boldsymbol{r}^{(k)} = \langle r_1^{(k)}, \ldots, r_{n_R}^{(k)} \rangle$ contains the values of the n_R attributes of the road. The matrix $\boldsymbol{R}$ may, therefore, be expressed as

$$\boldsymbol{R} = \begin{bmatrix} \boldsymbol{r}^{(1)} \\ \vdots \\ \boldsymbol{r}^{(i)} \\ \vdots \\ \boldsymbol{r}^{(m_R)} \end{bmatrix} = \begin{bmatrix} r_1^{(1)} & \cdots & r_k^{(1)} & \cdots & r_{n_R}^{(1)} \\ \vdots & & \vdots & & \vdots \\ r_1^{(i)} & \cdots & r_k^{(i)} & \cdots & r_{n_R}^{(i)} \\ \vdots & & \vdots & & \vdots \\ r_1^{(m_R)} & \cdots & r_k^{(m_R)} & \cdots & r_{n_R}^{(m_R)} \end{bmatrix}.$$

Since there exists a many-to-one relationship between accidents and roads (*i.e.* an accident is assumed to take place on a single road segment), the aim of Sub-module 4.1 is to associate the accident-vehicle record $\boldsymbol{a}_V^{(i)}$ with the corresponding road record $\boldsymbol{r}^{(k)}$.

Unlike in the case of the accident and vehicle data, it is assumed that there is *no* defined association between the accident and road data sets (as these two data sets are typically constructed independently). In other words, no explicit link is assumed between an accident record and a road record. A strategy is therefore devised in Sub-module 4.1 to associate accident and road objects spatially by means of their respective geometric elements.

During the investigation of each of the RAs, it is the responsibility of the data recorder to record the latitude and longitude coordinates of the location at which the accident occurred. As previously stated, the ith accident may, therefore, be represented geometrically by a single two-dimensional Cartesian point $\boldsymbol{p}^{(i)} = (x_A, y_A)^{(i)}$, defined in $\mathcal{A}\{\hat{\mathcal{A}} \in \hat{\mathcal{X}}_S\}$, where the point locations of the m_A accidents in $\mathcal{A}$ are contained in the set $\mathcal{P} = \{\boldsymbol{p}^{(1)}, \ldots, \boldsymbol{p}^{(m_A)}\}$. As previously described, the attributes defining an accident point are denoted by the compulsory attributes $(\hat{A}_x, \hat{A}_y)$.

In the case of a road network, a vector data model is, in practice, typically used to construct its spatial representation. A road is typically approximated by a cartographer in the form of a *series* of connected straight line segments. In this way, the curvature of the road is captured at a certain level of abstraction assumed by the cartographer. In this study, however, a road object is spatially represented by a *single* line segment denoted by ℓ, where a line segment must satisfy the five geometric conditions proposed by Okabe and Sugihara [194], as described in §3.2.1. In reality, a road would, therefore, comprise multiple road objects in this study. The kth road record $\boldsymbol{r}^{(k)} = \langle r_1^{(k)}, \ldots, r_{n_R}^{(k)} \rangle$ corresponds to its spatial representation assumed by road segment $\ell^{(k)}$.

The set of all road line segments is assumed to form part of a connected network $\mathcal{N} = (\mathcal{L}, \mathcal{E})$, like that presented in Chapter 3. The kth road may, therefore, be represented geometrically by a set of points $\ell^{(k)} = \{(x_R^{(1)}, y_R^{(1)})^{(k)} + u(x_R^{(2)}, y_R^{(2)})^{(k)} \mid u \in [0,1]\}$, where the vector pair $(\boldsymbol{e}_1^{(k)}, \boldsymbol{e}_2^{(k)}) = ((x_R^{(1)}, y_R^{(1)})^{(k)}, (x_R^{(2)}, y_R^{(2)})^{(k)})$ denotes the two endpoints of the line segment in two-dimensional Cartesian space, defined in $\mathcal{R}\{\hat{\mathcal{R}} \in \hat{\mathcal{X}}_S\}$. The attributes defining the road segment start and end points are denoted by $(\hat{R}_{x_1}, \hat{R}_{y_1})$ and $(\hat{R}_{x_2}, \hat{R}_{y_2})$, respectively. All line segment representations corresponding to the m_R roads in $\mathcal{R}$ are contained in the set $\mathcal{L} = \{\ell^{(1)}, \ldots, \ell^{(m_R)}\}$.

In an ideal situation, one would require all the points in $\mathcal{P}$ to coincide with some line segment in $\mathcal{L}$ (*i.e.* $(x_A, y_A)^{(i)} \in \ell^{(1)} \cup \ldots \cup \ell^{(m_R)}$ for all $i \in \{1, \ldots, m_A\}$). If this were the case, one could simply associate an accident object with the coinciding road object. In reality, however, this is often not the case for any single point in $\mathcal{P}$. This is simply due to inaccuracies associated with how the data are captured and represented. A geoprocessing procedure must, therefore, be designed to project each of the points in $\mathcal{P}$ onto the closest line segment in $\mathcal{L}$ so that this relationship can be established.

In more formal terms, given a point $\boldsymbol{p}^{(i)}$, one must determine

$$\min_{k \in \{1,\ldots,m_R\}} d_{\perp}(\boldsymbol{p}^{(i)}, \ell^{(k)}),$$

where $d_{\perp}(\boldsymbol{p}^{(i)}, \ell^{(k)})$ is the perpendicular distance between the point $\boldsymbol{p}^{(i)}$ and some point strictly on the line segment $\ell^{(k)}$. It is, however, obvious that, for all point and line segment combinations, $d_{\perp}(\boldsymbol{p}^{(i)}, \ell^{(k)})$ is not always defined. Consider the point $\boldsymbol{p}^{(i+1)}$ and line segment $\ell^{(k)}$ in Figure 6.7, for example. In this case, there are no points on $\ell^{(k)}$ that, when joined to $\boldsymbol{p}^{(i+1)}$, are perpendicular to $\ell^{(k)}$. The only circumstance where this is possible is if a point $\boldsymbol{p}^{(i)}$ is contained within the shaded region $\mathcal{D}^{(k)}$ defined by two lines which are perpendicular to $\ell^{(k)}$ and pass through its endpoints.

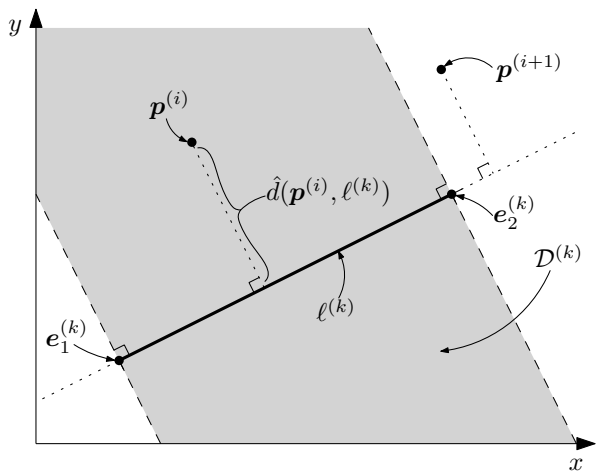
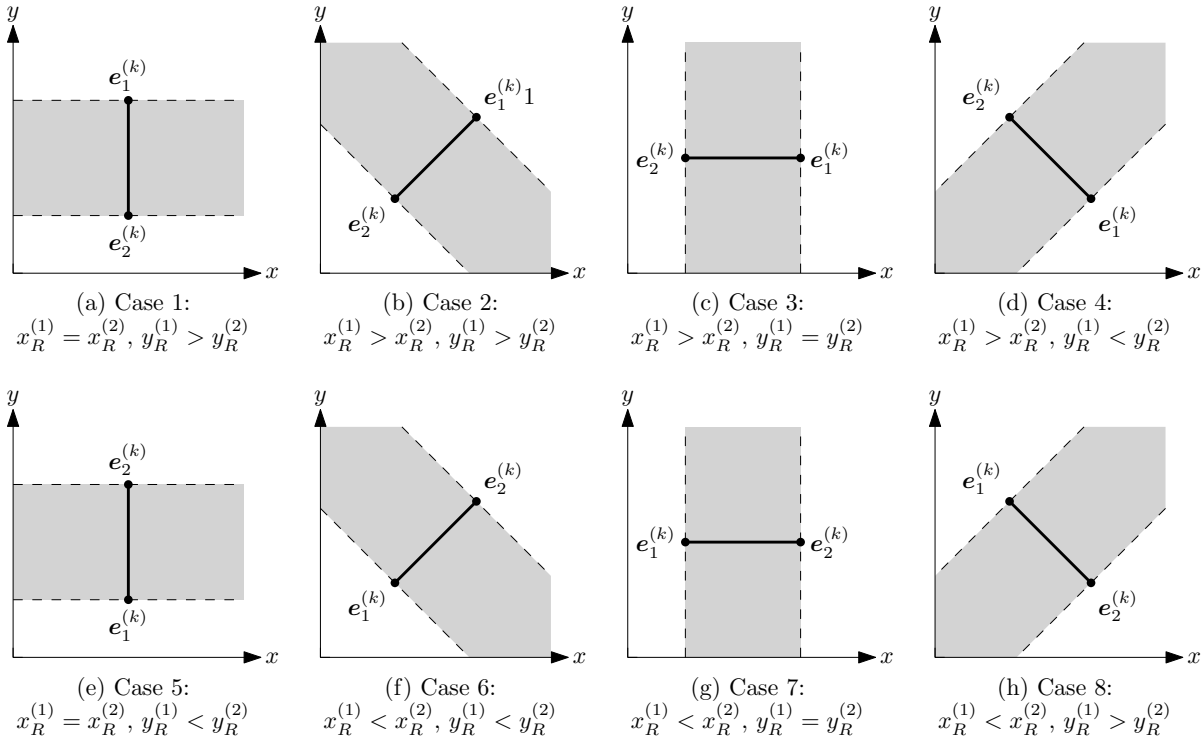


FIGURE 6.7: *Defining the region $\mathcal{D}^{(k)}$.*

In order to define the region $\mathcal{D}^{(k)}$ mathematically, the orientation of a line segment has to be established, based on the locations of its two endpoints. The eight possible cases in which a line segment may be orientated are illustrated in Figure 6.8. One may, as a result, define the set of points in the region $\mathcal{D}^{(k)}$ for line segment $\ell^{(k)}$ as

$$\mathcal{D}^{(k)} = \begin{cases} \{(x,y) \mid y_R^{(2)} \leq y \leq y_R^{(1)}\} & \text{if } x_R^{(1)} = x_R^{(2)} \text{ and } y_R^{(1)} > y_R^{(2)}, \\ \{(x,y) \mid -m(y - y_R^{(2)}) + x_R^{(2)} \leq x \leq -m(y - y_R^{(1)}) + x_R^{(1)}, & \text{if } x_R^{(1)} > x_R^{(2)} \text{ and } y_R^{(1)} > y_R^{(2)}, \\ \qquad -\frac{1}{m}(x - x_R^{(2)}) + y_R^{(2)} \leq y \leq -\frac{1}{m}(x - x_R^{(1)}) + y_R^{(1)}\} & \\ \{(x,y) \mid x_R^{(2)} \leq x \leq x_R^{(1)}\} & \text{if } x_R^{(1)} > x_R^{(2)} \text{ and } y_R^{(1)} = y_R^{(2)}, \\ \{(x,y) \mid -m(y - y_R^{(2)}) + x_R^{(2)} \leq x \leq -m(y - y_R^{(1)}) + x_R^{(1)}, & \text{if } x_R^{(1)} > x_R^{(2)} \text{ and } y_R^{(1)} < y_R^{(2)}, \\ \qquad -\frac{1}{m}(x - x_R^{(1)}) + y_R^{(1)} \leq y \leq -\frac{1}{m}(x - x_R^{(2)}) + y_R^{(2)}\} & \\ \{(x,y) \mid y_R^{(1)} \leq y \leq y_R^{(2)}\} & \text{if } x_R^{(1)} = x_R^{(2)} \text{ and } y_R^{(1)} < y_R^{(2)}, \\ \{(x,y) \mid -m(y - y_R^{(1)}) + x_R^{(1)} \leq x \leq -m(y - y_R^{(2)}) + x_R^{(2)}, & \text{if } x_R^{(1)} < x_R^{(2)} \text{ and } y_R^{(1)} < y_R^{(2)}, \\ \qquad -\frac{1}{m}(x - x_R^{(1)}) + y_R^{(1)} \leq y \leq -\frac{1}{m}(x - x_R^{(2)}) + y_R^{(2)}\} & \\ \{(x,y) \mid x_R^{(1)} \leq x \leq x_R^{(2)}\} & \text{if } x_R^{(1)} < x_R^{(2)} \text{ and } y_R^{(1)} = y_R^{(2)}, \\ \{(x,y) \mid -m(y - y_R^{(1)}) + x_R^{(1)} \leq x \leq -m(y - y_R^{(2)}) + x_R^{(2)}, & \text{if } x_R^{(1)} < x_R^{(2)} \text{ and } y_R^{(1)} > y_R^{(2)}, \\ \qquad -\frac{1}{m}(x - x_R^{(2)}) + y_R^{(2)} \leq y \leq -\frac{1}{m}(x - x_R^{(1)}) + y_R^{(1)}\} & \end{cases}$$



(a) Case 1: $x_R^{(1)} = x_R^{(2)}$, $y_R^{(1)} > y_R^{(2)}$

(b) Case 2: $x_R^{(1)} > x_R^{(2)}$, $y_R^{(1)} > y_R^{(2)}$

(c) Case 3: $x_R^{(1)} > x_R^{(2)}$, $y_R^{(1)} = y_R^{(2)}$

(d) Case 4: $x_R^{(1)} > x_R^{(2)}$, $y_R^{(1)} < y_R^{(2)}$

(e) Case 5: $x_R^{(1)} = x_R^{(2)}$, $y_R^{(1)} < y_R^{(2)}$

(f) Case 6: $x_R^{(1)} < x_R^{(2)}$, $y_R^{(1)} < y_R^{(2)}$

(g) Case 7: $x_R^{(1)} < x_R^{(2)}$, $y_R^{(1)} = y_R^{(2)}$

(h) Case 8: $x_R^{(1)} < x_R^{(2)}$, $y_R^{(1)} > y_R^{(2)}$

FIGURE 6.8: *The eight possible orientations of a line segment, where the shaded region denotes all points contained in the set* $\mathcal{D}^{(k)}$.

for $x, y \in \mathbb{R}$, where $m = (y_R^{(2)} - y_R^{(1)})/(x_R^{(2)} - x_R^{(1)})$ is the gradient of the line segment $\ell^{(k)}$. If the point $\boldsymbol{p}^{(i)}$ is indeed contained within the region $\mathcal{D}^{(k)}$ and $d_\perp(\boldsymbol{p}^{(i)}, \ell^{(k)})$ is relatively small, it is reasonable to assume that the ith accident may have occurred on the kth road. Consequently, if $\boldsymbol{p}^{(i)} \in \mathcal{D}^{(k)}$, road k is called an *link candidate* for the ith accident.

A disadvantage of the above approach is that the region $\mathcal{D}^{(k)}$ may, in some circumstances, be too restrictive when defining link candidates. To illustrate such an scenario, consider the accident point $\boldsymbol{p}^{(i+1)}$ in Figure 6.9. This point is situated between regions $\mathcal{D}^{(k)}$ and $\mathcal{D}^{(k+1)}$ defined by line segments $\ell^{(k)}$ and $\ell^{(k+1)}$, respectively (*i.e.* the point falls within a so-called "blind-spot"). It is, however, reasonable to assume that the point $\boldsymbol{p}^{(i+1)}$ may have occurred on either road k or road $k+1$. The problem is that, because $\boldsymbol{p}^{(i+1)} \notin \mathcal{D}^{(k)} \cup \mathcal{D}^{(k+1)}$, the line segments $\ell^{(k)}$ and $\ell^{(k+1)}$ are never considered as possible candidates for the assignment of point $\boldsymbol{p}^{(i+1)}$. An additional procedure must, therefore, be devised to address this problem.

With reference to Figure 6.9, consider the two dotted lines subtended from the point $\boldsymbol{p}^{(i+1)}$ to the two respective endpoints $\boldsymbol{e}_1^{(k)}$ and $\boldsymbol{e}_2^{(k)}$ of line segment $\ell^{(k)}$. One can clearly see that the Euclidean distance $d_\triangle(\boldsymbol{p}^{(i+1)}, \boldsymbol{e}_2^{(k)})$ between the points $\boldsymbol{p}^{(i+1)}$ and $\boldsymbol{e}_2^{(k)}$ is smaller than the distance $d_\triangle(\boldsymbol{p}^{(i+1)}, \boldsymbol{e}_1^{(k)})$ between the points $\boldsymbol{p}^{(i+1)}$ and $\boldsymbol{e}_1^{(k)}$. Either way, if one of these distances is relatively small, it is reasonable to assume that accident $i+1$ could have occurred on road k. In the same way, one may also argue that this accident could have taken place on road $k+1$. In this case, both roads k and $k+1$ are called *node candidates* for the accident $i+1$. Although road $k+2$ is a link candidate for the point $\boldsymbol{p}^{(i+1)}$ (*i.e.* $\boldsymbol{p}^{(i+1)} \in \mathcal{D}^{(k+2)}$), it holds that $\hat{d}(\boldsymbol{p}^{(i+1)}, \ell^{(k+2)}) > d_\triangle(\boldsymbol{p}^{(i+1)}, \boldsymbol{e}_2^{(k)}) = d_\triangle(\boldsymbol{p}^{(i+1)}, \boldsymbol{e}_1^{(k+1)})$, and so one would therefore be more inclined to conclude that accident $i+1$ occurred on either road k or road $k+1$.

In order to ensure that an accident object is only associated with a single road object, it is required that, in the context of Figure 6.9, accident $i+1$ is assigned to either road k or $k+1$. It

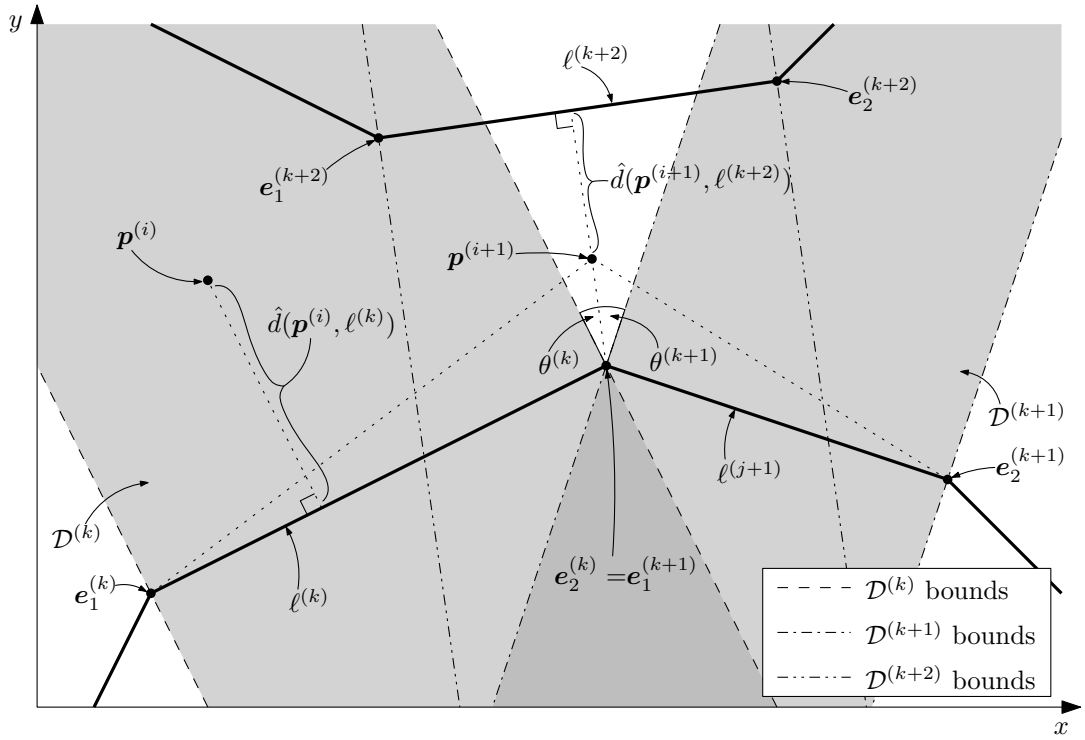


FIGURE 6.9: *A scenario where a point $\boldsymbol{p}^{(i+1)}$ does not fall in regions $\mathcal{D}^{(k)}$ or $\mathcal{D}^{(k+1)}$ but should presumably be assigned to either road k or $k+1$.*

is worth noting that line segments $\ell^{(k)}$ and $\ell^{(k+1)}$ both share a common endpoint $\boldsymbol{e}_2^{(k)} = \boldsymbol{e}_1^{(k+1)}$, and thus also share the line subtended between this endpoint and the point $\boldsymbol{p}^{(i+1)}$. Consequently, it is not clear to which of these two roads accident $i+1$ should be assigned. A general rule must, therefore, be devised to make this road selection. Such a rule may be as follows: Consider the angles $\theta^{(k)}$ and $\theta^{(k+1)}$ formed between the line subtended between the points $\boldsymbol{p}^{(i+1)}$ and $\boldsymbol{e}_2^{(k)} = \boldsymbol{e}_1^{(k+1)}$, and the two respective bounds of regions $\mathcal{D}^{(k)}$ and $\mathcal{D}^{(k+1)}$. One may argue that, if $\theta^{(k)} < \theta^{(k+1)}$, the point $\boldsymbol{p}^{(i+1)}$ is "closer" to region $\mathcal{D}^{(k)}$ than to region $\mathcal{D}^{(k+1)}$ and should, therefore, be assigned to line segment $\ell^{(k)}$. In general, given a set of line segments $\mathcal{J} \subset \mathcal{R}_G$ joined at a shared endpoint $\boldsymbol{e} \in \mathcal{T}$ and a point $\boldsymbol{p}^{(i)} \notin \mathcal{D}^{(1)} \cup \cdots \cup \mathcal{D}^{(|\mathcal{J}|)}$, the accident i should be assigned to road

$$\arg\min_{k \in \mathcal{J}} \theta^{(k)} \tag{6.3}$$

if $d_\triangle(\boldsymbol{p}^{(i)}, \boldsymbol{e})$ is sufficiently small. In the context of accident i, a sufficiently small distance d^* is one in which d^* is the smallest of all perpendicular *and* Euclidean distances produced by the link and node candidates, respectively.

With the above discussion in mind, Algorithm 6.2 is proposed as the procedure according to which an accident object should be assigned to a road object. The input to this algorithm is the point $\boldsymbol{p}^{(i)}$ corresponding to the accident i, together with the line segments corresponding to the set of all roads. The algorithm begins by considering, in turn, each road $k \in \{1, \ldots, m_R\}$. The if-condition in Step 2 tests whether road k is a link candidate. If this is the case, the perpendicular distance $d_\perp(\boldsymbol{p}^{(i)}, \ell^{(k)})$ between the point $\boldsymbol{p}^{(i)}$ and the line segment $\ell^{(k)}$ is appended to an array $\boldsymbol{c}$. If road k is, however, a node candidate, the minimum of the two Euclidean distances $d_\triangle(\boldsymbol{p}^{(i)}, \boldsymbol{e}_1^{(k)})$ and $d_\triangle(\boldsymbol{p}^{(i)}, \boldsymbol{e}_2^{(k)})$, from the point $\boldsymbol{p}^{(i)}$ to each of the respective endpoints of $\ell^{(k)}$, is appended to the array $\boldsymbol{c}$. In Step 6, the set of all roads which result in the minimum distance d^* are assigned to the set $\mathcal{J}$. If this set contains only one road, d^* is as a result of an link candidate, and the

Algorithm 6.2: **Find closest road**

Input : The point $\boldsymbol{p}^{(i)}$ corresponding to the accident i, and the set of all road line segments $\mathcal{R}_G$.

Output: The closest road k^* to accident i.

```
foreach k ∈ {1, ..., m_R} do
    if p^(i) ∈ D^(k) then
        c[k] ← d_⊥(p^(i), ℓ^(k));
    else
        c[k] ← min{d_△(p^(i), e_1^(k)), d_△(p^(i), e_2^(k))};
J ← argmin_{k∈{1,...,m_R}} c[k];
if |J| = 1 then
    k* ← k ∈ J;
else
    k* ← argmin_{k∈J} θ^(k);
return k*;
```

closest road k^* is simply assigned the only element value in the singleton set. If, however, $\mathcal{J}$ contains more than one road, the minimum distance d^* is as a result of multiple node candidates. The value of k^* is then calculated according to (6.3). Finally, the algorithm returns the closest road k^* to accident i. The overall time-complexity of this procedure is, therefore, $\mathcal{O}(m_A m_R)$.

Using the results of Algorithm 6.2, one can define the merged row in the AVR matrix $\boldsymbol{A}_{VR}$ as the concatenation of the vectors $\boldsymbol{a}_V^{(i)}$ and $\boldsymbol{r}^{(k^*)}$, expressed mathematically as

$$\boldsymbol{a}_{VR}^{(i)} = \boldsymbol{a}_V^{(i)} \frown \boldsymbol{r}^{(k^*)} = [a_1^{(i)}, \ldots, a_{n_A}^{(i)}, (v_1^{(i,1)}, \ldots, v_1^{(i,k_i)}), \ldots, (v_{n_V}^{(i,1)}, \ldots, v_{n_V}^{(i,k_i)}), r_1^{(k^*)}, \ldots, r_{n_R}^{(k^*)}].$$

In this way, an $m_A \times (n_A + n_V)$ matrix of accident-vehicle records $\boldsymbol{A}_V$ and the $m_R \times n_R$ matrix of road records may be combined into a single $m_A \times (n_A + n_V + n_R)$ accident-vehicle matrix

$$\boldsymbol{A}_{VR} = \begin{bmatrix} \boldsymbol{a}_{VR}^{(1)} \\ \vdots \\ \boldsymbol{a}_{VR}^{(i)} \\ \vdots \\ \boldsymbol{a}_{VR}^{(m_A)} \end{bmatrix},$$

where $\boldsymbol{a}_{VR}^{(i)}$ is a row vector containing all information associated with the ith accident record $\boldsymbol{a}^{(i)}$ and corresponding vehicle and road records. The matrix $\boldsymbol{A}_{VR}$ is used to define the attribute values of a new set of combined data, denoted by the set $\mathcal{A}_{VR}$. This set is represented spatially by the set of points in $\mathcal{P}$ and the set of attributes is given by $\hat{\mathcal{A}}_{VR} = \{\hat{\mathcal{A}} \cup \hat{\mathcal{V}} \cup \{\hat{\mathcal{R}} \in \hat{\mathcal{X}}_{NS}\} \cup \{\hat{A}_R\}\}$, where the new attribute $\hat{A}_R$ is the road with which an accident has been associated. This new set of data is stored in its own designated table, denoted by Datastore D6.

6.5 Creating severity attributes

The primary function of Module 5.0 in the RARE framework of Figure 5.4 is to facilitate the creation of both continuous and categorical accident severity attributes using the value of the victim severity attribute $\hat{T}_S$, and is shown in greater detail in Figure 6.10. The continuous accident severity measure is necessary to determine the severity of accidents along road segments and at road junctions, while the categorical accident severity measure may serve as a target variable during a classification investigation. This process is now described in greater detail.

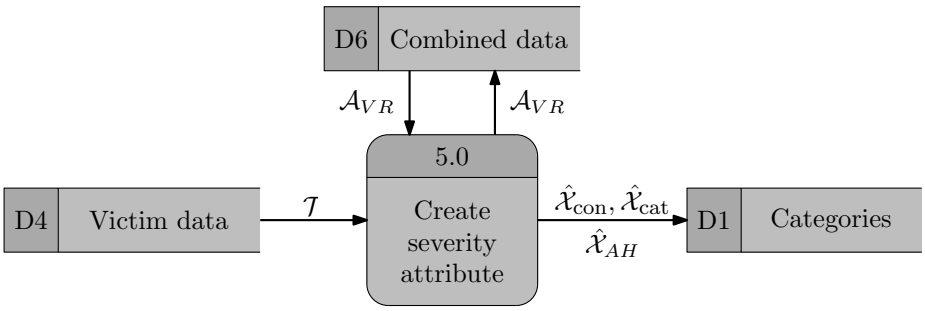


FIGURE 6.10: *A diagram showing in greater detail the flow of data in Module 5.0.*

The task is creating a categorical severity attribute $\hat{A}_S \in \mathcal{X}_{\text{cat}} \cup \mathcal{X}_{AH}$ is relatively straight forward. A user may, for example, easily create a new categorical accident severity attribute `AccidentSeverity` $\in$ {`MoreSevere`, `LessSevere`}, where the decision is left to the user to decide which categories in $\hat{T}_S \in \mathcal{X}_{\text{cat}}$ constitutes a `LessSevere` and `MoreSevere` accident. The procedure for constructing a continuous accident severity attribute $\hat{A}_S \in \mathcal{X}_{\text{cont}} \cup \mathcal{X}_{AH}$ is, on the other hand, not as straight forward. A problem lies in the fact that, as previously mentioned, the accident's severity measure $\hat{T}_S \in \mathcal{X}_{\text{cat}}$ for a specific victim is typically ordinal in nature. As described in §4.2.1, if an attribute is nominal, the categories are all of equal importance and there exists no meaningful sequence in which to order these categories. In this case, every nominal attribute category can be thought of as being separated by a distance of one, as illustrated in Figure 6.11(a), for the example where categories reflect the weather condition. In contrast, an ordinal attribute contains categories that are intrinsically ranked. In this case, each of the possible categories have an associated numerical value (e.g values of 0, 1, 2 and 3). It should, however, be noted that, although categories assume an intrinsic order, it cannot be said, for example, that the true difference between the severity categories `Slight` and `Serious` in Figure 6.11(b) is same as the difference between the categories `Serious` and `Fatal`. It is, therefore, necessary to assign specific costs associated with each of the possible categorical values assumed by the severity attribute $\hat{T}_S$.

Suppose the number of victims associated with c distinct severity types for accident i are captured in the vector $\boldsymbol{s}^{(i)}$ in descending order of severity, where the first entry in $\boldsymbol{s}^{(i)}$ denotes the number of uninjured victims and the last entry in $\boldsymbol{s}^{(i)}$ denotes the number of deceased victims. While the severity types assume an intrinsic order, it can of course not be said that the *true* difference between these types are the same (*e.g.* the difference between a "slight" and "serious" injury is not necessarily the same as the difference between "serious" and "fatal" injury). Consequently, it is necessary that the values in $\boldsymbol{s}^{(i)}$ are weighted to reflect some cost (possibly monetary) associated with each severity type, captured by the vector $\boldsymbol{m}$ hereinafter referred to as the *weight vector*. Then the value of a continuous accident severity attribute $a_S^{(i)}$ may be calculated for accident i as $a_S^{(i)} = \boldsymbol{m}^T \boldsymbol{s}^{(i)}$ for all $i \in \{1, \ldots, m_A\}$. A new updated set of combined data may then be constructed by the concatenation $\mathcal{A}_{VR} := \{\boldsymbol{a}_{VR}^{(i)} \frown a_S^{(i)} \mid i \in \{1, \ldots, m_A\}\}$

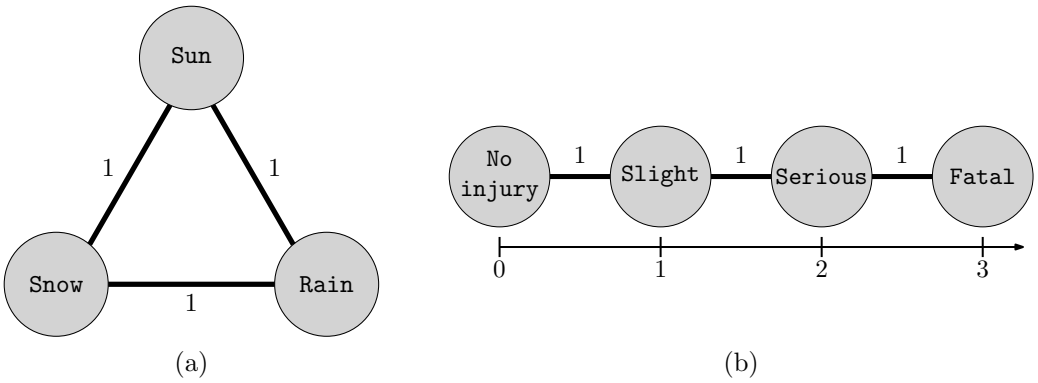


FIGURE 6.11: *A representation of three categories associated with (a) a nominal attribute, where each category is thought of as being separated by a distance of one, and (b) an ordinal attribute, where the distances between categories depend on the rank of each category.*

whereby each new accident severity attribute value is appended with the current accident record. The result subsequently overwrites the combined data set currently stored in the database.

If, however, the weight vector $\boldsymbol{m}$ is unknown, an alternative mechanism must be developed for defining the values in $\boldsymbol{m}$ in such a way that these values better represent the qualitative perception of the severity categories in relation to one another. It is, therefore, proposed that this be achieved by employing a so-called *scaling function* $\varphi(\cdot)$ which may effectively be used to adjust the distance between the categories of the severity attribute $\hat{T}_S$ by defining the weight vector as

$$\boldsymbol{m} = \left\langle \varphi\left(0\right), \varphi\left(\frac{1}{c}\right), \cdots, \varphi\left(\frac{1-c}{c}\right), \varphi\left(1\right) \right\rangle,$$

where $c = |\boldsymbol{m}|$ is the number of categorical values the severity attribute $\hat{T}_S$ may assume.

A natural question is how the function $\varphi(\cdot)$ should be defined. There are three natural characteristics of this function:

1. The function $\varphi(x)$ must be defined everywhere on the interval $x \in [0, 1]$,
2. the function $\varphi(x)$ must be strictly increasing on the interval $[0, 1]$ (*i.e.* $\varphi(b) > \varphi(a)$ for all $a, b \in [0, 1]$ satisfying $b > a$), and
3. the coordinates of the points $(0, 0)$ and $(1, 1)$ must be associated by the function $\varphi(x)$ (*i.e.* $\varphi(0) = 0$ and $\varphi(1) = 1$).

Property 1 ensures, for an arbitrary value between 0 and 1, that there always exists a corresponding scaled value. Property 2 ensures that the ranking integrity of the values in $\boldsymbol{m}$ is maintained. Property 3 finally ensures that the most and least severe categories maintain unit and zero values, respectively, and that all other values are scaled within the interval $(0, 1)$.

The function $\varphi(x) = x^\eta$ is proposed as an appropriate scaling function on the unit interval $[0, 1]$ for some constant $\eta \in \mathbb{R}^+$. A number of benefits are associated with choosing this simple function. First, the function allows one to define a family of strictly increasing linear, convex or concave functions by merely specifying the value of η. In this way, a user may weight the values in $\boldsymbol{s}$ in such a way that they are deemed to better represent the relative qualitative perception of severity between the categories. Secondly, this function can be reverse-engineered by allowing a user to specify any value (x^*, y^*) in the unit square $x^*, y^* \in [0, 1] \times [0, 1]$ and taking the value of η as $\eta = \log_{x^*}(y^*)$, which would ensure that $(x^*)^\eta = y^*$. Lastly, as $\eta \to \infty$, the function

$x^\eta \to 0$ in the interval $x \in (0,1)$. In other words, a large positive choice of η allows a user to specify the weight vector as $\boldsymbol{m} = \langle 0, 0, \cdots, 0, 1 \rangle$, which effectively considers only the most extreme severity category as relevant in the context of accident consequence. Conversely, as $\eta \to 0$, the function $x^\eta \to 1$ in the interval $x \in (0,1)$. In other words, a small positive choice of η allows a user to specify the weight vector as $\boldsymbol{m} = \langle 1, 1, \cdots, 1, 1 \rangle$, which effectively considers all severity categories as equal.

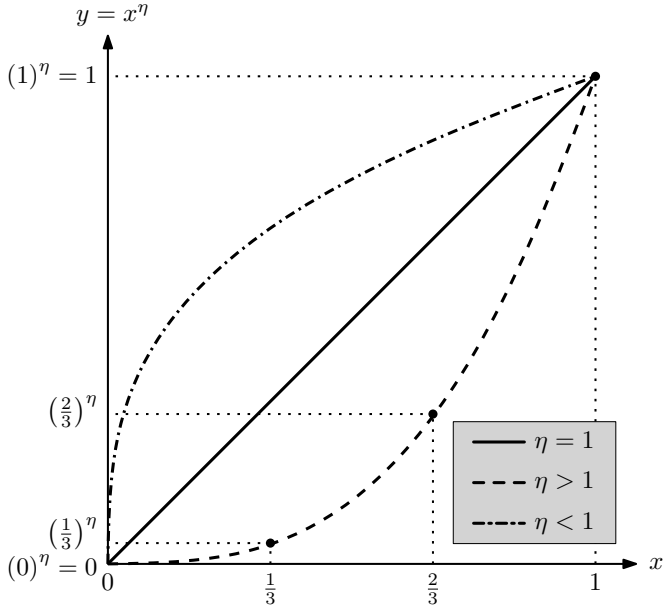


FIGURE 6.12: *A plot of the function x^η where the attribute $\hat{T}_S$ has $c = 4$ distinct categories.*

6.6 Constructing a derived attribute

The primary function of Module 6.0 in the RARE framework of Figure 5.4 is to allow a user to enhance the data currently stored in the database by constructing new derived attributes, and is shown in greater detail in Figure 6.13. The user is first required to define an algorithmic procedure whereby one or more of the attributes are used to construct the new attribute. This algorithmic expression is then performed to produce a new attribute column which is appended to the combined data set currently stored in the database. This process is now described in greater detail.

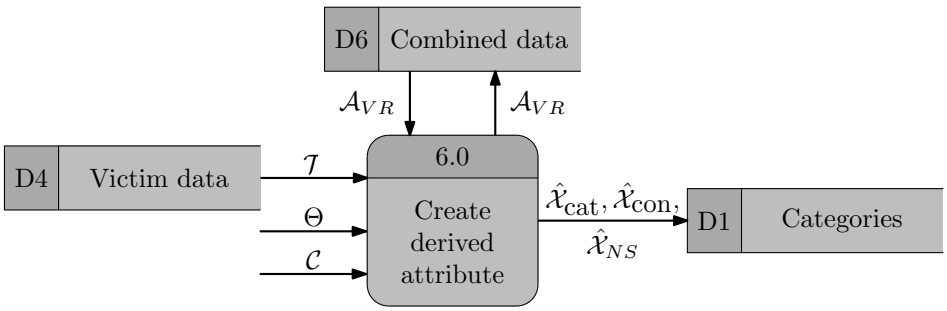


FIGURE 6.13: *A diagram showing in greater detail the flow of data in Module 6.0.*

The user is first required to establish the data type of the attribute being constructed. If this new attribute, denoted by $\hat{A}_{\text{new}}$, is categorical in nature, the user is then also required to specify a set of C_{new} categories $\mathcal{C} = \{\hat{A}^{(1)}_{\text{new}}, \ldots, \hat{A}^{(C_{\text{new}})}_{\text{new}}\}$ corresponding to the new attribute. Thereafter, the user is expected to define an algorithmic procedure Θ which is, in essence, a function of one or more of the current attributes in $\hat{\mathcal{A}}_{VR}$ or $\hat{\mathcal{T}}$. One may, for example, wish to define a new accident attribute `PedestrianAccident` $\in \{$`Yes`, `No`$\}$ indicating whether any of the victims involved were pedestrians. In this case, one would need to first identify an attribute $\hat{T} \in \hat{\mathcal{T}} \cup \hat{\mathcal{X}}_{\text{cat}}$ which designates the victim type (*i.e.* driver, passenger or pedestrian) which may be used to construct the new attribute $\hat{A}_{\text{new}}$. Again, in the case where $\hat{\mathcal{A}}_{\text{new}}$ is categorical, the algorithmic procedure $\Theta(\hat{\mathcal{A}}_{VR})$ (or $\Theta(\hat{\mathcal{T}})$) must produce strictly integer values bounded between the values of one and C_{new} inclusive (*i.e.* $\Theta(\hat{\mathcal{A}}_{VR}) \in \{1, \ldots, C_{\text{new}}\} \cap \mathbb{Z}$).

Once this has been done, an attribute value $a^{(i)}_{\text{new}}$ can be calculated for each accident $i \in \{1, \ldots, m_A\}$ according to the algorithmic procedure $\Theta(\hat{\mathcal{A}}_{VR})$ (or $\Theta(\hat{\mathcal{T}})$). In the case where $\hat{A}_{\text{new}}$ is continuous, $a^{(i)}_{\text{new}} = \Theta(\hat{\mathcal{A}}_{VR})$ (or $a^{(i)}_{\text{new}} = \Theta(\hat{\mathcal{T}})$) for all $i \in \{1, \ldots, m_A\}$. Otherwise, in the case of a categorical attribute, the algorithmic procedure should produce a value $c = \Theta(\hat{\mathcal{A}}_{VR})$ (or $c = \Theta(\hat{\mathcal{T}})$) which defines the category value $a^{(i)}_{\text{new}} = \hat{A}^{(c)}_{\text{new}} \in \mathcal{C}$ for all $i \in \{1, \ldots, m_A\}$. A new updated set of combined data may then be constructed by the concatenation $\mathcal{A}_{VR} = \{\boldsymbol{a}^{(i)}_{VR} \frown a^{(i)}_{\text{new}} \mid i \in \{1, \ldots, m_A\}\}$ whereby each new attribute value is appended to the current accident record. These result subsequently overwrites the combined data set currently stored in the database.

6.7 Modelling road accident risk along a road segment

The primary function of Module 7.0 in the RARE framework of Figure 5.4 is to propose a descriptive modelling approach for quantifying the rate and severity of road accidents along a road segment, thereby accomplishing two of the eight elements of the envisioned RARE framework, as illustrated in Figure 1.4. This is achieved by means of three sub-modules, as illustrated in Figure 6.14. In Sub-module 7.1, a novel modelling approach is employed for quantifying the RA rate and severity along a road segment which takes inspiration from the three dimensions of road safety, as described in §2.2.2. Thereafter, these may be visualised by means of Sub-modules 7.2 and 7.3 according to a spatial and attribute distribution representation, respectively. These sub-modules are now described in greater detail.

Consider the set of m_A historical RAs $\mathcal{P} = \{\boldsymbol{p}^{(1)}, \ldots, \boldsymbol{p}^{(m_A)}\}$ currently stored in Datastore D2, where an accident $\boldsymbol{p}^{(i)}$ is spatially represented by the geographical point $\boldsymbol{p}^{(i)} = (x_A, y_A)^{(i)}$. Furthermore, consider the set of m_R roads $\mathcal{L} = \{\ell^{(1)}, \ldots, \ell^{(m_R)}\}$ currently stored in Datastore D4, where, as previously stated, a road is spatially represented by the geographical straight line segment $\ell^{(k)} = \{(x^{(1)}_R, y^{(1)}_R)^{(k)} + u(x^{(2)}_R, y^{(2)}_R)^{(k)} \mid u \in [0, 1]\}$. As previously mentioned in §6.2, it is also assumed that the kth road $\ell^{(k)}$ has an attribute value of $r^{(k)}_T$ indicating the number of vehicles that have passed along the road during a specified time period.

The first step in the proposed methodology for modelling RA risk (executed in Sub-module 7.1) is to partition each road in $\mathcal{L}$ into a set of subintervals of approximately equal length. If a maximum subinterval length ψ is defined, then the road $\ell^{(k)}$ is partitioned into

$$g_k = \begin{cases} \frac{|\ell^{(k)}|}{\psi}, & \text{if } |\ell^{(k)}| \text{ is divisible by } \psi, \\ \left\lfloor \frac{|\ell^{(k)}|}{\psi} \right\rfloor + 1, & \text{otherwise,} \end{cases}$$

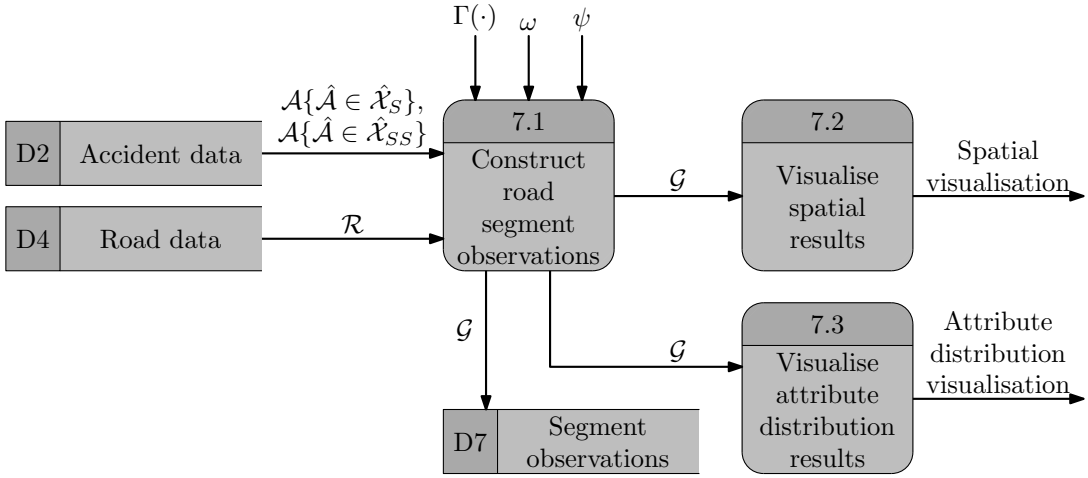


FIGURE 6.14: *A child diagram of Module 7.0 showing in greater detail the flow of data between sub-modules.*

subintervals, each with a length of

$$h_k = \frac{|\ell^{(k)}|}{g_k},$$

where $|\ell^{(k)}|$ denotes the length of line segment $\ell^{(k)}$. This ensures that the combined length of each g_k subintervals never exceeds the maximum subinterval length ψ for all roads in $\mathcal{L}$ (*i.e.* $h_k < \psi$ for all $i \in \{1, \ldots, m_R\}$). A *road observation* $\ell_j^{(k)}$ is defined as the set of points satisfying

$$\left\{ \left(x_R^{(1)}, y_R^{(1)}\right)^{(k)} + u \left(x_R^{(2)}, y_R^{(2)}\right)^{(k)} \mid u \in \left[\frac{j-1}{g_k}, \frac{j}{g_k}\right] \right\},$$

where all $m_G = \sum_{k=1}^{m_R} g_k$ road observations containing RAs are contained in the set $\mathcal{G} = \{\ell_j^{(k)} \mid j \in \{1, \ldots, g_k\},\ k \in \{1, \ldots, m_R\}\}$. Notice that if $\psi \to 0$, then $|\mathcal{G}| \to \infty$ and $|\ell_j^{(k)}| \to \psi$ for all $j \in \{1, \ldots, g_k\}$ and $k \in \{1, \ldots, m_R\}$. The quantity ψ must, therefore, be chosen such that the variance in road observation lengths is sufficiently small (to ensure that the set of road observations is unbiased) and so that $\mathcal{G}$ contains an adequate number of road observations.

It follows that the frequency $F(\ell_j^{(k)})$ of RAs that have occurred along road observation $\ell_j^{(k)}$ is

$$F(\ell_j^{(k)}) = M \int_{\ell_j^{(k)}} \Gamma(q)\, \mathrm{d}q,$$

where $\Gamma(q)$ is a user-defined network kernel density estimator with a bandwidth of ω, as described in §3.3.2. The accident rate $R(\ell_j^{(k)})$ along road observation $\ell_j^{(k)}$ is then simply the normalised RA frequency

$$R(\ell_j^{(k)}) = \frac{F(\ell_j^{(k)})}{r_T^{(k)}}. \tag{6.4}$$

The next step is to define a continuous severity measure $S(\ell_j^{(k)})$ for each road observation in $\mathcal{L}$ as a combined severity effect of all the accidents occurring along a road observation $\ell_j^{(k)}$. As previously described in §6.5, it is assumed that the number of victims associated with c distinct severity types for accident i are captured in the vector $\boldsymbol{s}^{(i)}$ in descending order of severity,

where $a_S^{(i)} = \boldsymbol{m}^T \boldsymbol{s}^{(i)}$ denotes the severity of accident i. The accident severity $S(\ell_j^{(k)})$ along road observation $\ell_j^{(k)}$ is defined as

$$S(\ell_j^{(k)}) = \frac{\int_{\ell_j^{(k)}} \sum_{i=1}^{m_A} a_S^{(i)} \Gamma_{\boldsymbol{p}^{(i)}}(q)\, \mathrm{d}q}{F(\ell_j^{(k)})}, \tag{6.5}$$

where each base kernel function $\Gamma_{\boldsymbol{p}^{(i)}}(q)$ is weighted by the scaled measure $a_S^{(i)}$ and the individual effects are summed and normalised for all $i \in \{1, \ldots, m_A\}$ along $\ell_j^{(k)}$.

The final step in this RA risk modelling approach is concerned with inferring a set of attributes, for each road observation in $\mathcal{G}$, which describe inherent characteristics of the roadway from the accidents that have occurred on a road observation $\ell_j^{(k)}$. Given that each RA has a set of $n_G = |\mathcal{X}_{SS} \cup \mathcal{X}_{\text{cat}}|$ associated semi-spatial categorical attributes, where the nth attribute has a set of associated values collected in $\mathcal{C}^{(n)}$, the proportion of RAs along road observation $\ell_j^{(k)}$ describing the road according to a specific feature value is given by

$$x_n^{(c)}(\ell_j^{(k)}) = \frac{\sum_{\substack{i=1,\\ x_n^{(i)}=c}}^{m_A} \int_{\ell_j^{(k)}} \Gamma_{\boldsymbol{p}^{(i)}}(q)\, \mathrm{d}q}{F(\ell_j^{(k)})} \quad \text{for } c \in \mathcal{C}^{(n)}. \tag{6.6}$$

It must be noted that this formulation is only valid if the accident attributes considered describe inherent properties of the roadway (and not the accident itself), as a majority consensus must be established between the accidents occurring along a road observation $\ell_j^{(k)}$.

Having established quantitative expressions for $F(\ell_j^{(i)})$, $R(\ell_j^{(i)})$ and $S(\ell_j^{(i)})$ over the network $\mathcal{N}$, together with the proportion of accidents $x_n^{(c)}(\ell_j^{(k)})$ which possess a value of c for the nth categorical semi-spatial attribute, the results obtained *via* these expressions can be communicated to the user in the form of a geographical plot of the road network $\mathcal{N}$ in Sub-module 7.2 or a two-dimensional scatter plot where RA rate is plotted against RA severity for the different attribute category subsets in Sub-module 7.3. This may, therefore, allow a user to form an intuitive understanding of the RAR characterised by the accidents occurring on the road network $\mathcal{N}$.

6.8 Modelling road accident risk at a road junction

The primary function of Module 8.0 in the RARE framework of Figure 5.4 is to propose a descriptive modelling approach for quantifying the rate and severity of road accidents at a road junction, thereby accomplishing a further two of the eight elements of the envisioned RARE framework, as illustrated in Figure 1.4. This is achieved by means of five sub-modules, as illustrated in Figure 6.15. In Sub-modules 8.1, 8.2 and 8.3, a novel modelling approach is employed for grouping a set of accidents occurring at the same junction for the purpose of quantifying the RA rate and severity at these junctions. Thereafter, these results may be visualised by means of Sub-modules 8.4 and 8.5 according to a spatial and attribute distribution representation, respectively. These sub-modules are now described in greater detail.

As a way of explanation, consider the hypothetical example illustrated in Figure 6.16, where a set of accidents are geographically distributed around two road junctions, labelled A and B. In this

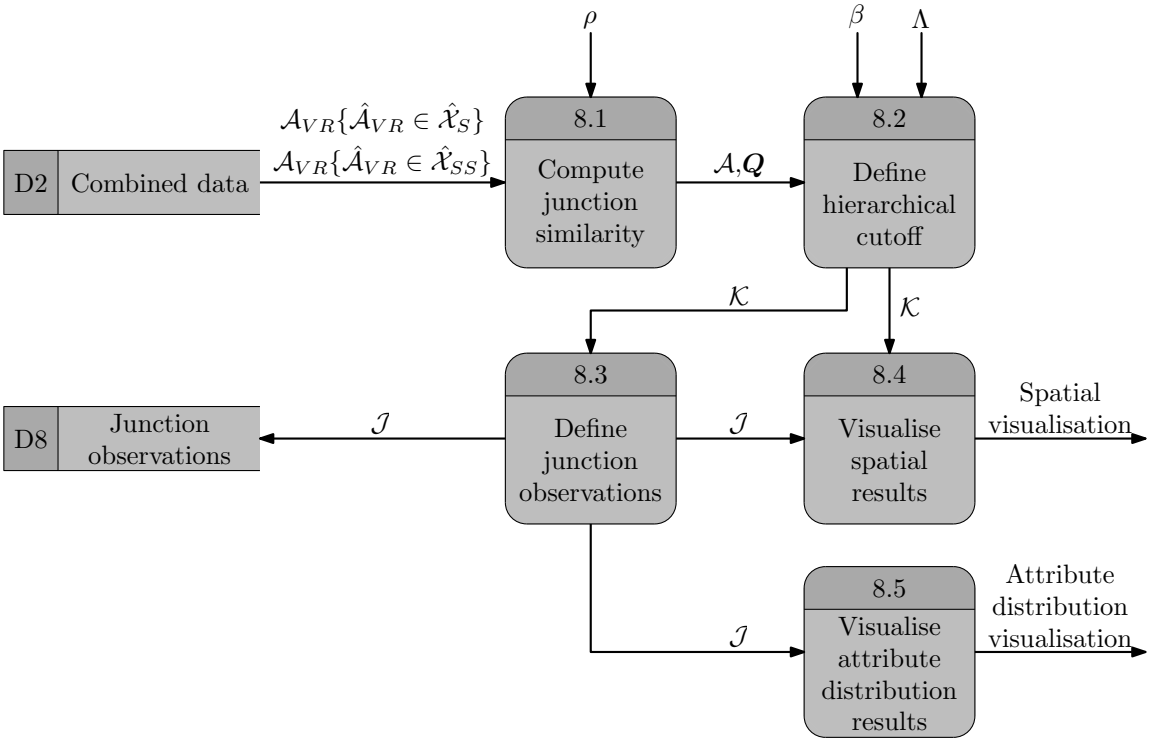


FIGURE 6.15: *A child diagram of Module 8.0 showing in greater detail the flow of data between sub-modules.*

example, it is not always clear from this spatial representation whether an accident occurred at either junction A or B. If one were, however, to consider the set of semi-spatial attributes of the accidents occurring at some form of junction, denoted by $\mathcal{A}_J = \mathcal{A}\{\hat{\mathcal{A}} \in \hat{\mathcal{X}}_{SS} \cap \hat{A}_J = \texttt{AtJunction}\}$, one may be better able to make this distinction. Accident $\boldsymbol{p}^{(10)}$, for example, can initially be excluded from this analysis since both the attributes `JunctionType` and `JunctionControl` for this accident have a value of `NotAtJunction`. Next, consider accident $\boldsymbol{p}^{(5)}$ which is within a

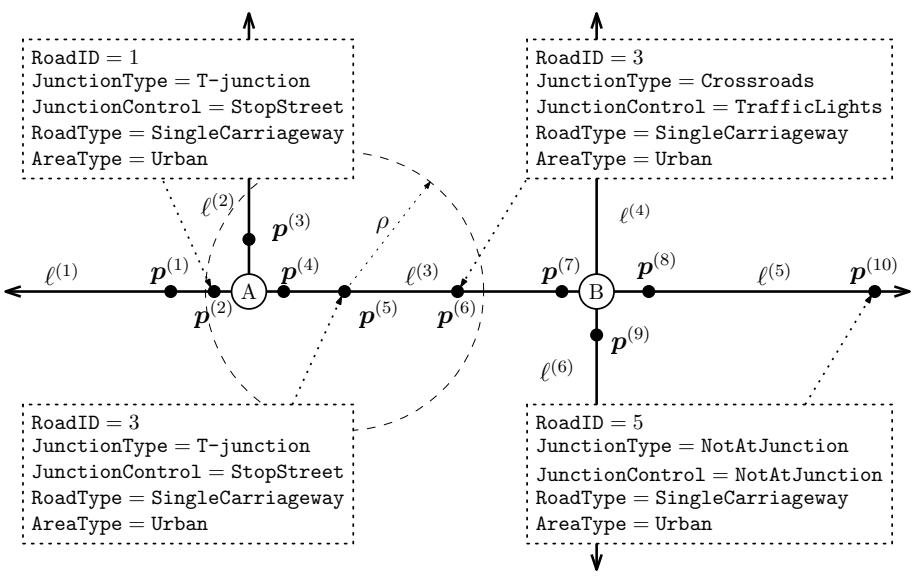


FIGURE 6.16: *A hypothetical example of accident points occurring at two road junctions.*

distance ρ from both accidents $\boldsymbol{p}^{(2)}$ and $\boldsymbol{p}^{(6)}$. Although both accidents $\boldsymbol{p}^{(5)}$ and $\boldsymbol{p}^{(6)}$ occurred on the same road segment, the values of their `JunctionType` attribute are `T-junction` and `Crossroads`, respectively. One should, therefore, be more inclined to believe that accident $\boldsymbol{p}^{(5)}$ occurred at junction A and accident $\boldsymbol{p}^{(6)}$ occurred at junction B. Conversely, although accident $\boldsymbol{p}^{(2)}$ is slightly further from accident $\boldsymbol{p}^{(5)}$ than accident $\boldsymbol{p}^{(6)}$, one should be more inclined to believe that accidents $\boldsymbol{p}^{(5)}$ and $\boldsymbol{p}^{(2)}$ occurred at the same junction since they possess the same values for all their associated semi-spatial attributes. In light of this hypothetical example, it is presupposed that two sources of information should be utilised to make the distinction as to which of the $m_S = |\mathcal{A}_J|$ accidents occurred at the same road junction, namely: Accidents occurring at the same junction should be (1) geographically close to one another and (2) have the same (or similar) semi-spatial attributes which describe inherent properties of the roadway.

Sub-module 8.1 leverages these two notions by computing a so-called *junction dissimilarity score* q_{ij} between accidents i and j (which were initially considered to have occurred at some form of road junction) for all $i, j \in \{1, \ldots, m_S\}$, defined mathematically as

$$q_{ij} = \begin{cases} \frac{d_\triangle(\boldsymbol{p}^{(i)}, \boldsymbol{p}^{(j)})^2}{\Lambda(\boldsymbol{a}_J^{(i)}, \boldsymbol{a}_J^{(j)})} + 1, & \text{if } d_\triangle(\boldsymbol{p}^{(i)}, \boldsymbol{p}^{(j)}) < \rho \text{ and } i \neq j, \\ \infty, & \text{otherwise,} \end{cases} \tag{6.7}$$

where $d_\triangle(\boldsymbol{p}^{(i)}, \boldsymbol{p}^{(j)})$ is the Euclidean distance between points $\boldsymbol{p}^{(i)}$ and $\boldsymbol{p}^{(j)}$, ρ is a user-defined radius, and $\Lambda(\boldsymbol{a}_J^{(i)}, \boldsymbol{a}_J^{(j)}) \in [0, 1]$ is a user-defined similarity measure between the vectors $\boldsymbol{a}_J^{(i)} \in \mathcal{A}_J$ and $\boldsymbol{a}_J^{(j)} \in \mathcal{A}_J$ of semi-spatial attribute values corresponding to the ith and j^{th} accidents, respectively. This dissimilarity measure must be defined by the user in such a way so that, if $\boldsymbol{a} = \boldsymbol{b}$, then $\Lambda(\boldsymbol{a}, \boldsymbol{b}) = \Lambda(\boldsymbol{a}, \boldsymbol{a}) = 1$. If, on the other hand, $\boldsymbol{a} \neq \boldsymbol{b}$, then $0 < \Lambda(\boldsymbol{a}, \boldsymbol{b}) < 1$. As a way of example, consider the vectors $\boldsymbol{a} = \langle 0\,2\,1\,0\,1\,3 \rangle$, $\boldsymbol{b} = \langle 2\,2\,1\,0\,1\,3 \rangle$ and $\boldsymbol{c} = \langle 3\,1\,2\,1\,0\,3 \rangle$. One would expect $\Lambda(\boldsymbol{a}, \boldsymbol{b}) \approx 1$ since only the first entries in $\boldsymbol{a}$ and $\boldsymbol{b}$ are not equal, and $\Lambda(\boldsymbol{a}, \boldsymbol{c}) \approx 0$ as only the last entries in $\boldsymbol{a}$ and $\boldsymbol{c}$ are equal. There exists a wide variety of choices in the literature for defining the similarity measure $\Lambda(\boldsymbol{a}, \boldsymbol{b})$ between two vectors $\boldsymbol{a}$ and $\boldsymbol{b}$. Chandola *et al.* [41], however, noted that the *Lin*, *Goodall* and *Occurrence Frequency* measures achieve consistently better clustering performance across a wide variety of benchmark data sets.

Notice that if accidents i and j have almost identical semi-spatial attribute values (*i.e.* $\Lambda(\boldsymbol{a}_J^{(i)}, \boldsymbol{a}_J^{(j)}) \approx 1$) and are geographically close to one another (*i.e.* $\rho > d_\triangle(\boldsymbol{p}^{(i)}, \boldsymbol{p}^{(j)}) \approx 0$), their junction dissimilarity score q_{ij} will be relatively small (*i.e.* $q_{ij} \approx 0$) which means that accidents i and j are more likely to have occurred at the same junction. Conversely, if $\Lambda(\boldsymbol{a}_J^{(i)}, \boldsymbol{a}_J^{(j)}) \approx 0$ or $d_\triangle(\boldsymbol{p}^{(i)}, \boldsymbol{p}^{(j)}) \geq \rho$, then q_{ij} will be relatively large, indicating that accidents i and j are more likely to have occurred at different junctions. All junction dissimilarity scores corresponding to the m_S accidents that were considered to have occurred at some form of junction are contained in a symmetric *junction dissimilarity matrix* $\boldsymbol{Q}$, like that shown in Figure 6.17(a), where in the figure, a white cell indicates a small dissimilarity value and a dark blue cell indicates a large dissimilarity value. The parameter ρ in (6.7) serves as a way of reducing the space and time complexity associated with computing all $m_S(m_S - 1)/2$ pairwise comparisons (if the symmetry of $\boldsymbol{Q}$ is exploited), since, if two accidents are geographically far from one another (as suspected by the user), one can immediately exclude the possibility that the accidents occurred at the same junction. One may, therefore, represent $\boldsymbol{Q}$ in the form of a sparse matrix containing only pairwise comparisons achieving a geographical distance less than ρ. If $d_\triangle(\boldsymbol{p}^{(i)}, \boldsymbol{p}^{(j)}) < \rho$, the Euclidean distance between accidents i and j is squared in (6.7) so as to penalise accidents more severely that are geographically far from one another.

Once computed, the junction dissimilarity matrix $\boldsymbol{Q}$ can be represented as a tree-like diagram called a *dendrogram*, such as that shown in Figure 6.17(b), which is often employed in *hierarchical*

cluster analysis (HCA). In this diagram, accidents with small pairwise junction dissimilarity scores branch closer to the leaves of the dendrogram while accidents with large pairwise junction dissimilarity scores branch closer to the root of the dendrogram. A user may, therefore, define a set of m_J geographical clusters, contained in the set $\mathcal{K}$, as illustrated in Figure 6.7(c), by merely specifying a so-called *junction dissimilarity threshold* β in Sub-module 8.2, as illustrated in Figure 6.7(b). In other words, if $q_{ij} < \beta$, accidents i and j form part of the same junction, and, if $q_{ij} \geq \beta$, accidents i and j are considered to have occurred at different junctions. A new attribute $\hat{A}_C \in \mathcal{X}_{\text{meta}}$ is subsequently appended to each accident in $\mathcal{K}$, indicating to which cluster the accident belongs (which serves as a link attribute). The quality of these clusters may be assessed visually by the user in Sub-module 8.4 and, once a suitable junction similarity threshold has been determined, a set of *junction observations* $\mathcal{J} = \{u^{(1)}, \ldots, u^{(m_J)}\}$ may be represented spatially as the centroids of all accidents contained in the same geographical cluster (computed in Sub-module 8.3), as illustrated in Figure 6.7(d), where the kth junction observation $u^{(k)}$ assumes a new attribute $\hat{J}_I \in \mathcal{X}_{\text{meta}}$ which serves as the junction identifier and an attribute $\hat{J}_N \in \mathcal{X}_{\text{cont}}$ denoting the number of accidents in the junction observation cluster.

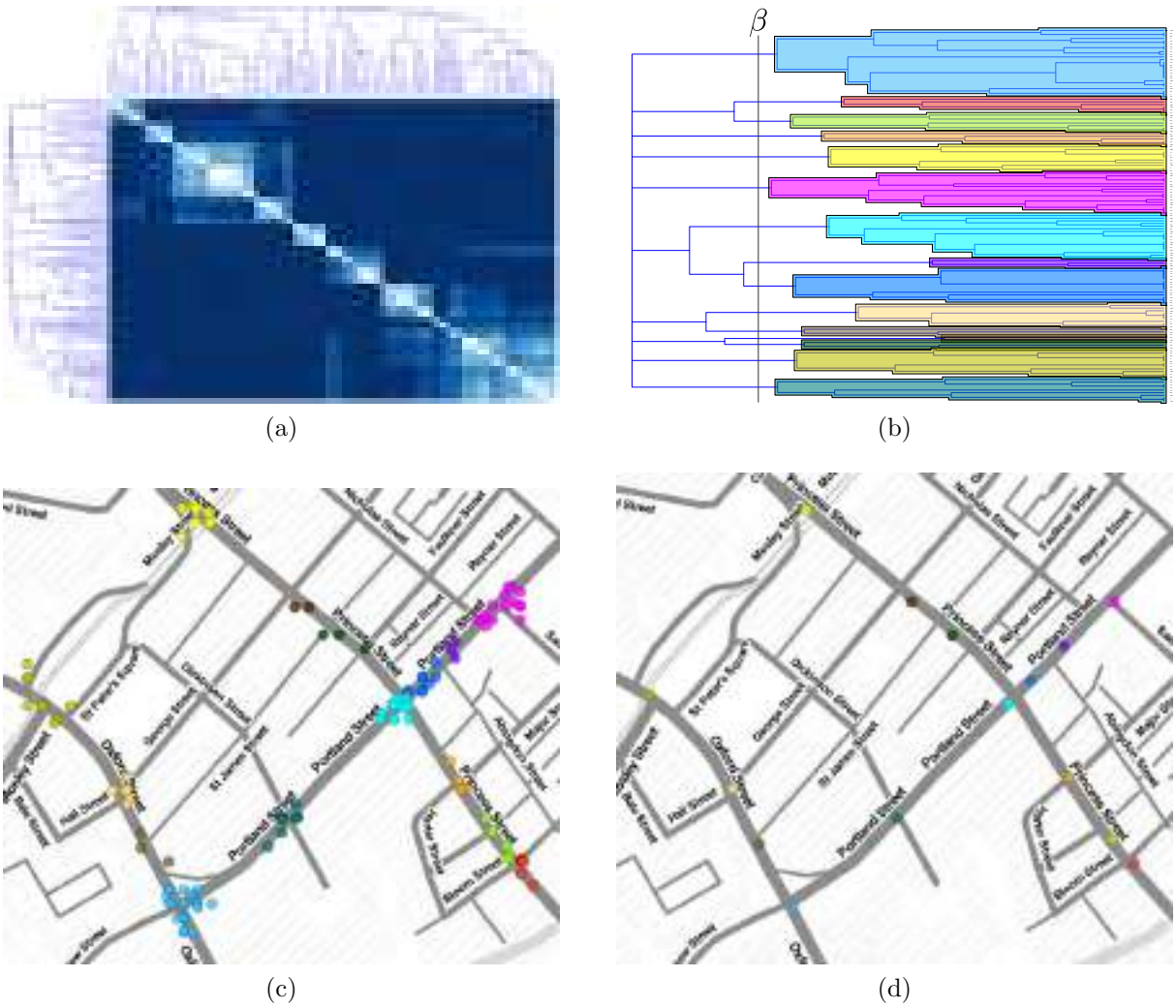


FIGURE 6.17: *The first step in Module 8.0 involves (a) computing the junction similarity matrix so that (b) one may define a junction similarity threshold β which (c) defines a set of geographical clusters, so that (d) the centroid of a cluster defines the spatial representation of a junction observation.*

Given the set of junction observations $\mathcal{J} = \{u^{(1)}, \ldots, u^{(m_J)}\}$, the next step in Module 8.3 is to define an RA rate and severity attribute value for each junction observation by aggregating the accidents in the set of clusters $\mathcal{K} = \{\mathcal{K}_1, \ldots, \mathcal{K}_{m_J}\}$ which belong to that junction, where the kth cluster is defined by the set of accidents $\mathcal{K}_k = \{\boldsymbol{a}_J^{(i)} \mid i \in m_S, \hat{A}_C = k\}$. The RA frequency at the kth junction is simply the number of accidents in a cluster, or $F(u^{(k)}) = |\{\boldsymbol{a}_J^{(i)} \mid i \in m_S, \hat{A}_C = k\}|$. Given that the number of vehicles that have passed through junction k may be approximated as

$$t_k = \frac{\sum_{i=1,\ \hat{A}_C=k}^{m_S} r_T^{(i)}}{F(u^{(k)})}, \tag{6.8}$$

the RA rate at junction observation $u^{(k)}$, denoted by $R(u^{(k)})$, may be computed in a similar way to (6.4), expressed mathematically as

$$R(u^{(k)}) = \frac{F(u^{(i)})}{t_k}. \tag{6.9}$$

Likewise, the RA severity at a junction observation $u^{(k)}$, denoted by $S(u^{(k)})$, may also be computed in a similar way to (6.5), expressed mathematically as

$$S(u^{(k)}) = \frac{\sum_{i=1,\ \hat{A}_C=k}^{m_S} a_S^{(i)}}{F(u^{(i)})}, \tag{6.10}$$

where $a_S^{(i)} = \boldsymbol{m}^T \boldsymbol{s}^{(i)}$ again denotes the severity of accident i.

Given that the nth categorical accident attribute has a set of associated values collected in $\mathcal{C}^{(n)}$, the proportion of RAs at a junction observation $u^{(k)}$ describing the road according to a specific feature value, denoted by $x_n^{(c)}(u^{(k)})$, relies on the same notions as (6.6), expressed mathematically as

$$x_n^{(c)}(u^{(k)}) = \frac{|\{\boldsymbol{a}_J^{(i)} \mid i \in m_S, \hat{A}_C = k, x_n^{(i)} = c\}|}{F(\ell_j^{(k)})} \quad \text{for } c \in \mathcal{C}^{(n)}. \tag{6.11}$$

It should again be noted that this formulation is only valid if the accident attributes considered describe inherent properties of the junction (and not the accident itself), as a majority consensus must be established between the accidents occurring at a junction observation $u^{(k)}$.

Having established quantitative expressions for $F(u^{(k)})$, $R(u^{(k)})$ and $S(u^{(k)})$ over the network $\mathcal{N}$, together with the proportion of accidents $x_n^{(c)}(u^{(k)})$ possessing a value of c for the nth categorical semi-spatial attribute, the results obtained *via* these expressions can again be communicated to the user in the form of a geographical plot of the road network $\mathcal{N}$ in Sub-module 7.2 or a two-dimensional scatter plot where RA rate is plotted against RA severity for the different attribute category subsets in Sub-module 7.3. This may, therefore, allow a user to form an intuitive understanding of the RAR characterised by the accidents occurring at the road junctions on the network $\mathcal{N}$.

6.9 Predicting road accident risk

The primary function of Module 9.0 in the RARE framework of Figure 5.4 is to propose a predictive modelling approach for characterising the rate and severity of road accidents along road segments and at road junctions, thereby accomplishing the remaining four of the eight

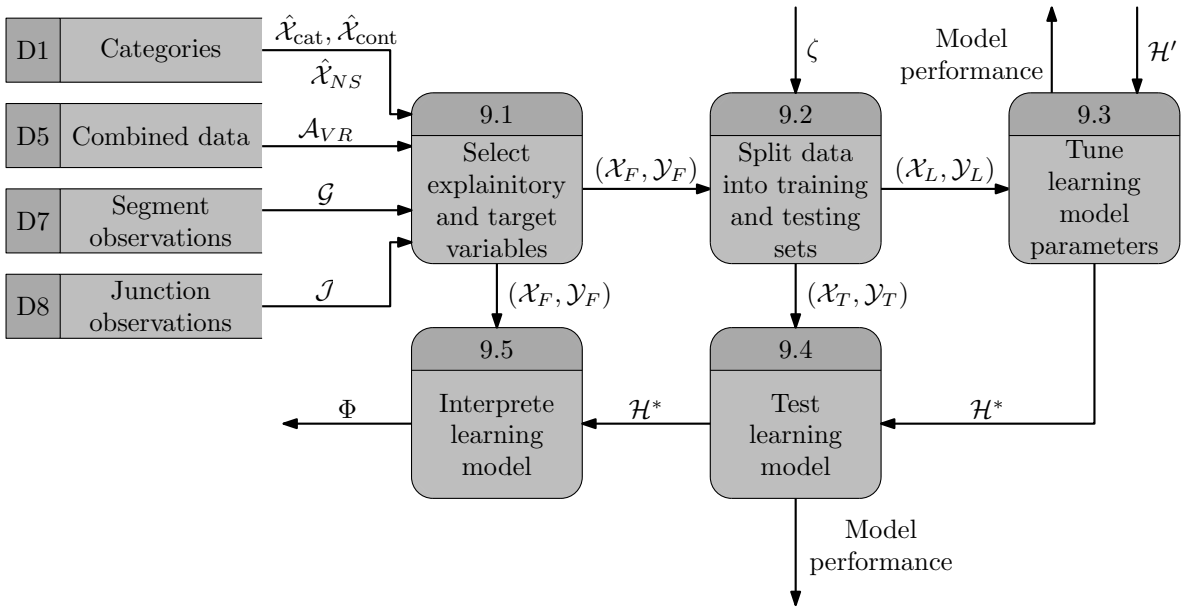


FIGURE 6.18: *A child diagram of Module 9.0 showing in greater detail the flow of data between sub-modules.*

elements of the envisioned RARE framework, as illustrated in Figure 1.4. This is achieved by means of five sub-modules, as illustrated in Figure 6.18, which collectively allow the user to conduct a standard supervised learning analysis, as described in §4.1.2. The user is first allowed to filter the set of selected data in Sub-module 9.1 such that only the records/attributes that are deemed relevant to the investigation at hand are considered. In Sub-module 7.2, the data set is partitioned into a learning data set and a test data set, as described in §4.2.3. The learning data set is used in Sub-module 7.3 to tune the hyper-parameters of various learning models so as to ascertain the best functional mapping between the defined sets of explanatory and target variables. The performance of the tuned learning model(s) is then evaluated in Sub-module 9.4 against the test data set so as to obtain a final unbiased performance evaluation. Insight into the rule formulation of the specified learning models are finally communicated to the user in Sub-module 9.5 using the SHAP method described in §4.3.3 so that an intuitive understanding may be formed of how the values of specific explanatory attributes affect the response of the target attribute. This process is now described in greater detail.

As mentioned, the user is first required, in Sub-module 9.1, to construct a set of filtered data $\mathcal{X}_F = \mathcal{X}\{\bullet\}$ (with a corresponding set of target variable values $\mathcal{Y}_F$) by specifying one or more condition(s), denoted by the place holder $\bullet$, which establishes a subset of attributes/records contained in an arbitrary set of data $\mathcal{X}$. In this formulation, the user may choose either the combined data $\mathcal{A}_{VR}$, segment observations $\mathcal{G}$ or junction observations $\mathcal{J}$ as an appropriate set of input data and define either an associated RA rate or severity measure as the target variable. The selection of various attributes and/or records is purely up to the user — this selection will largely be based on the research question under consideration coupled with specific domain-knowledge held by the user. When, however, deciding which attributes to include in this filtered accident set, the user may also utilise the attribute groupings currently stored in Datastore D1. In this way, the user may select an attribute based on its data type (*i.e.* whether the attribute is discrete or continuous), its position in the Haddon matrix (*i.e.* whether the attribute describes the situation before, during or after the accident and whether a human, vehicle or road factor is being considered), or the type of spatial information represented by the attribute (*i.e.* whether the attribute is spatial, semi-spatial or non-spatial).

The sets of filtered explanatory and target variables $(\mathcal{X}_F, \mathcal{Y}_F)$ are then partitioned, in Sub-module 9.2, into two subsets: A random $(100 \cdot \zeta)\%$ of the data forms the *training* data set and the remaining $100(1-\zeta)\%$ forms the *test* data set, denoted by $(\mathcal{X}_L, \mathcal{Y}_L)$ and $(\mathcal{X}_T, \mathcal{Y}_T)$, respectively. The learning data set $(\mathcal{X}_L, \mathcal{Y}_L)$ is then passed to Sub-module 9.3 where the user is first required to select a set of n_H learning models $\mathcal{H} = \{H_1, \ldots, H_{n_H}\}$, like those described in Chapter 4, which can be used to explain the defined target values as a function of the explanatory values. The learning data set is used to tune the hyper-parameter values of a candidate learning model H_i according to a grid-based search technique using either the cross-validation or the bootstrap re-sampling splitting criterion (as described in §4.2.3) so as to find a model which achieves the best performance in respect of the learning data set. Each learning model in $\mathcal{H}$ is assessed with respect to a specified performance measure, as described in §4.3.1, where all (approximately) optimally-tuned models are contained in the set $\mathcal{H}^*$. In Sub-module 9.4, the performance of the (approximately) optimally-tuned learning models $\mathcal{H}^*$ are evaluated in respect of the unbiased test data set $(\mathcal{X}_T, \mathcal{Y}_T)$ so as to obtain a final performance evaluation. Both the learning and testing set results are communicated to the user.

Thereafter, the model-agnostic SHAP method, described in §4.3.3, may be employed so as to facilitate an interpretation of a selected learning model's inner rule formulation. More specifically, a set of SHAP values $\Phi = \{\phi_{11}, \ldots, \phi_{NM}\}$ are computed for all the observations $(\mathcal{X}_F, \mathcal{Y}_F)$ according to the selected learning model H_i^*. This rule formulation, therefore, allows the user to gain an intuitive understanding of which attributes in $\hat{\mathcal{X}}$ are the most informative when describing the response of a target attribute $\hat{Y}$. In this way, a user may characterise the risk of a certain outcome (*i.e.* the target attribute) in terms of the presence or combination of various attributes (*i.e.* the explanatory attribute values).

6.10 Chapter summary

The purpose of this chapter was to provide an in-depth description of the modules comprising the RARE framework proposed in Chapter 5. The chapter opened in §6.1 with a description of how one may facilitate the categorisation of attributes according to their data type, together with assisting in the imputing of missing values in the data sets. This was followed in §6.2 by a discussion on further formatting and attribute categorisation steps whereby attributes could be classified as being either spatial, semi-spatial or non-spatial, and according to the Haddon matrix. The focus then shifted in §6.3 and §6.4 to how one may merge vehicle and road records with a corresponding accident record for the purpose of creating an all-encompassing data set. Thereafter, the method for creating accident severity attributes is described in §6.5, followed, in §6.6, by a description of a procedure for facilitating the construction of a new attribute by the user, based on one or more current attribute value(s). The focus next shifted in §6.7 and §6.8 to a description of novel approach towards quantifying RAR along road segments and at road junctions, respectively. These results were then used in §6.9 as input to a supervised learning procedure for identifying the most indicative factors which contribute to higher levels of RAR.

Part III

Case study

CHAPTER 7

Framework development and data preparation

Contents

7.1 Development of the proposed RARE framework 127
7.2 Accident and road data 128
7.3 Configuring the data 130
7.4 Combining the data 134
7.5 Creating new data 138
7.6 Chapter summary 138

The purpose of this chapter is to serve as a form of validation of Modules 1.0–6.0 in the proposed RARE framework of Figure 5.4 through its application to a real-life case study. The chapter opens in §7.1 with an elucidation of how the proposed RARE framework was implemented, and this is followed in §7.2 by a description of the real-life data employed during this case study. The focus then shifts in §7.3 to how Modules 1.0 and 2.0 were used to configure the data appropriately according to the various attribute groupings. This is followed in §7.4 by an application of Modules 3.0 and 4.0 to construct a combined data set. Thereafter, the creation of new attributes according to the processes described in Modules 5.0 and 6.0 are described in §7.5. The chapter finally closes in §7.6 with a brief summary of its contents.

7.1 Development of the proposed RARE framework

The purpose of this section is to describe how the proposed RARE framework of Part II was implemented for the purpose of conducting a real-life case study. As previously mentioned, the primary aim of this dissertation is to design, develop and demonstrate the practical workability of a generic framework for evaluating risk in the context of RAs. The development of such an framework, therefore, serves as a concept demonstrator of how such a system would function in the presence of real-life data.

The first consideration was to establish an *application programming interface* (API) which could act as the framework basis for implementing the individual modules of the proposed RARE framework. When selecting an existing API, the focus was to ensure that researchers and analysts would be able to alter or append the proposed framework if they deemed this necessary.

The criteria for API selection were two-fold: First, the API should be well-documented open-source software with a specific focus on the data mining process, and, secondly, the API should allow for the development and seamless integration of the proposed modules of the RARE framework. Meta Brown, a prominent researcher in the field of data mining, maintains that the best data mining tool is an application with a visual programming interface [28]. These applications typically present the researcher with a blank *canvas*, together with a number of pre-built *widgets* that the researcher can drag and connect together on the canvas to achieve some data mining task. Visual programming is thus ideal since it allows researchers to gain a holistic picture of their work flow, facilitates rapid prototyping and allows for the easy exchange and reuse of developed widgets.

With these criteria in mind, the Orange Data Mining API (version 3.19) [70] was selected as the framework basis for the development of the proposed RARE framework. This software environment includes the implementation of numerous data preprocessing techniques, ML algorithms and data visualisation capabilities which aided the author in the implementation of the proposed RARE framework. It is, therefore, proposed that the modules of the RARE framework implementation should be in line with the Orange Data mining API so that they may be re-used or re-purposed in other problem application domains in the context of this software environment. A screenshot of the RARE framework concept demonstrator built in the Orange Data Mining Software is shown in Figure 7.1. All purpose-built sub-modules were subsequently developed using the Python 3.7.0 programming language, indicated by the red widgets in the figure, and may be viewed online (please refer to Appendix A for additional information).

7.2 Accident and road data

As a way of validating the effectiveness of the proposed RARE framework of Part II, the concept demonstrator implementation of the proposed framework was tested in respect of data made available by the UK Department of Transport. The data identified for this study comprises a set of accident, vehicle, road and victim data for Greater Manchester in the *United Kingdom* (UK) over the period 2005–2017 [275, 277]. Further insights into the values that each categorical attribute may assume, together with the proportion of accidents with a specific categorical attribute value, may be found in Appendix B. These data were selected for three main reasons, namely: (1) Both accident and traffic flow data are publicly available, (2) the data were deemed of a high quality, and (3) the data conformed to the entity relationship assumed in the RARE framework, as described in Figure 5.3.

The process by which the accident, vehicle and victim data are collected in the UK was first investigated so as to verify the quality of the data considered for this validation step. If a road user is involved in an RA in the UK, a police officer is required to attend the scene of the accident and complete a STATS19 accident report form (this form is provided in Appendix C) [278]. Detailed instructions for the appropriate completion of this form are stipulated in the STATS20 manual, while the STATS21 documentation describes the process of electronically transferring the data into the national RA database [279]. The latitude and longitude position of an accident is electronically captured at the first point of impact on the road or, in the case where an accident impact occurs after leaving the road, it is captured at the point at which the vehicle first left the road. In light of this information, the accident data employed in this case study was deemed of a high quality. For a more detailed description of the nomenclature used by the UK Department for Transport, the reader is referred to Appendix D.

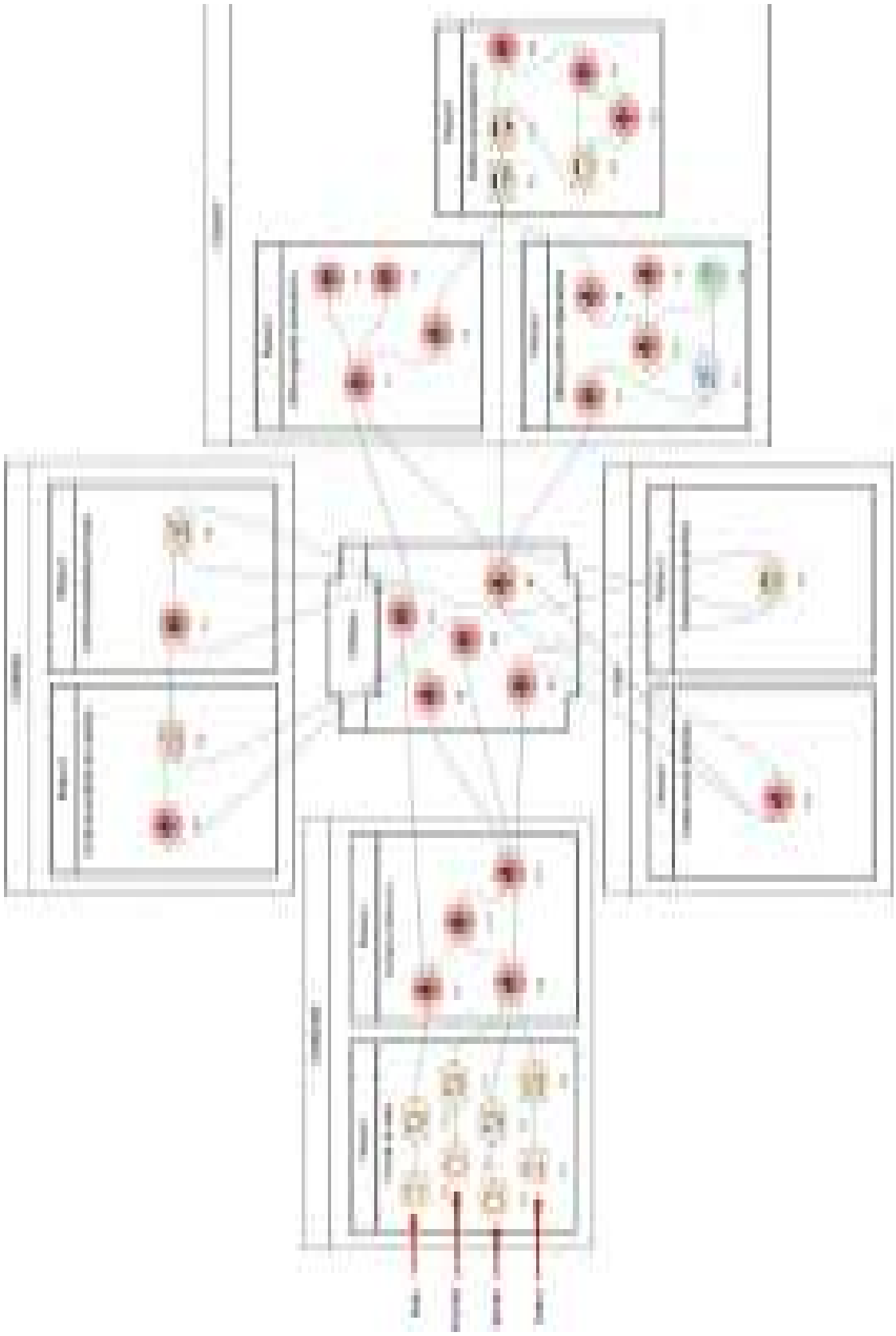


FIGURE 7.1: *The RARE framework concept demonstrator built in the Orange Data Mining Software. The red widgets indicate the sub-modules of the RARE framework developed by the author in the Python 3.7.0 programming language.*

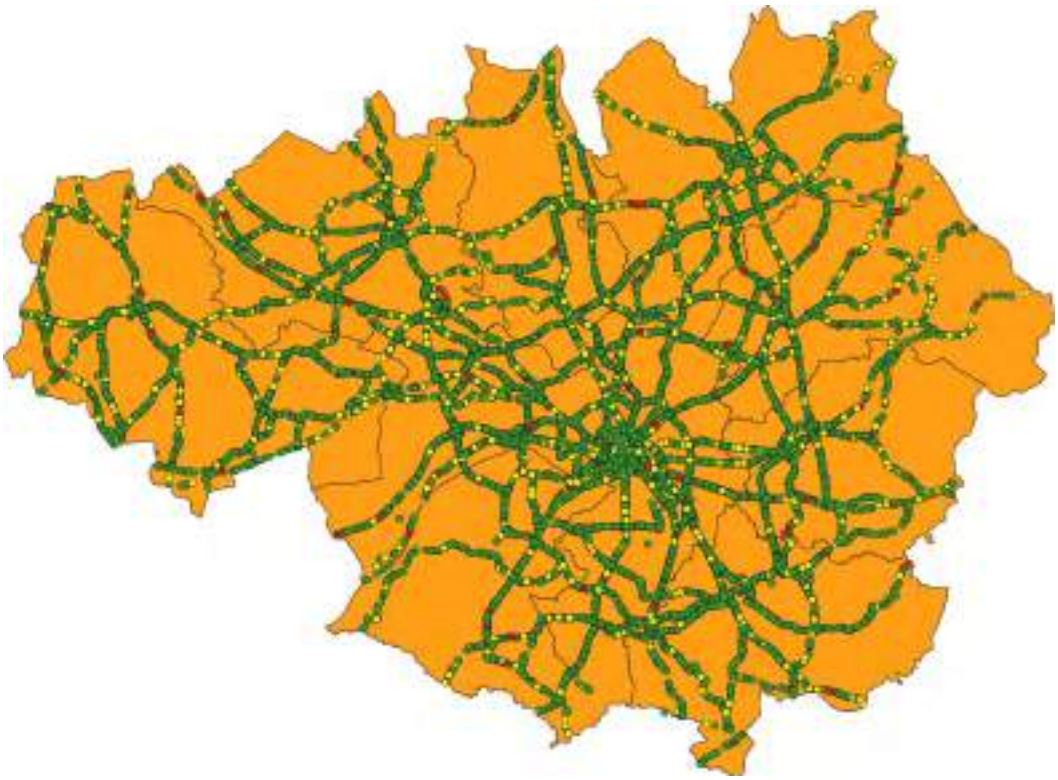
A geographical representation of the Greater Manchester accident data set is shown in Figure 7.2(a), and consists of $m_A = 37\,795$ accident records with $m_V = 73\,852$ associated vehicle records and $m_T = 54\,843$ associated victim records. The green, yellow an red dots in the figure indicate RAs in which the maximum sustained victim injury was "Slight," "Serious" and "Fatal," respectively. According to the UK Department for Transport, fatal injuries include only those cases where death occurs no more than 30 days after the accident [279]. Victims are classified as seriously injured if they sustain one or more of the following injuries: Broken neck or back, severe head injury (or unconscious), severe chest injury, any difficulty breathing, internal injuries, multiple severe injuries (or unconscious), loss of arm or leg (or part), other chest injury (not bruising), deep penetrating wound, fracture, deep cuts/lacerations, other head injury, crushing, burns (excluding friction burns), concussion, severe general shock requiring hospital treatment, detention in hospital as an in-patient (either immediately or later), or injuries to victims who die 30 or more days after the accident from injuries sustained during that accident [279]. Victims are classified as slightly injured if they sustain one or more of the following injuries: Whiplash or neck pain, shallow cuts/lacerations/abrasions, sprains and strains (not necessarily requiring medical treatment), bruising or slight shock requiring roadside attention [279]. Persons who were merely shaken and did not sustain other injuries are not classified as slightly injured unless they receive or appear to require medical treatment [279]. An accident is, therefore, only recorded by the UK Department for Transport if at least one involved person sustained some form of injury.

The road data obtained for this study only contain AADT measures for type A and A(M) roads as well as motorways. Consequently, only accidents occurring on these road types were considered during this investigation. The road data set is illustrated geographically in Figure 7.2(b), and comprises $m_R = 755$ road segments with AADT data collected over the period 2005–2017, where red and dark green segments indicate roads with high and low vehicle counts, respectively. As may be seen in the figure, motorways (*e.g.* the M60 motorway circling Manchester) have significantly higher traffic flows, indicated by the red line segments, than the other road types. Edge biases, as discussed in §3.2.5, were assumed negligible since all major urban settings were located well within the artificial study region boundary.

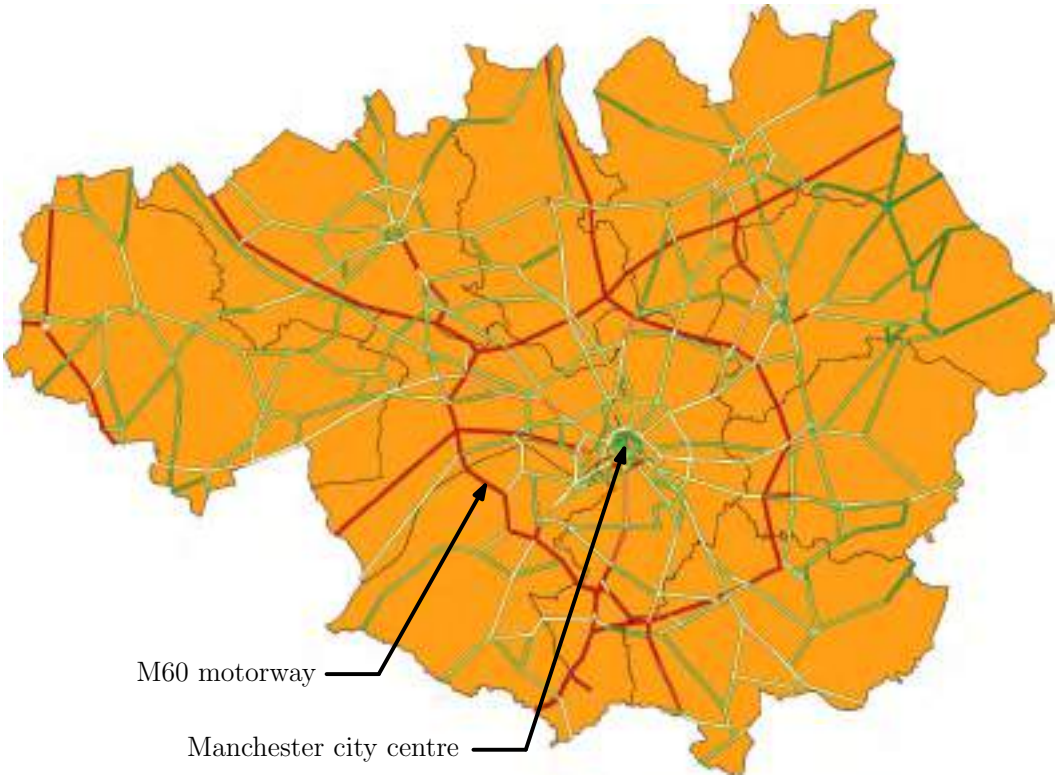
7.3 Configuring the data

After obtaining the necessary accident, vehicle, road and victim data sets, the first consideration was an appropriate configuration of the data through attribute categorisation and data cleaning. The first categorisation step was to configure the data sets according to their data type, as per Sub-module 1.1. Attributes were classified as either categorical, continuous or a meta-attribute, as indicated in Tables 7.1–7.4. Thereafter, the first five rows of compulsory attributes described in Table 6.1 were identified. The unique accident identifier was taken as the attribute $\hat{A}_I =$ `AccidentIndex`, which was subsequently linked to the attribute $\hat{V}_L =$ `AccidentIndex` in the set of vehicle attributes and the attribute $\hat{T}_L =$ `AccidentIndex` in the set of victim attributes. Furthermore, the vehicle and victim attributes $\hat{V}_I =$ `VehicleReference` and $\hat{T}_I =$ `VictimReference` were taken as the unique vehicle and victim identifiers, respectively. Lastly, the road AADT attribute was defined as $\hat{R}_T =$ `AADT`, the severity attribute as $\hat{T}_S =$ `VictimSeverity` and the junction attribute as $\hat{A}_J =$ `JunctionControl`.

The percentages of missing values contained in the accident, vehicles, roads and victims data sets were then ascertained, the results of which may also be found in Tables 7.1–7.4, respectively. From these results, it was concluded that the data sets are, for the most part, complete. It was, however, surprising to discover that 40.44% of the `JunctionControl` at-



(a)



(b)

FIGURE 7.2: *A geographic representation of the 2005–2017 Greater Manchester (a) accidents data set and (b) roads data set.*

TABLE 7.1: *The attributes and data types of the accident data set.*

Accidents			
No.	**Attribute**	**Data type**	**% missing**
1	AccidentIndex	Meta	0.00%
2	Longitude	Continuous	0.00%
3	Latitude	Continuous	0.00%
4	AccidentSeverity	Categorical	0.00%
5	NumberOfVehicles	Continuous	0.00%
6	Date	Meta	0.00%
7	DayOfWeek	Categorical	0.00%
8	Time	Meta	0.00%
9	FirstRoadClass	Categorical	0.00%
10	RoadType	Categorical	0.00%
11	SpeedLimit	Categorical	0.00%
12	JunctionDetail	Categorical	0.00%
13	JunctionControl	Categorical	40.44%
14	HumanPedestrianCrossing	Categorical	0.00%
15	PhysicalPedestrianCrossing	Categorical	0.00%
16	LightConditions	Categorical	0.00%
17	WeatherConditions	Categorical	0.32%
18	RoadSurfaceConditions	Categorical	0.00%
19	SpecialConditionsAtSite	Categorical	0.00%
20	CarriagewayHazards	Categorical	0.00%
21	UrbanOrRuralArea	Categorical	0.00%

TABLE 7.2: *The attributes and data types of the vehicle data set.*

Vehicles			
No.	**Attribute**	**Data type**	**% missing**
1	AccidentIndex	Meta	0.00%
2	VehicleReference	Meta	0.00%
3	VehicleType	Categorical	0.00%
4	TowingAndArticulation	Categorical	0.00%
5	VehicleManoeuvre	Categorical	0.00%
6	VehicleLocationRestrictedLane	Categorical	0.00%
7	JunctionLocation	Categorical	0.00%
8	SkiddingAndOverturning	Categorical	0.00%
9	HitObjectInCarriageway	Categorical	0.00%
10	VehicleLeavingCarriageway	Categorical	0.00%
11	HitObjectOffCarriageway	Categorical	0.00%
12	FirstPointOfImpact	Categorical	0.00%
13	WasVehicleLeftHandDrive	Categorical	3.41%
14	JourneyPurposeOfDriver	Categorical	68.33%
15	SexOfDriver	Categorical	5.70%
16	AgeOfDriver	Continuous	6.55%
17	EngineCapacityCC	Continuous	25.08%
18	PropulsionCode	Categorical	24.53%
19	AgeOfVehicle	Continuous	28.25%
20	DriverIMDDecile	Categorical	25.55%
21	DriverHomeAreaType	Categorical	21.28%

TABLE 7.3: *The attributes and data types of the road data set.*

Roads			
No.	**Attribute**	**Data type**	**% missing**
1	`RoadIndex`	Meta	0.00%
2	`AADT`	Continuous	0.00%
3	`RoadLength`	Continuous	0.00%
4	`StartLatitude`	Continuous	0.00%
5	`StartLongitude`	Continuous	0.00%
6	`EndLatitude`	Continuous	0.00%
7	`EndLongitude`	Continuous	0.00%

TABLE 7.4: *The attributes and data types of the victim data set.*

Victims			
No.	**Attribute**	**Data type**	**% missing**
1	`AccidentIndex`	Meta	0.00%
2	`VehicleReference`	Meta	0.00%
3	`VictimReference`	Meta	0.00%
4	`VictimSeverity`	Categorical	0.00%
5	`VictimClass`	Categorical	0.00%

tribute values were missing. Upon further investigation, it was discovered that, in the majority of these instances, this attribute was intentionally left blank by the police officer for accidents that did not occur at a junction. All missing values for this attribute, which also possess a `JunctionDetail` attribute value of `NotAtJunctionOrWithin20Metres`, were subsequently replaced with the category `NotAtJunctionOrWithin20Metres`. In the vehicles data set, attributes containing more than 5% missing values were discarded altogether (*i.e.* the `JourneyPurposeOfDriver`, `SexOfDriver`, `AgeOfDriver`, `EngineCapacityCC`, `PropulsionCode`, `AgeOfVehicle`, `DriverIMDDecile` and `DriverHomeAreaType` attributes). Upon closer examination of the `WasVehicleDriverLeftHanded` attribute, it was found that 99.65% of all vehicle records contained the value of `No` for this attribute. This statistic is somewhat surprising since a recently conducted survey found that approximately 11% of people living in the UK are left handed [50]. This result suggests that, during the data collection process, a police officer likely assumed that the driver's dominant hand is right if he or she was not entirely certain. Consequently, this attribute was also excluded from the vehicles data set. All records which contained missing attribute values were subsequently discarded from the data set. At this point, the procedure of Sub-module 1.2 was deemed complete and each respective data set was considered appropriately formatted.

A single representative record was then created for each road record according to the procedure outlined in Sub-module 2.1. The four steps of Sub-module 2.2 for removing associated records (which may have been removed in Sub-module 1.2) was then completed. After this procedure, $m_A = 35\,874$ of the original accident records remained, together with $m_V = 70\,198$ of the associated vehicle records and $m_T = 52\,323$ victim records.

The next step in the attribute categorisation phase, stipulated in Sub-module 2.3, required the identification of spatial attributes currently present in the data which were used to define the geographical plots in Figure 7.2. During this stage, the accident attributes $\hat{A}_x =$ `Latitude` and $\hat{A}_y =$ `Longitude` signifying the point of an accident, were defined, together with the road

attributes $\hat{R}_{x_1}$ = `StartLatitude`, $\hat{R}_{y_1}$ = `StartLongitude`, $\hat{R}_{x_2}$ = `EndLatitude` and $\hat{R}_{y_2}$ = `EndLongitude`, which indicate the start and end points of road segments. These attributes were appropriately identified as the set of spatial attributes. The remaining attributes were all initially deemed non-spatial in nature. In this way, all compulsory attributes stipulated in Table 6.1 were appropriately defined in the context of this case study. The spatial scores (*i.e.* the quantity S_j for attribute j) of all accident attributes were then computed and are shown Figure 7.3. A more detailed breakdown of the individual category scores may be found in Appendix E. A spatial threshold of $\gamma = 0.45$ was then chosen. Consequently, a total of seven semi-spatial attributes were defined, namely `FirstRoadClass`, `RoadType`, `SpeedLimit`, `JunctionDetail`, `JunctionControl`, `PhysicalPedestrianCrossing` and `UrbanOrRuralArea`. All other attributes remained non-spatial in nature.

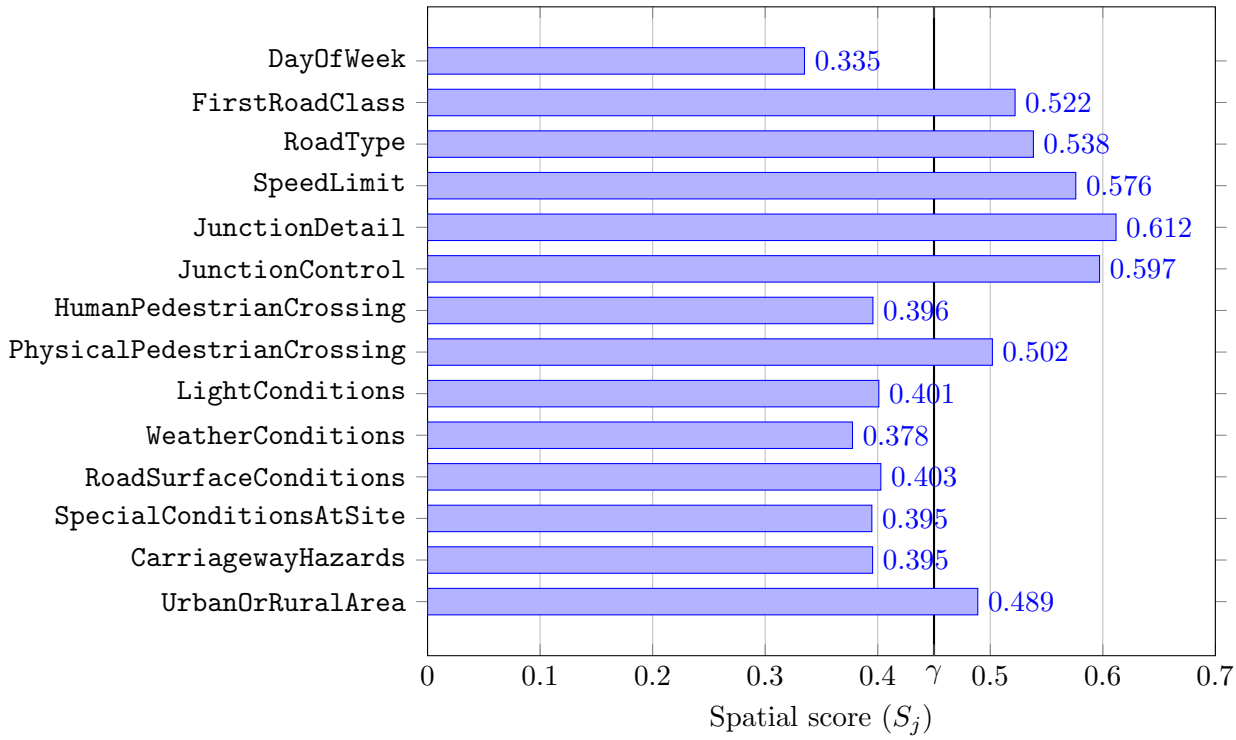


FIGURE 7.3: *The spatial scores achieved by all accident attributes.*

During the final categorisation step, stipulated in Sub-module 2.4, all non-spatial attributes in the formatted accident, vehicles, roads and victims data sets were classified according to the Haddon matrix, as shown in Table 7.5. The `WeatherConditions` attribute was, for example, placed in the cell defined by the *before* (pre-accident) row and *environment* column in the Haddon matrix since it describes an *environmental* factor which can first be labelled in the time period *before* the accident's occurrence. Upon completion of this final step, the accident, vehicle, road and victim data were considered appropriately configured.

7.4 Combining the data

After all data sets had been configured appropriately and stored in a database, the next step required the merging of vehicle and road records with accident records according to the procedures described in §6.3 and §6.4, respectively. The first step, stipulated in Sub-module 3.1,

Table 7.5: *Classification of attributes according to the Haddon matrix.*

	Human	Vehicle	Environment
Pre-accident		1. `VehicleType` 2. `TowingAnd-Articulation`	1. `Date` 2. `DayOfWeek` 3. `Time` 4. `FirstRoadClass` 5. `RoadType` 6. `SpeedLimit` 7. `JunctionDetail` 8. `JunctionControl` 9. `HumanPedestrian-Crossing` 10. `PhysicalPedestrian-Crossing` 11. `LightConditions` 12. `WeatherConditions` 13. `RoadSurface-Conditions` 14. `SpecialConditions-AtSite` 15. `CarriagewayHazards` 16. `UrbanOrRuralArea` 17. `AADT` 18. `RoadLength`
Accident	1. `VictimClass`	1. `NumberOfVehicles` 2. `VehicleManoeuvre` 3. `VehicleLocation-RestrictedLane` 4. `JunctionLocation` 5. `SkiddingAnd-Overturning` 6. `FirstPointOf-Impact`	1. `HitObjectIn-Carriageway` 2. `HitObjectOff-Carriageway`
Post-accident	1. `VictimSeverity`	1. `VehicleLeaving-Carriageway`	

required the ordering of the set of vehicle records associated with an accident record according to Algorithm 6.1. As described in §6.3, a vector $\boldsymbol{\rho}$ must first be specified, stipulating the order (or precedence) in which vehicle attributes should be used to order the rows of a set of vehicle records. In this case study, this array was chosen as $\boldsymbol{\rho} = \langle 3\ 4\ 7\ 6\ 5\ 12\ 9\ 8\ 10\ 11 \rangle$ corresponding to the attributes `VehicleType`, `TowingAndArticulation`, `JunctionLocation`, `VehicleLocationRestrictedLane`, `VehicleManoeuvre`, `FirstPointOfImpact`, `HitObjectInCarriageway`, `SkiddingAndOverturning`, `VehicleLeavingCarriageway` and `HitObjectOffCarriageway`, respectively, in Table 7.2. This order was chosen in the sequential order that events occur during the course of an RA.

The frequency of the orderings accomplished using Algorithm 6.1 were tracked according to which of the aforementioned attributes resulted in the collapse of the recursive function (*i.e.* the value of r if Step 4 in Algorithm 6.1 evaluates to true) for the purpose of providing insight into the functioning of the proposed algorithm. The results of this investigation may be found in Figure 7.4, where the bottom bar plot (labelled "All attribute values equal") indicates those cases in which two or more vehicle records had the exact same values for all of the aforementioned vehicle attributes (*i.e.* if Step 1 in Algorithm 6.1 evaluates to true). From this plot, one may conclude that just over half of all vehicle records were ordered according to their `VehicleType` attribute value. Attributes with a large proportion of accidents of a specific categorical attribute value (as illustrated in Appendix B) resulted in lower vehicle ordering counts, as may be expected. After all vehicle records had been ordered appropriately, they were reduced to single, representative vehicle records and concatenated to their respective accident records, as stipulated in Sub-module 3.2.

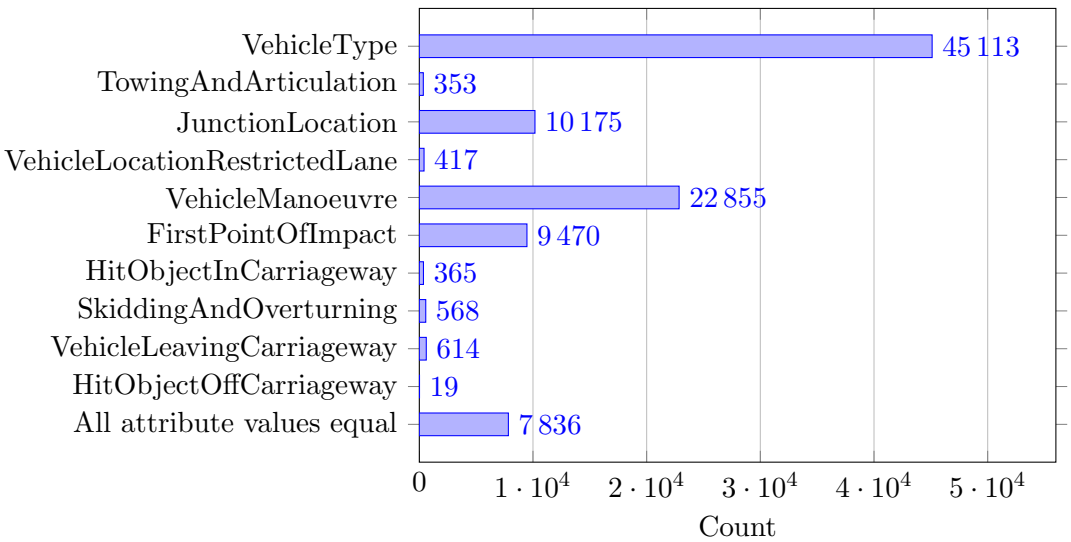
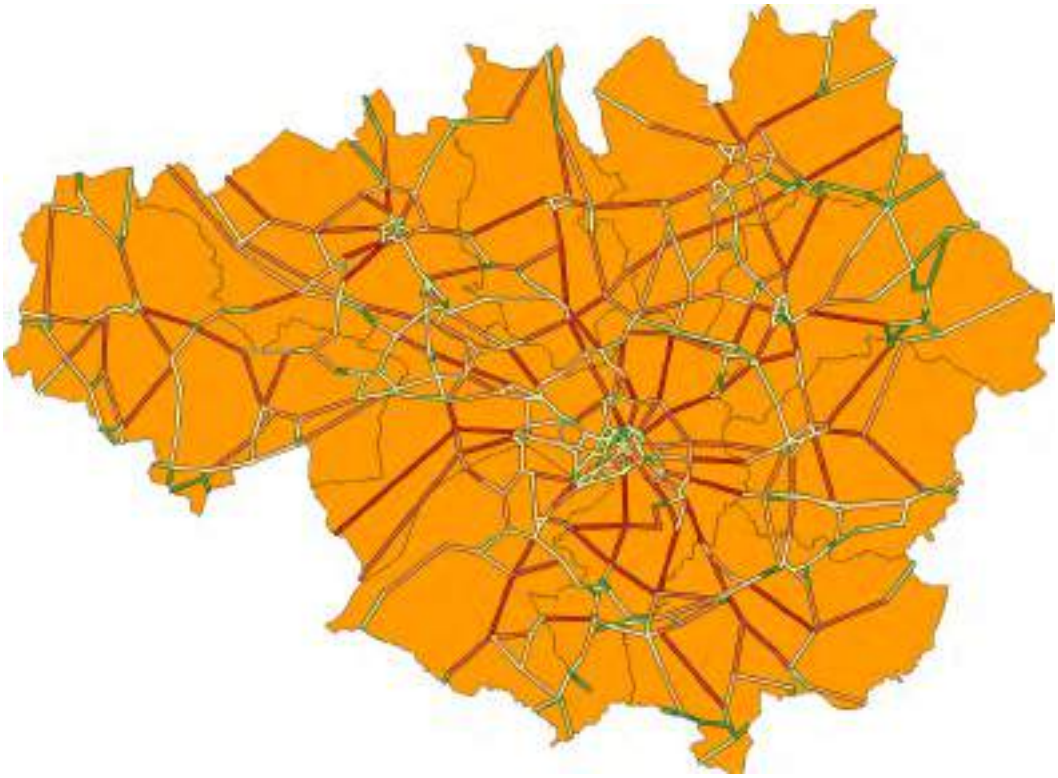
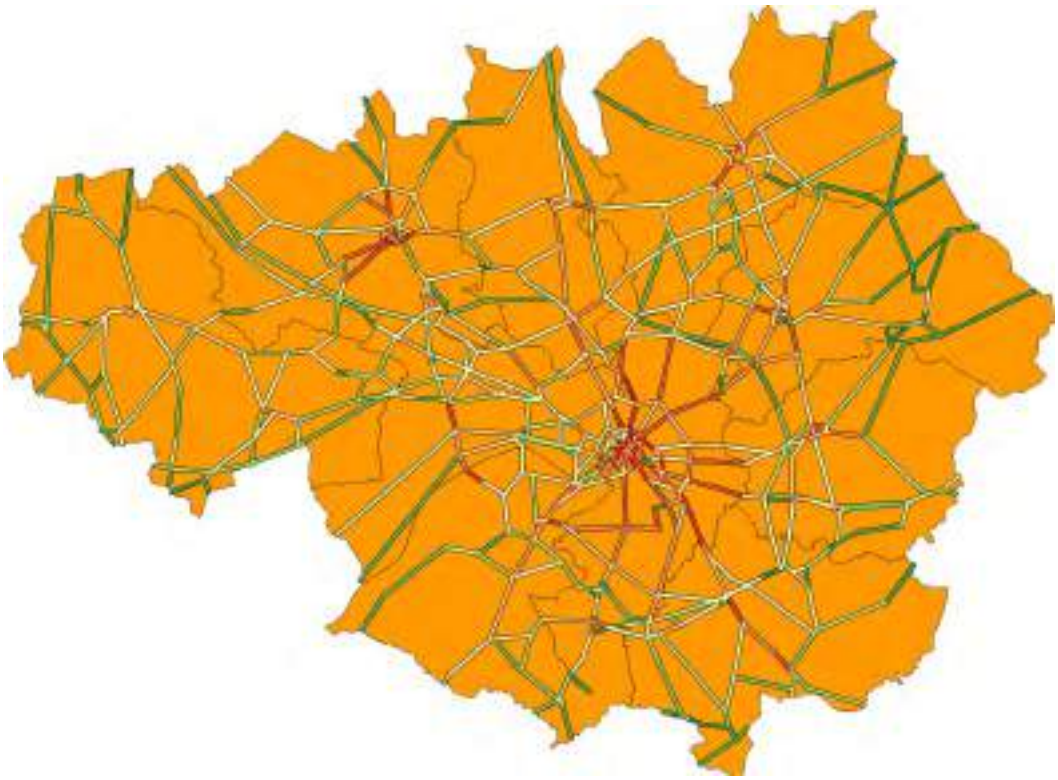


FIGURE 7.4: *The frequency of collapses in the recursive function of Algorithm 6.1, according to vehicle attributes.*

After all vehicle records had been merged with their corresponding accident records, each accident record was associated with a corresponding road record, as stipulated in Sub-module 4.1. The association of accident points with road segments was achieved by employing Algorithm 6.2, as described in §6.4. A geographical visualisation of the results obtained through the application of this algorithm may be found in Figure 7.5. As a way of verifying the intended functionality of Algorithm 6.2, the number of accident points associated with each road segment in the network is shown in Figure 7.5(a), where red road segments indicate high accident counts and green road segments denote low accident counts. The accident counts in Figure 7.5(a) are normalised



(a)



(b)

FIGURE 7.5: *A geographical representation of roads classified according to (a) accident count and (b) accident count normalised by road length.*

according to the length of a road segment in Figure 7.5(b). One would expect longer road segments to contain more RAs, while urbanised areas (like Manchester city centre) should contain a larger number of accidents per road segment length — this is indeed observed in Figure 7.5(a) and (b), respectively. Algorithm 6.2 was, therefore, deemed to be a useful tool for associating accident records with appropriate road records.

7.5 Creating new data

In accordance with the procedures stipulated in Sub-module 5.0, a binary categorical severity attribute was created, called `CategoricalSeverity` $\in \hat{\mathcal{X}}_{\text{cat}}$, where and accident was classified as `MoreSevere` if any of the associated victim records had the value `Serious` or `Fatal` for the `VictimSeverity` attribute, or `LessSevere` otherwise. Furthermore, it has been estimated that, as of 2017, the average cost in *Great British Pounds* (GBPs) per victim sustaining a `Slight`, `Serious` and `Fatal` injury is £16 434, £213 184 and £1 897 129 [250], respectively. This allowed for the construction of the weight vector $\boldsymbol{m}$ and thus a continuous severity attribute, called `ContinuousSeverity` $\in \hat{\mathcal{X}}_{\text{cont}}$, according to the procedure described in §6.5.

During the final step of the data preparation stage, two new attributes were constructed by means of defined algorithmic procedures, as stipulated in Sub-module 6.0. The first created attribute was a binary categorical attribute, called `PedestrianAccident`, which took the value `Yes` if any of the associated victim records had value of `Pedestrian` for the `VictimClass` attribute, or `No` otherwise. Since the `VictimClass` attribute was categorised as a during-human attribute in the Haddon matrix, the new `SevereAccident` $\in \hat{\mathcal{X}}_{DH}$ attribute also inherited this property. The second attribute created was a twenty four class categorical attribute, called `TimeBand`, which classifies an accident into hourly bins according to its `Time` attribute (*i.e.* if an accident occurred at 14:35, it is classified as occurring during the interval 14:00–14:59). Again, since the `Time` attribute was categorised as a before-environment attribute in the Haddon matrix, the new `TimeBand` $\in \hat{\mathcal{X}}_{BE}$ attribute also inherited this property. After the construction of these new attributes, as stipulated in Sub-module 6.0, the set of combined data was updated. At this point, the combined data stored in the database was deemed of a high quality and considered acceptable for conducting further analysis.

7.6 Chapter summary

The purpose of this chapter was to demonstrate how each of the data preparation sub-modules of the proposed RARE framework would function in the context of real-life data. The chapter opened in §7.1 with a description of how a concept demonstrator of the proposed RARE framework of Part II was implemented for the purpose of conducting a real-life case study. This was followed in §7.2 with a description of the Greater Manchester accident, vehicle, road and victim data sets employed during this case study. The focus then shifted in §7.3 to how attributes were appropriately categorised, together with how missing values in the data were imputed appropriately. Thereafter, the application and verification of the two proposed algorithms for merging vehicle and road records with their corresponding accident records were described in §7.4. The chapter finally concluded in §7.5 with a description of how additional attributes were created and appended to the combined data set.

CHAPTER 8

Data modelling and evaluation

Contents

8.1 Analysis of road accident risk along road segments 140

8.1.1 Parameter choices for descriptive analysis of road segments 140

8.1.2 Spatial analysis of road segments 141

8.1.3 Non-spatial analysis of road segments 147

8.1.4 Parameter choices for predictive analysis of road segments 150

8.1.5 Predictive analysis of road segments 151

8.2 Analysis of road accident risk at road junctions 155

8.2.1 Parameter choices for descriptive analysis of road junctions 155

8.2.2 Spatial descriptive analysis of road junctions 156

8.2.3 Non-spatial descriptive analysis of road junctions 156

8.2.4 Parameter choices for predictive analysis of road junctions 160

8.2.5 Predictive analysis of road junctions 162

8.3 Predictive analysis of road accident severity 163

8.3.1 Road accident severity at crossroads controlled by traffic lights 166

8.3.2 Road accident severity on a motorway 169

8.4 Chapter summary 170

The purpose of this chapter is to serve as a form of validation of Modules 7.0–9.0 in the proposed RARE framework by conducting multiple investigations in the context of the real-life data prepared in the previous chapter. In order to demonstrate the intended working of the three modules comprising the Convert class in the proposed RARE framework, numerous investigative questions (indicated by italicised sentences) are posed throughout this chapter. Through these demonstrations, the hope is that the reader is provided with insight into how a user may conduct his or her own investigations according the RARE framework in the context of real-world RA data. The chapter opens in §8.1 with a descriptive and predictive analysis of RAR experienced along road segments due to various road characteristics. This is followed in §8.2 by a similar investigation into RAR experienced at road junctions. Thereafter, a final investigation is carried out in §8.3, where observations are neither road segments nor junctions, but rather accidents themselves. The chapter finally closes in §8.4 with a brief summary of its contents.

8.1 Analysis of road accident risk along road segments

This first section is primarily concerned with demonstrating how Modules 7.0 and 9.0 may be used to answer the following overarching investigative question: *What road characteristics influence the rate and severity of RAs along road segments?* As previously described in §6.7, Module 7.0 aims to answer this question by providing descriptive insights through viewing results it produces in both a spatial and non-spatial manner. First and foremost, the parameter choices made during the construction of road segment observations in Module 7.0 are stated and substantiated in §8.1.1. The focus then shifts in §8.1.2 to quantifying the accident rate and severity along the road segments in Greater Manchester and interpreting these results in terms of the spatial representation produced by Sub-module 7.2. These findings are supported in §8.1.3, by interpreting the same results from a non-spatial perspective by employing Sub-module 7.3. Attention then shifts in §8.1.4 to demonstrating how Module 9.0 may be employed to obtain a predictive analytic view of the results, which begins by first describing the parameter choices associated with Module 9.0. This is followed, in §8.1.5, by an assessment of the predictive performance of six learning models, together with an interpretation of the SHAP values of a selected learning model.

8.1.1 Parameter choices for descriptive analysis of road segments

As described in §6.7, Sub-module 7.1 requires that a base kernel function $K(x)$ with associated bandwidth ω is specified. In this investigation, the well-known Epanechnikov kernel function $K(x) = \frac{3}{4}(1-x^2)$ was employed as an appropriate base kernel function as it is optimal in a mean-square error sense and has the lowest loss of efficiency [170]. A bandwidth of $\omega = 50$ metres was chosen as it is the maximum distance considered by the UK Department for Transport when determining whether a road crossing may have contributed to an accident [279]. As previously described in §6.7, the quantity ψ must also be chosen such that the variance in road segment observation lengths is sufficiently small and so that $\mathcal{G}$ contains an adequate number of road segment observations. The trade-off between these two considerations is shown in Figure 8.1 as

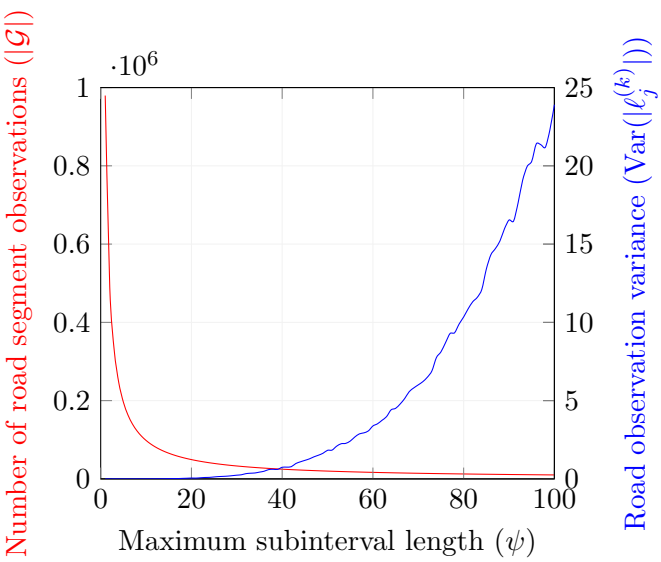


FIGURE 8.1: *The effects of the maximum subinterval length choice on the number of road segment observations and the variance in road segment observation length.*

a function of the maximum subinterval length for this case study. As may be seen in the figure, as the quantity ψ approaches zero, the road segment observation variance also approaches zero, but the number of road segment observations grows significantly large — this would likely lead to large learning model computation times. Conversely, if ψ is chosen too large, the variance in road segment observation lengths becomes significantly large (leading to biases) and too few road segment observations are defined. In this study, a value of $\psi = 20$ metres was chosen as an appropriate trade-off between these two considerations, coupled with the fact that a junction is not considered, by the UK Department for Transport, to have contributed to an accident if the accident was not within 20 metres of the junction [279].

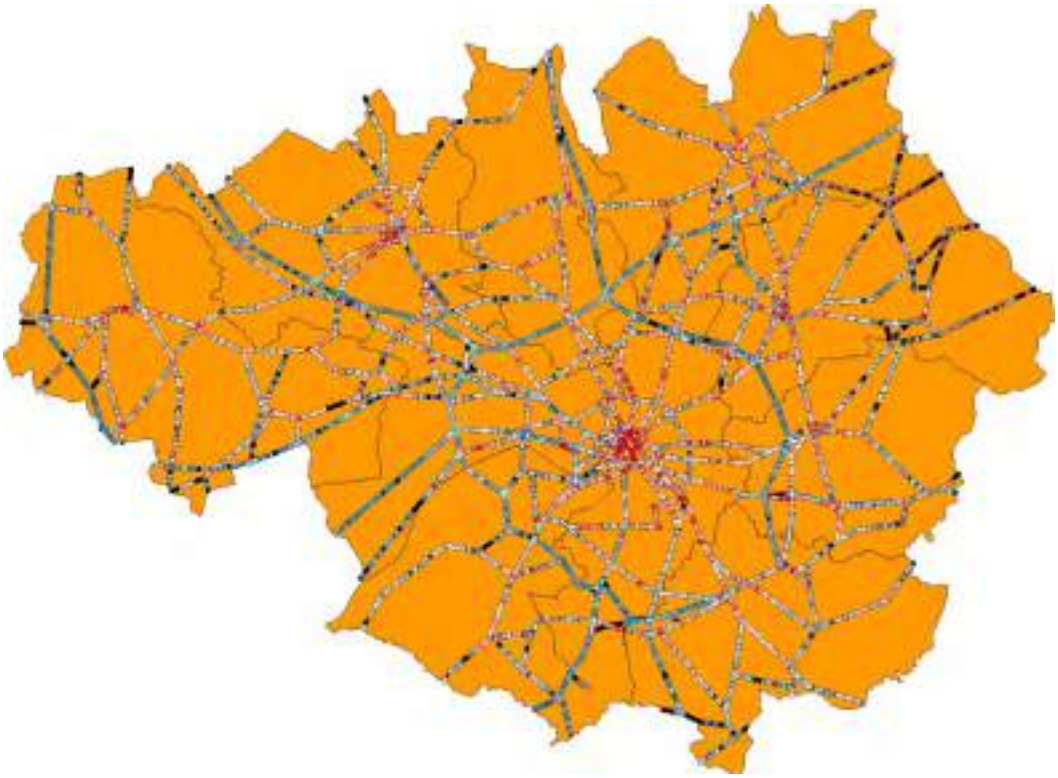
8.1.2 Spatial analysis of road segments

The first step in this investigation was to quantify the accident rate and severity measures along the defined road segment observations. A geographical representation of these results for the Greater Manchester road network are illustrated in Figures 8.2(a) and (b), respectively, where red lines represent roads with higher accident rates/severities and blue lines represent roads with lower accident rates/severities[1]. Sub-module 7.2 may, therefore, be used to provide insight into the question: *Which road segments exhibit a significantly high accident rates and severities in Greater Manchester?* From these results, one may deduce that significantly higher accident rates appear to occur in Manchester city centre, which may be due to the higher density of roads (and thus more junctions). In contrast, the M60 motorway circling Manchester appears to exhibit somewhat lower rates of road accidents, which is likely due to its high traffic volumes and lack of traffic conflict points. This finding is also consistent with the conclusion of the Road Safety Foundation which has rated the M60 motorway as one of the safest roads in Great Britain [266]. In the case of accident severity, it is not entirely clear whether one can draw any preliminary insights from the spatial representation in Figure 8.2(b), although there does appear to be a slightly higher accident severity in the Manchester city centre and along some sections of the M60 motorway. These results may also serve as the starting point for answering the question: *What is the safest route when travelling from point A to point B in Greater Manchester?*

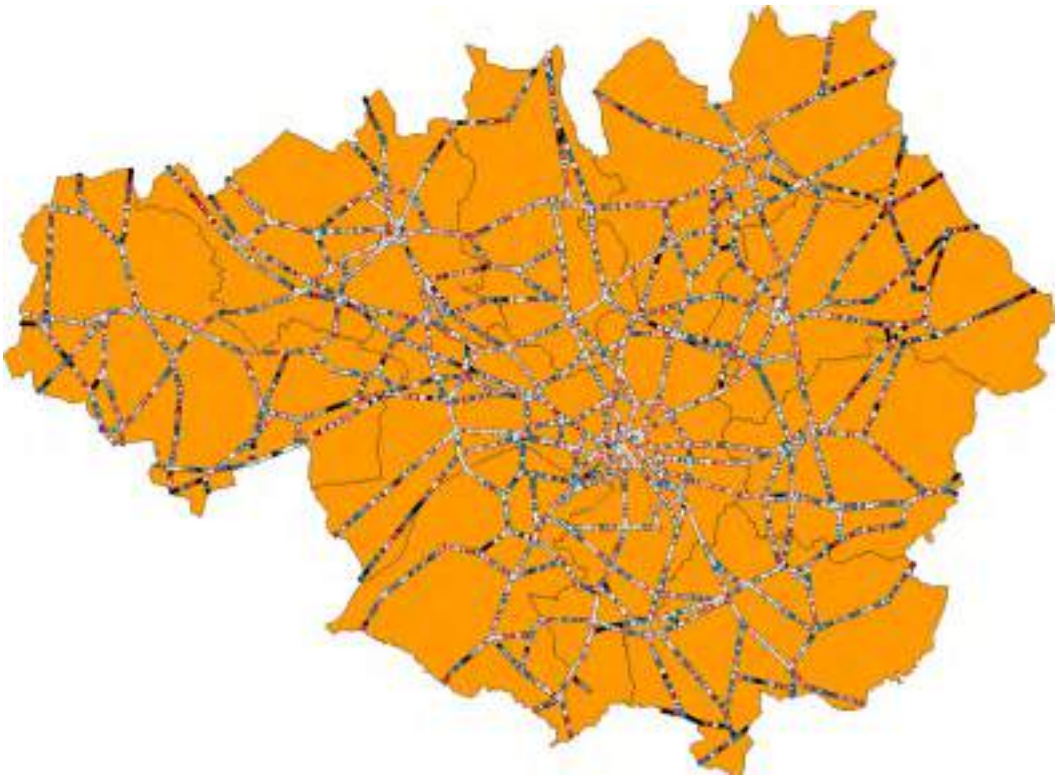
Given this information, an analyst may also conduct his or her own lower-level investigation as to the possible causes of high accident rates or severities at specific locations on the road. Consider, as an example, the enlarged version of the RA severity results of Figure 8.2(b) for the Manchester city centre shown in Figure 8.3(b) for which an accompanying map is also provided in Figure 8.3(a), where red and blue circles denote accidents with a value of `Yes` and `No` for the `PedestrianAccident` attribute (constructed in §7.5), respectively. In Figure 8.3(b), it is clear that a high accident severity is realised at specific points in the city centre, presumably due to the large number of pedestrian-related accidents. One may, for example, ask the question: *Why is there a large number of pedestrian-related accidents at the intersection of Bridge Street (A34) and Deansgate (A56)?* This specific location is circled in green in Figure 8.3(a).

Presented with a real-life depiction of the scene, as illustrated in Figure 8.4, a domain knowledge expert may perform logical reasoning to form a hypothesis as to the likely causes of pedestrian-related accidents at this location. With reference to Figure 8.4, one may, for example, argue that the position of the traffic light (labelled "A") and faded stop road marking (labelled "B") on the A34 are situated abnormally far from the intersection, which, during the traffic light's red phase, requires a driver to stop at the position labelled "C". In this position, the field of vision to the left and right of the driver is obstructed by buildings (labelled "D" and "E"). A driver may unwittingly be inclined to pass the faded stop road marking and continue to the position

[1] Black lines in Figure 8.2 represent road segment observations an accident frequency of zero.

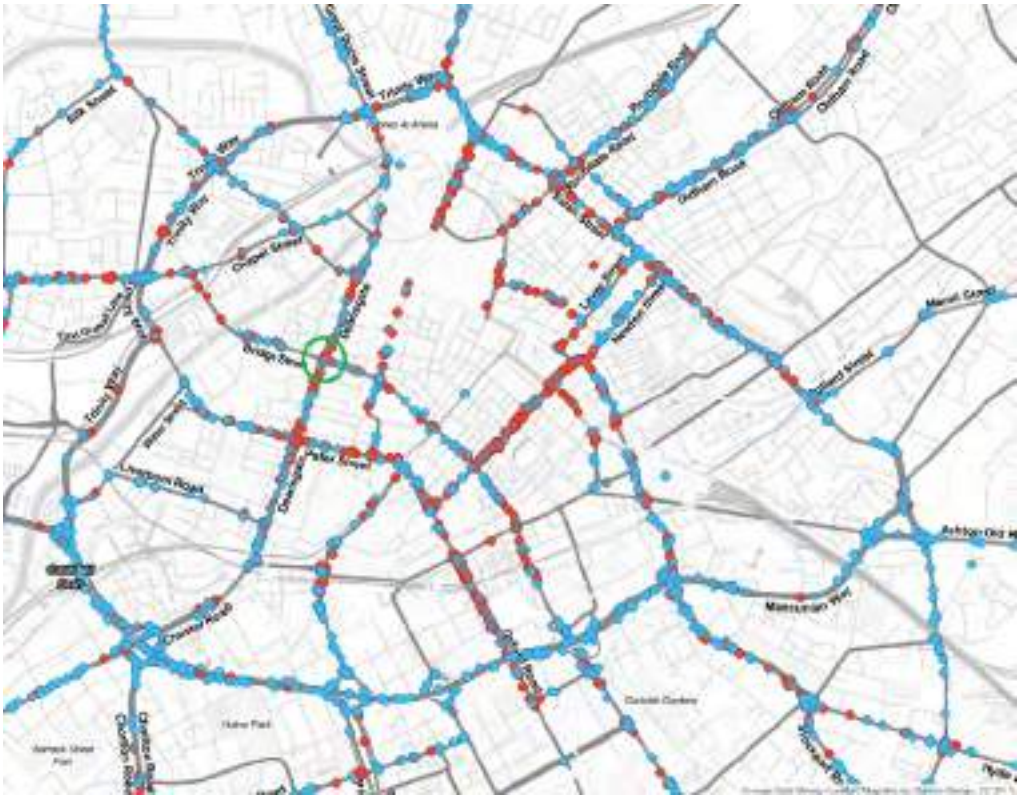


(a)



(b)

FIGURE 8.2: *A spatial representation of the Greater Manchester (a) accident rate and (b) accident severity measurements along road segments.*



(a)

(b)

FIGURE 8.3: *(a) A map of the Manchester city centre and (b) and a corresponding enlarged version of accident severity results shown in Figure 8.2(b).*

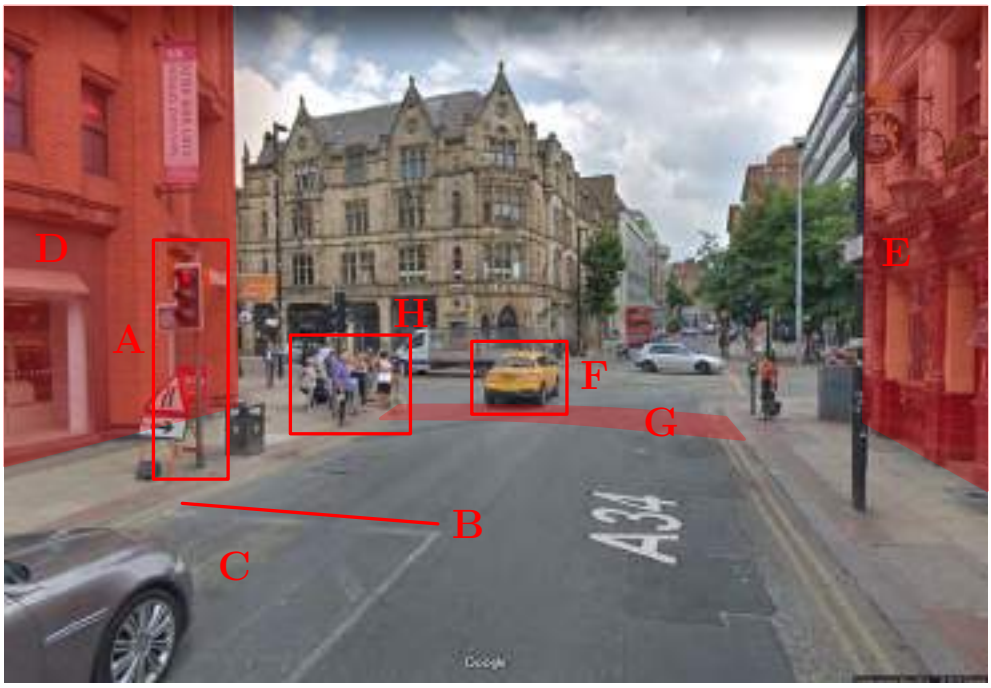
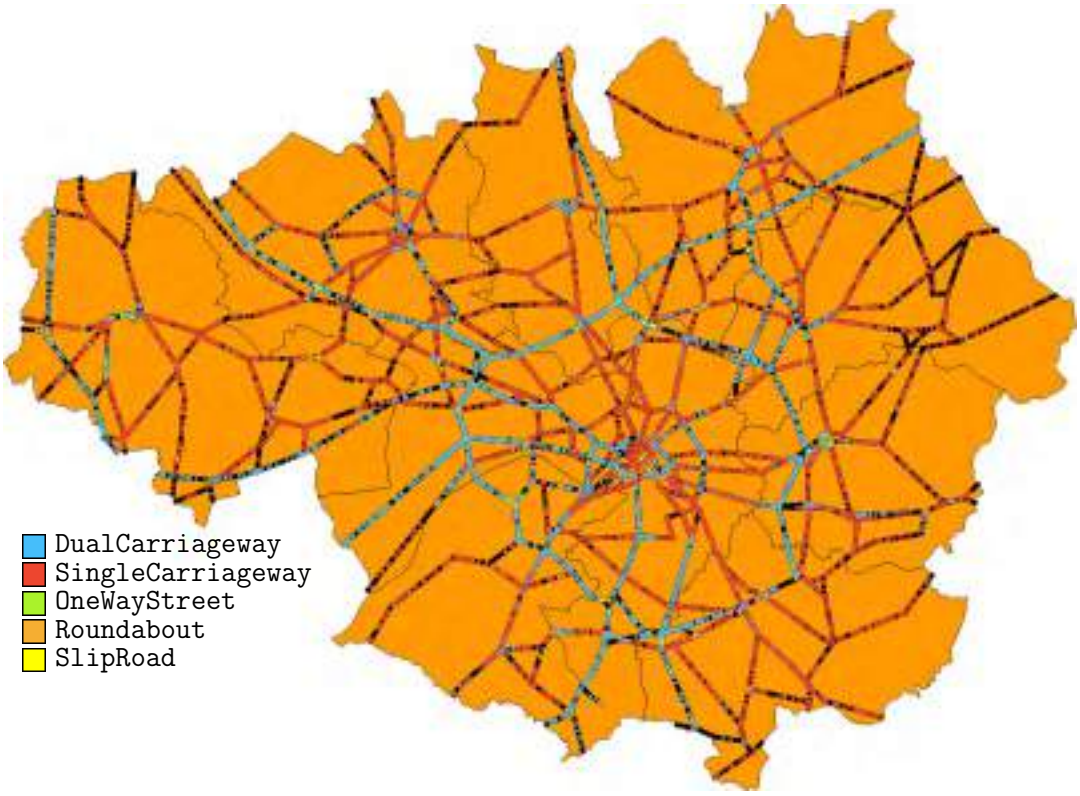


FIGURE 8.4: *A real-life scene at the Bridge Street and Deansgate intersection in Manchester facing due east [90].*

assumed by the yellow vehicle (labelled "F"), in order to gain a better view of the vehicles travelling on the A56. This may, in turn, result in a situation where a distracted pedestrian (labelled "H"), who is attempting to cross the road at the pedestrian crossing (labelled "G") during the A34's red phase, is hit by an approaching vehicle. Although this is only one probable hypothesis, this same kind of reasoning may be performed by a domain knowledge expert to develop multiple hypotheses in an attempt to explain why accidents may occur at a specific road location.

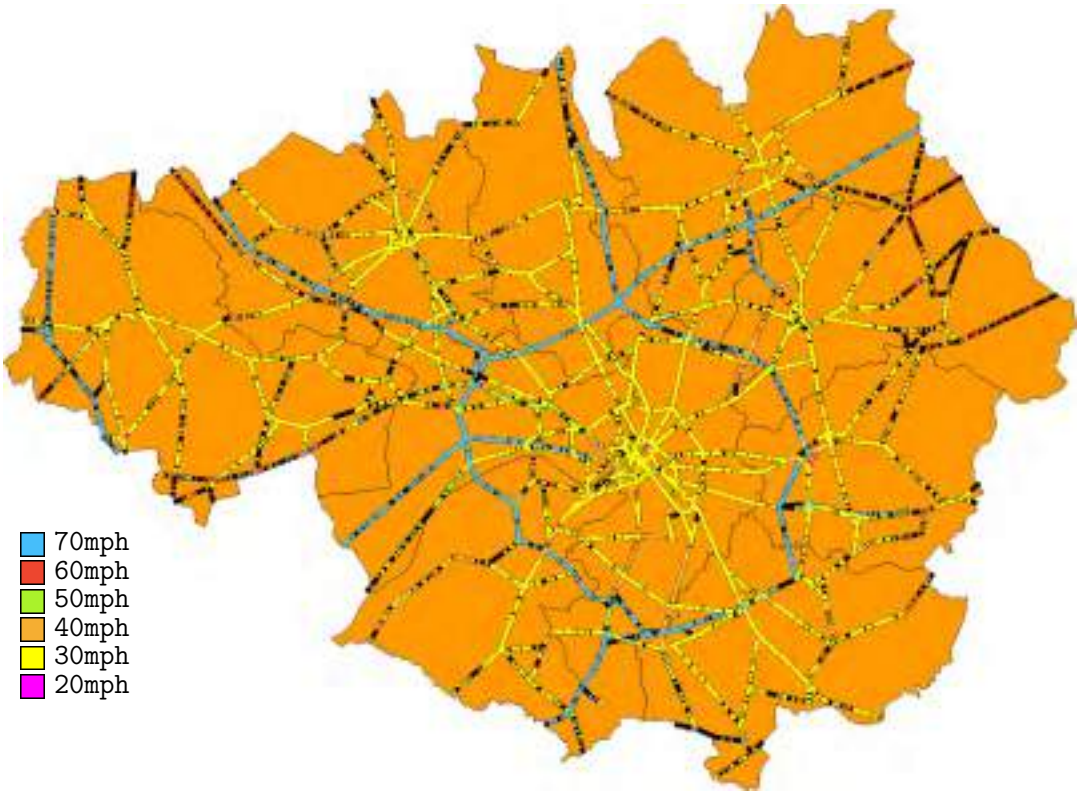
Sub-module 7.2 may also be used to spatially view the semi-spatial attribute values corresponding to each of the road segment observations. The spatial results corresponding to the most interesting semi-spatial attributes are shown in Figure 8.5, while the spatial results of the remaining semi-spatial attributes may be found in Appendix F. Figure 8.5(a) serves as a form of verification of Sub-module 7.1 as it provides insights into the module's ability to infer the semi-spatial attribute values from RAs to a set of road segment observations — one can more clearly see the M60 motorway circling the city of Manchester. It is also clear from the spatial results shown in Figure 8.5(b) that motorways (like the M60) are, as expected, dual carriageways, while A type roads are typically single carriageways. One can also confirm from this figure that slip roads are located at motorway intersections, while the majority of one way streets are found in Manchester city centre. One may deduce from Figure 8.5(c) that motorways have a set speed limit of 70mph — a fact confirmed by the UK Department for Transport [279]. Figures 8.5(c) and (d) also suggest that the majority of major roads in Greater Manchester are in urban settings with a set speed limit of 30mph. It may also be deduced that 60mph roads are typically single carriageways located in rural areas, while 20mph roads are found in highly urbanised settings like Manchester city centre. Although junctions are considered in their own right in the next section, the spatial results of Figure 8.5(e) provide some preliminary insights into the distribution of junctions in Greater Manchester. As may be seen in the figure, a large proportion of the junctions in Greater Manchester are T- or staggered junctions — the majority of which are located in urban settings (when taking Figure 8.5(d) into consideration).

(a) FirstRoadClass



(b) RoadType

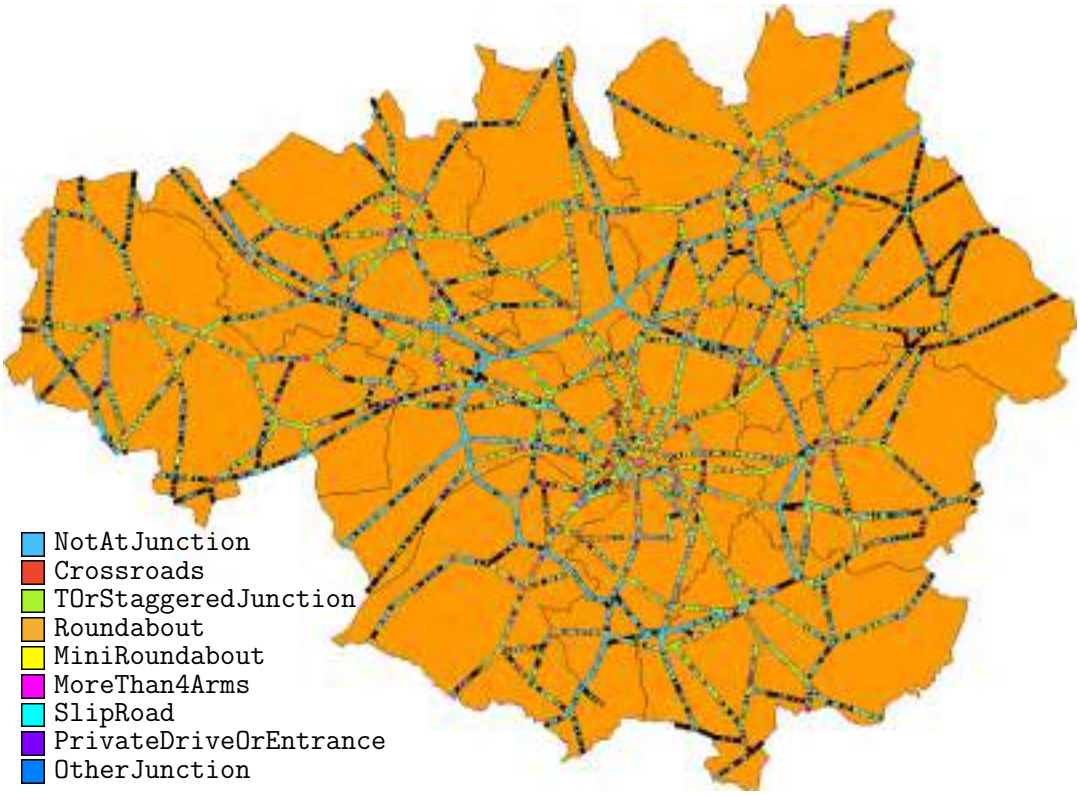
FIGURE 8.5: *A spatial representation of different road segment attribute values in Greater Manchester.*



(c) `SpeedLimit`

(d) `UrbanOrRuralArea`

FIGURE 8.5: *A spatial representation of different road segment attribute values in Greater Manchester.*



(e) `JunctionDetail`

FIGURE 8.5: *A spatial representation of different road segment attribute values in Greater Manchester.*

8.1.3 Non-spatial analysis of road segments

Sub-module 7.3 may also be employed by an analyst to view previously described accident rate and severity results in a non-spatial manner. This is achieved through the use of a scatter plot, like that shown in Figures 8.6, where the average accident rate is plotted against the average accident severity for the various categories corresponding to a specified attribute. In this figure, supplementary information is also provided by plotting each point as a pie chart so as to illustrate what proportion of an attribute's categories coincide with the categories of a supplementary attribute, while the size of a pie chart denotes the accident frequency of the specified scattered attribute category.

As may be seen in Figure 8.6, the average rate of RAs appear to be lower on single and dual carriageways (when compared to one-way streets and interchange roundabouts), but experience a greater accident severity, on average. A large proportion of accidents occurring on single carriageways appear to occur at some form of junction, the most prominent being a T- or staggered junction, which corroborates the fact that the majority of junctions in Greater Manchester are, in fact, T- or staggered junction, as previously seen in Figure 8.5(e). The highest average accident rate occurs on one-way streets — a road type common to dense urban environments which supports the conclusions drawn from the spatial representations shown in Figures 8.2(a) and 8.5(b).

Figure 8.7 provides more insight into the accident rate and severity associated with each speed limit. The highest average rate of accidents occurs on 20mph roads — a speed limit common to dense urban environments, which supports the conclusions drawn from the spatial representations shown in Figures 8.2(a) and 8.5(c). The highest accident severity is observed on 60mph

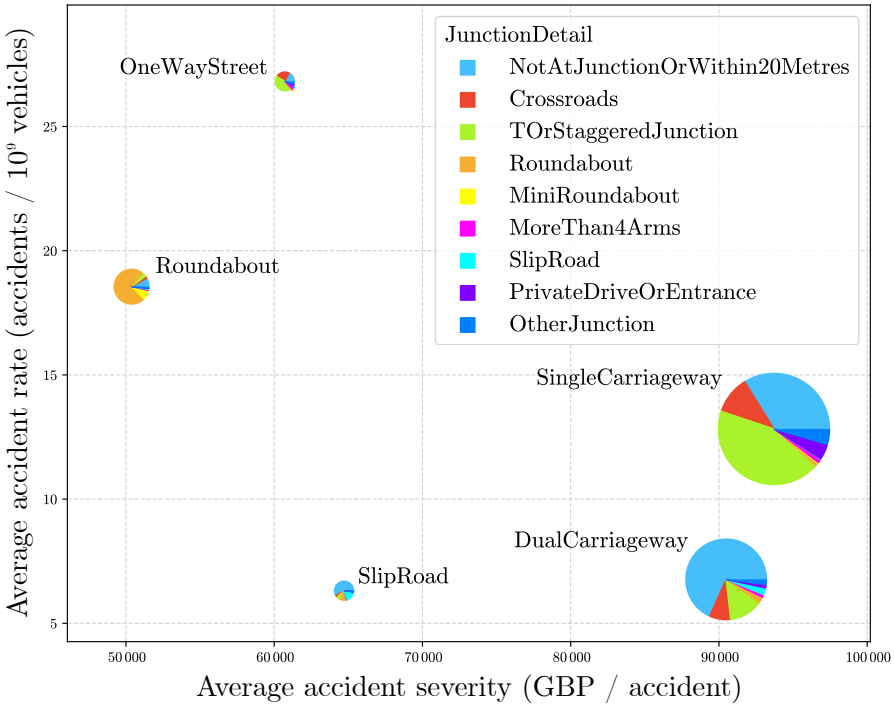


FIGURE 8.6: *An accident rate against accident severity scatter plot of the* `RoadType` *attribute categories, where each point is plotted as a pie chart showing the road segment proportions coinciding with the* `JunctionDetail` *attribute categories, and the size of a pie chart denotes accident frequency for the* `RoadType` *attribute category.*

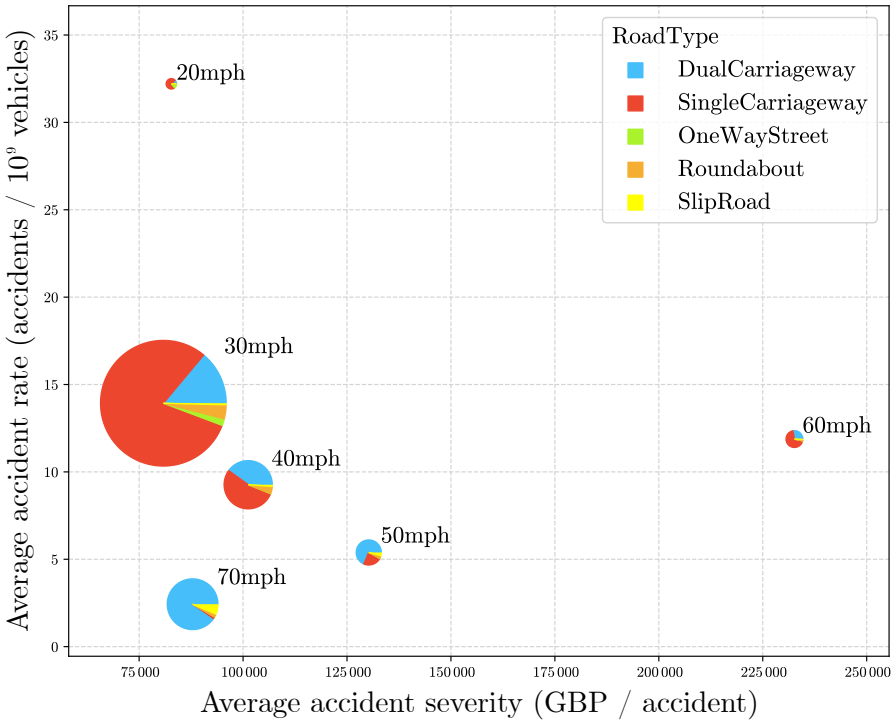


FIGURE 8.7: *An accident rate against accident severity scatter plot of the* `SpeedLimit` *attribute categories, where each point is plotted as a pie chart showing the road segment proportions coinciding with the* `RoadType` *attribute categories, and the size of a pie chart denotes accident frequency for the* `SpeedLimit` *attribute category.*

roads which may appear surprising when compared with the accident severity on 70mph roads. One would expect road segments with higher speed limits to result in higher accident severities. As previously deduced from Figures 8.5(b) and (c), however, motorways typically have a speed limit of 70mph in the UK and are commonly dual carriageways. As previously mentioned, 60mph roads, on the other hand, are typically single carriageways common to rural areas. A possible explanation for this observation may be that motorists are more likely to experience a severe accident on a rural 60mph single carriageway (than on a 70mph dual carriageway) because of the increased chance of head-on collisions on these roads (*i.e.* there is no physical barrier separating traffic flowing in opposite directions on a single carriageways).

Further insights into the accident rate and severity along road segments located in urban and rural areas may be seen in Figure 8.8. From this figure, one may conclude that the rate of being involved in an accident on an urban road is significantly higher than that on a rural road. Conversely, the severity of an accident on rural roads is, on average, more severe than on roads in an urban setting. Presented with this information, a domain knowledge expert may form the following hypothesis: The higher rate of accidents in urban areas is due to the higher density of roads and thus an increased chance of having an accident at a road junction. The majority of accidents on rural roads, on the other hand, do not occur at any form of junction, as may be seen in Figure 8.8, which may also suggest that accidents on these roads typically occur at higher speeds.

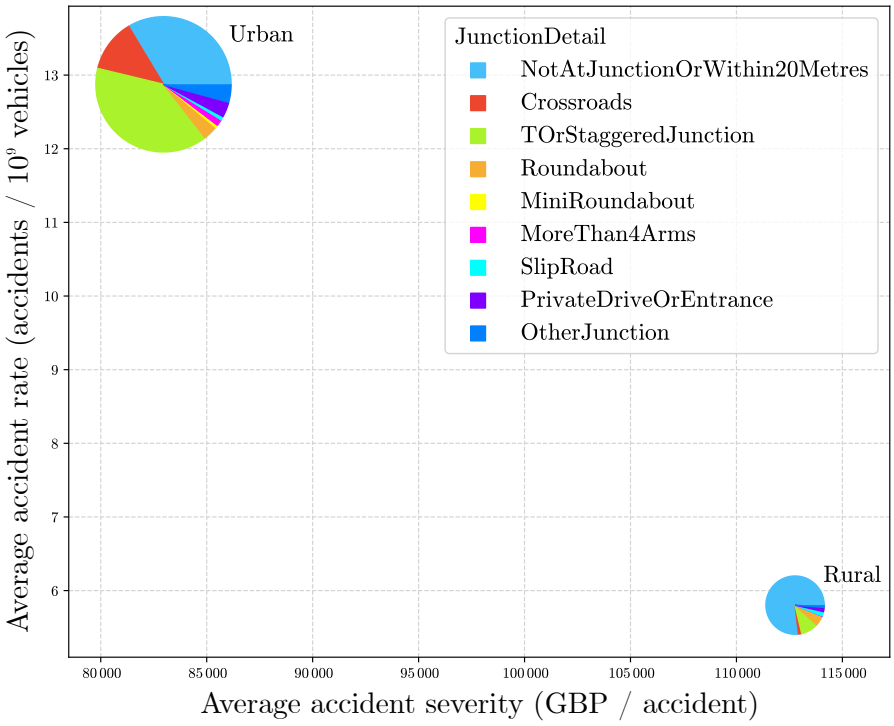


FIGURE 8.8: *An accident rate against accident severity scatter plot of the* `UrbanOrRuralArea` *attribute categories, where each point is plotted as a pie chart showing the road segment proportions coinciding with the* `SpeedLimit` *attribute categories, and the size of a pie chart denotes accident frequency for the* `UrbanOrRuralArea` *attribute category.*

8.1.4 Parameter choices for predictive analysis of road segments

Attention now shifts to demonstrating how Module 9.0 may be employed to provide a predictive analysis view of the results uncovered in the previous section. In accordance with the procedure described in §6.9, the first step involved selecting an appropriate set of road segment observation attributes. A total of thirty six road segment attributes were constructed according to the process described in §6.7 (one for each accident attribute category). Since the continuous road observation attributes are essentially a one-hot encoded form of the original accident categorical attributes, this leads to attributes that are highly correlated. To combat this problem, one accident attribute category is selected as a reference (as may be seen in Appendix B) and the corresponding road segment attribute is removed, resulting in a total of twenty nine road observation attributes. A random 70% of the road observations formed the learning set and the remaining observations the testing set.

A total of six regression learning models, described in Chapter 3, were selected for this regression task, namely the *Linear Regression* (LR), *Regression Tree* (RT), *Random Forest* (RF), kNN, SVM and MLP learning models. A constant model which always predicts the target variable mean was also included to serve as a benchmark. Attention then turned to (approximately) optimally tuning the hyper-parameters of each of the learning models by employing a grid-based search technique. The parameters considered in this investigation, together with the (approximately) optimal parameters found in the case of each learning model considered, are as follows:

LR. In the case of the LR learning model, the *L1 ratio parameter* (which controls the proportion of lasso regularisation to ridge regularisation) was considered, together with the regularisation penalty. For accident rate, the best LR learning model achieved an R^2 performance of 0.189, and had an L1-ratio of 0.04 and a regularisation penalty of 0.0005. In the case of accident severity, the best LR learning model achieved an R^2 performance of 0.022, and had an L1-ratio of 0.17 and a regularisation strength of 0.0002.

RT. In the case of the RT learning model, the MSE, MAE and Friedman MSE splitting criteria were considered and the depth of the tree was controlled by imposing a minimum sample number in each leaf. For accident rate, the best RT learning model achieved an R^2 performance of 0.281 and employed the MAE splitting criterion with a minimum of 37 samples per leaf. In the case of accident severity, the best RT learning model achieved an R^2 performance of 0.025 and employed the MSE splitting criterion with a minimum of 96 samples per leaf.

RF. In the case of the RF learning model, the number of regression trees and the maximum attribute subset considered when growing an individual tree were chosen as the appropriate set of parameters. For accident rate, the best RF learning model achieved an R^2 performance of 0.425 and employed 150 regression trees, each considering only one random attribute when growing the tree. For accident severity, the best RF learning model achieved an R^2 performance of 0.028 and employed 190 regression trees, each considering two random attributes when growing the tree.

kNN. In the case of the kNN learning model, the Euclidean, Manhattan and Chebyshev measures were considered as possible distance metrics weighted either uniformly or inversely by distance, together with the number of neighbours parameter. For accident rate, the best kNN learning model achieved an R^2 performance of 0.379 and considered the seven nearest neighbours based on an inverse weighted Manhattan distance. For accident severity, the best kNN learning model achieved an R^2 performance of 0.037 and considered the 97 nearest neighbours based on an inverse weighted Chebyshev distance.

SVM. In the case of the SVM learning model, a linear, polynomial, sigmoid and RBF kernel function where considered, together with the cost penalty term value. For accident rate, the best SVM learning model achieved an R^2 performance of 0.187 and employed an RBF kernel with a cost penalty term of 10.0. For accident severity, the best SVM learning model achieved an R^2 performance of -0.061 and employed a linear kernel with a cost penalty term of 9.0.

MLP. In the case of the MLP learning model, only network structures with at most two hidden layers and a maximum of fifty eight neurons (*i.e.* twice the number of input attributes) per hidden layer were considered, based on the literature reviewed in §4.9. The ReLu activation function was employed in the case of each MLP structure and optimised the weights of the network according to the Adam optimisation approach, as these are the most popular choices in the literature [180, 227]. For accident rate, the best structure achieved an R^2 performance of 0.308 and had two hidden layers, containing 28 neurons in each layer. In the case of accident severity, the best structure achieved an R^2 performance of 0.021 and had two hidden layers, again containing 28 neurons in both the first and second layers. It is interesting that, for both accident rate and severity, the best MLP structure contained 28 hidden neurons in both hidden layers.

8.1.5 Predictive analysis of road segments

After (approximately) optimal parameter settings had been determined, each learning model was evaluated according to the well-known MSE and R^2 performance metrics in respect of the learning and testing set, as per to the procedure described in §6.9. These results may be seen in Table 8.1 for both cases in which accident rate and accident severity were selected as the target variable.

TABLE 8.1: *The MSE and R^2 performances achieved by the six learning models and the constant model in respect of the learning and testing sets when the accident rate and severity along a road segment are selected as the target variables.*

		Learning		Testing	
Target	Model	MSE	R^2	MSE	R^2
Rate	LR	479.95	0.120	259.14	0.212
	RT	435.20	0.202	232.26	0.294
	RF	76.15	0.860	166.25	0.494
	kNN	14.56	0.973	238.09	0.275
	SVM	484.12	0.112	263.86	0.197
	MLP	396.95	0.272	230.07	0.300
	Const.	545.32	0.000	328.80	0.000
Severity	LR	520.79	0.027	537.11	0.022
	RT	509.40	0.047	536.13	0.024
	RF	359.40	0.328	532.75	0.030
	kNN	336.54	0.370	531.24	0.033
	SVM	567.12	-0.010	582.89	-0.062
	MLP	521.89	0.024	537.89	0.020
	Const.	534.56	0.000	549.17	0.000

As may be seen in this table, in the case when accident rate along road segments was selected as the target variable, the RF learning model achieved a significantly greater predictive performance

than the other learning models when evaluated on the unbiased test set. Although performing the best on the learning set, the kNN learning model significantly overfitted the data and failed to generalise on the test set. The LR, RT, SVM and MLP learning models were all able to generalise much better to the unseen test data set, although not comparable to the performance achieved by the RF learning model.

In the case when accident severity along road segments was selected as the target variable, it is clear that all learning models exhibited a lack of ability to predict the accident severity measure. This suggests that accident severity may be more strongly linked to other non-spatial attributes inherent to the accident itself (*e.g.* the vehicle's type and manoeuvre, or weather and lighting condition) as opposed to the attributes inherent to the road segment — a finding which warrents further investigation. Although the kNN learning model was the best performing learning model, it only managed to achieve an R^2 performance of 0.033.

As described in §6.9, the newly proposed SHAP values may be employed to provide insight into the working of each learning model. As a way of demonstration, only the SHAP values corresponding the two MLP models trained to predict accident rate and severity are shown in Figures 8.9 and 8.10, respectively. The MLP learning model was chosen as it is inherently black-box in nature and has, to the best of the author's knowledge, never been used in road accident analysis literature to facilitate an interpretation of the impact of road characteristics on the accident rate and severity along road segments. The SHAP values corresponding to all other models may be found in Appendix G. In both Figures 8.9 and 8.10, the SHAP value for each road segment observation (denoted as a point) is plotted according to each of the semi-spatial attributes, where each road segment observation is coloured according to its attribute value. Attributes are ranked according to the maximum absolute deviation in SHAP value $\sum_{i=1}^{M} |\phi_i^{(j)}|$ (indicated by the horizontal black bar) which may be viewed as a measure of importance for attribute j. In essence, a red point represents a road segment observation in which the indicated accident attribute category was present while a blue point represents a road segment observation where the indicated accident attribute category was absent. These figures should, therefore, be interpreted as follows: If a road segment observation lies to the right of the zero-line, the attribute value of that observation had a positive relative impact on the value of the defined target variable while, if it lies to the left of the zero-line, it had a negative relative impact on the value of the defined target variable.

As may be seen in Figure 8.9, the MLP trained to predict accident rate along road segments identified `JunctionControl=GiveWayOrUncontrolled` as the most important attribute. More specifically, a relatively higher accident rate is observed when road observations are at uncontrolled junctions. Furthermore, the `RoadType=DualCarriageway` attribute was identified as the second-most informative attribute. More specifically, a lower relative accident rate is observed when road segment observations are dual carriageways. The same conclusion can be made in the case where the road type is a single carriageway or roundabout. These observations are confirmed by the descriptive results observed in Figure 8.6, where single and dual carriageways, as well as roundabouts, exhibit a lower average accident rate than to one-way streets.

Although not a great predictor of accident severity along road segments, the SHAP values may still offer some insight into the rule formulation produced by the MLP model. In Figure 8.10, the MLP identified `RoadType=SingleCarriageway` as the most informative attribute. More specifically, a higher relative accident severity is observed when road observations are single carriageways. The same conclusion can be made in the case where the road type is a dual carriageway. These observations are again confirmed by the descriptive results observed in Figure 8.6 where single and dual carriageways exhibit higher average accident severities than the other road types. The second-most informative feature was `FirstRoadClass=A` where it is observed that many

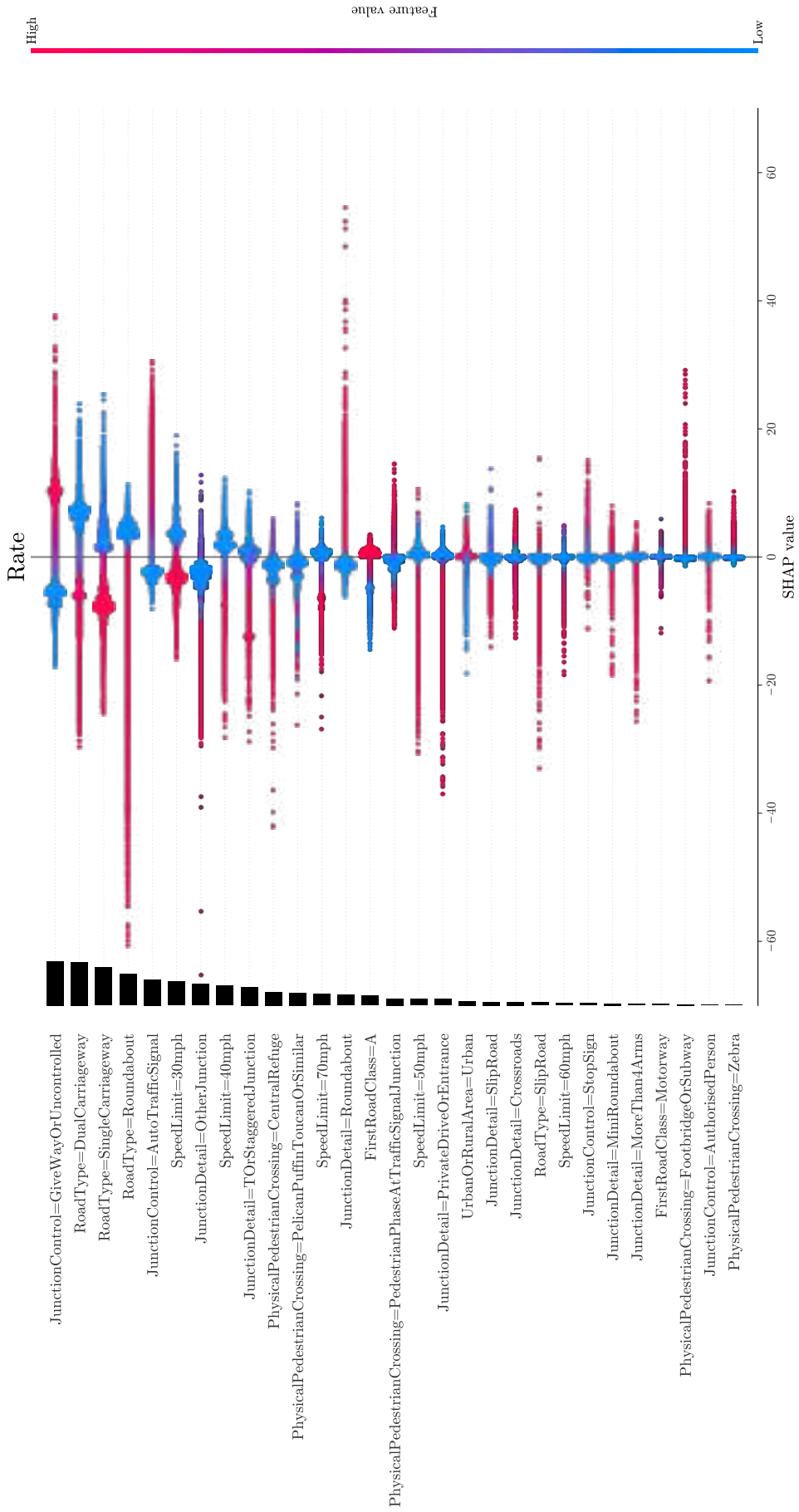


FIGURE 8.9: *SHAP value results for the MLP trained to predict accident rate, where each road segment observation is plotted according to the semi-spatial attribute categories.*

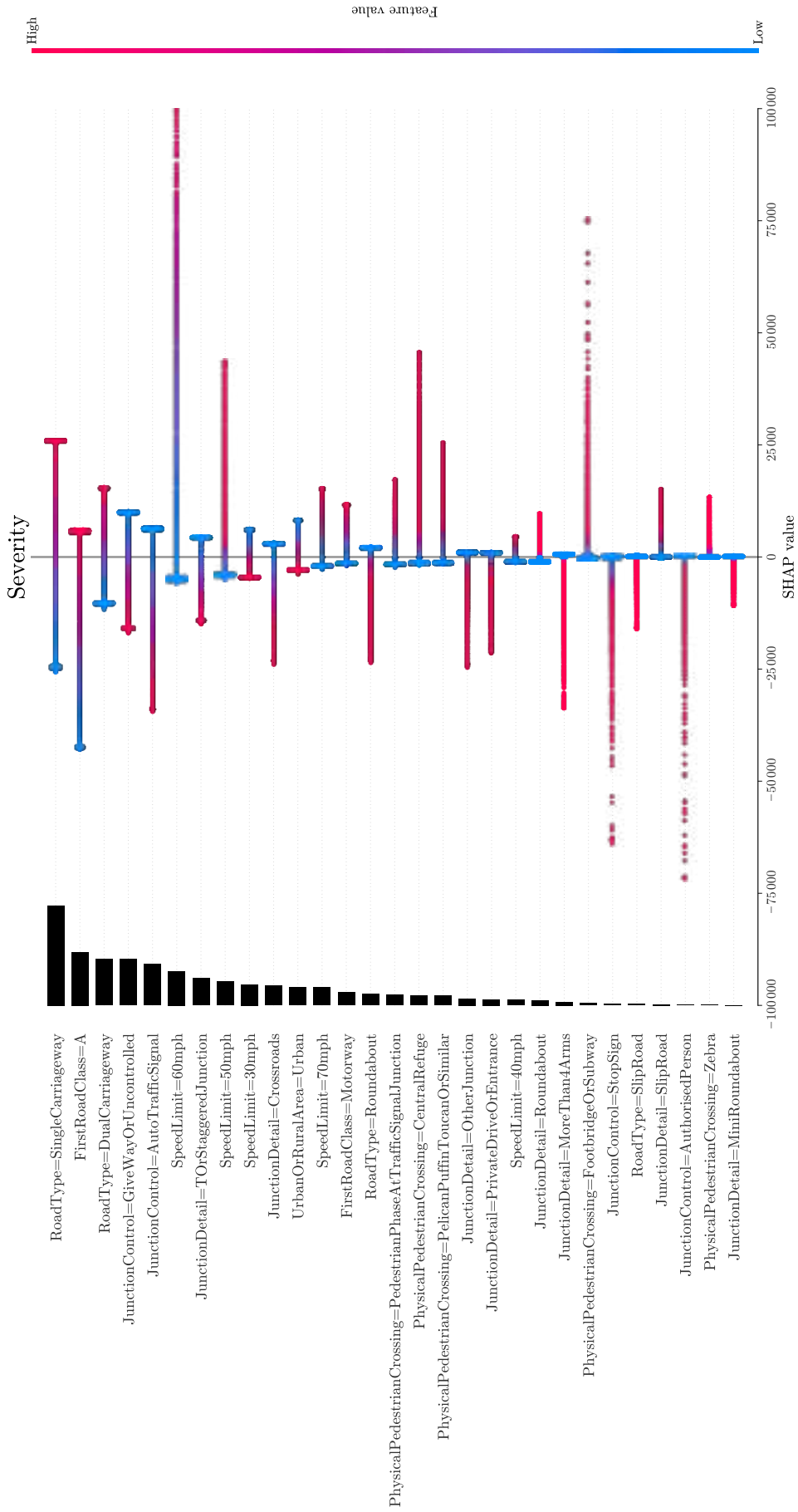


FIGURE 8.10: *SHAP value results for the MLP trained to predict accident severity, where each road segment observation is plotted according to the semi-spatial attribute categories.*

road observations which are not A-type roads are predicted to exhibit lower accident severities. The `JunctionControl=GiveWayOrUncontrolled` and `JunctionControl=AutoTrafficSignal` attributes are also present and suggest that lower accident severities are experienced at junctions with these control types. The sixth-most important attribute was `SpeedLimit=60mph` which suggests that significantly higher accident severities are experienced on 60mph roads — a finding corroborated by Figure 8.7.

This illustrates that both the descriptive and predictive modelling approaches proposed in this section should be carried out in unison to draw conclusions with respect to the relative rate and severity of accidents based on various road characteristics. It is proposed that the SHAP method may serve as an appropriate tool for the interpretation of black-box learning models by road safety researchers when used in conjunction with a descriptive view of the data being modelled. Road safety researchers may, therefore, develop more confidence in the predictions made by a learning model if their SHAP values align with the general intuition they have surrounding the phenomenon being modelled.

8.2 Analysis of road accident risk at road junctions

This section takes the same form as the previous section but is primarily concerned with demonstrating how Modules 8.0 and 9.0 may be used to answer the following overarching investigative question: *What road characteristics influence the rate and severity of RAs at road junctions?* As previously described in §6.8, Module 8.0 aims to answer this question in the same manner as Module 7.0 — by providing descriptive insights as a result of viewing the results in both a spatial and non-spatial manner. As in the previous section, the parameter choices made during the construction of road junction observations in Sub-module 8.0 are first stated and substantiated in §8.2.1. The focus then shifts in §8.2.2 to quantifying the accident rate and severity at road junctions in Greater Manchester and interpreting these results using the spatial representation produced by Sub-module 8.4. These findings are supported in §8.2.3, by interpreting the same results from a non-spatial perspective by employing Sub-module 8.5. Attention then shifts in §8.2.4 to demonstrating how Module 9.0 may be employed to obtain a prescriptive analytic view of the results, which begins by first describing the parameter choices associated with Module 9.0. This is followed, in §8.2.5, by an assessment of the predictive performance the same six learning models considered in the previous road segment investigation, as well as an interpretation of the SHAP values of the MLP learning model.

8.2.1 Parameter choices for descriptive analysis of road junctions

As described in §6.6, Sub-modules 8.1 and 8.2 require the user to specify a radius ρ, a junction dissimilarity threshold β and a similarity measure Λ. A value of $\rho = 100$ metres was selected as an appropriate radius of RA effect since it is reasonable to assume that RAs likely did not occur at the same junction if they are not within this distance from one another. The junction dissimilarity threshold β and similarity function Λ were determined through an empirical study — the results of which may be found in Appendix H. In this study, the Lin, Goodall and Occurrence Frequency similarity measures where considered as they achieve consistently better clustering performance across a wide variety of benchmark data sets [41]. Each similarity measure was assessed according to different junction dissimilarity threshold values. Overall, it was decided that the Goodall similarity measure with a junction dissimilarity threshold of $\beta = 0.7$ produced adequate junction clusterings.

8.2.2 Spatial descriptive analysis of road junctions

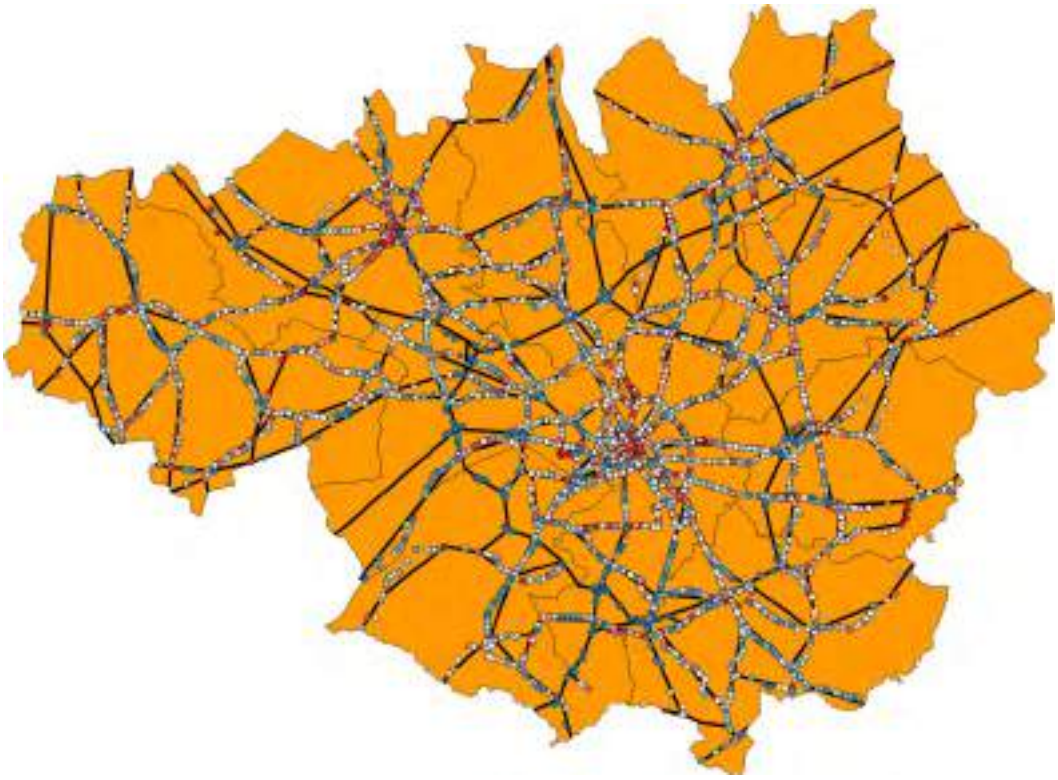
As was the case in the road segment investigation, the first step was to quantify the accident rate and severity measures at the defined road junction observations. A geographical representation of these results for the Greater Manchester road network are illustrated in Figures 8.11(a) and (b), respectively, where red dots represent junctions with higher accident rates/severities and blue dots represent junctions with lower accident rates/severities. Sub-module 8.4 may, therefore, be used to provide insight into the question: *Which road junctions exhibit significantly high accident rates and severities in Greater Manchester?* From these results, one may deduce that significantly higher accident rates appear to occur at junctions in highly urbanised areas like Manchester city centre. By drawing from the spatial results uncovered in Figure 8.5(d), one may reason that lower accident rates appear to occur at junctions located in rural areas. Again, it is not entirely clear, in the case of accident severity, whether one can draw any preliminary insight from the spatial representation in Figures 8.11(b).

Sub-module 8.4 may also be used to spatially view the semi-spatial attribute values corresponding to each of the road junction observations. The spatial results corresponding to the most interesting semi-spatial attributes are shown in Figure 8.12, while the spatial results of the remaining semi-spatial attributes may be found in Appendix I. Figure 8.12(a) confirms the preliminary insights drawn from Figure 8.5(e). As may be seen in the figure, a large proportion of the junctions in Greater Manchester are T- or staggered junctions and crossroads, which is expected, given that 80.03% of the road segments in Greater Manchester are located in urban areas, where the majority of these junctions are uncontrolled, as may be seen in Figure 8.12(b). There does, however, appear to be a higher concentration of junctions controlled by traffic lights in highly urbanised areas like Manchester city centre. Figure 8.12(c) also provides more insight into the physical pedestrian crossing facilities located at each junction, the most common being a pedestrian light phase which indicates whether a pedestrian may cross the road or not.

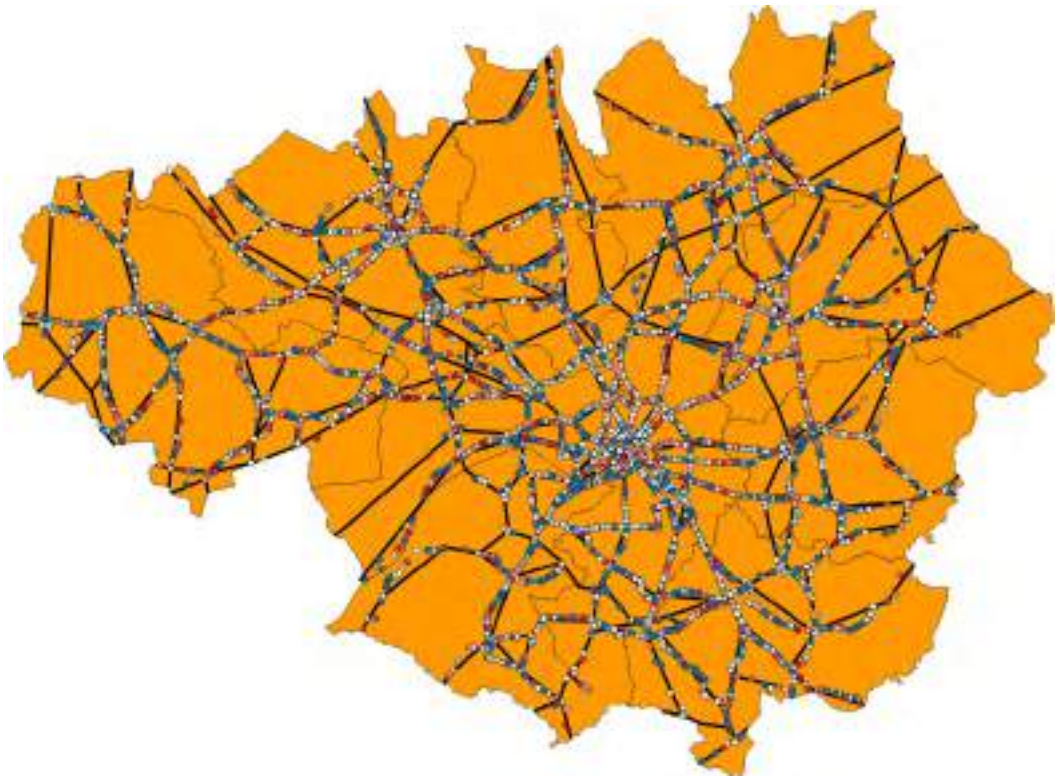
8.2.3 Non-spatial descriptive analysis of road junctions

In the same way as Sub-module 7.3, Sub-module 8.5 may be employed by an analyst to generate a scatter plot representation of the accident rate and severity road junction results. Figure 8.13 provides more insight into the accident rate and severity associated with each junction type. As may be seen in the figure, the highest average accident rate is experienced at crossroads. T- or staggered junctions, in comparison, exhibit lower accident rates but higher accident severities, on average. One possible explanation for this observation is that four-arm junctions have more conflict points between streams of traffic than three-arm junctions, which would likely lead to higher accident frequencies [300]. Uncontrolled T- or staggered junctions, on the other hand, typically require motorists travelling on minor roads to merge with traffic on a major road. This often requires motorists turning left to make judgements based on traffic flows from both directions, which may result in T-bone collisions impacting on the driver's side of the vehicle. It may also be deduced that accidents at a roundabout are the least severe of all other junction types. This result, therefore, confirms that the primary purpose of roundabouts, which is to reduce traffic speeds and minimise T-bone and head-on collisions [284], is indeed being realised in Greater Manchester.

Further insight into the accident rate and severity at junctions with different control facilities may be seen in Figure 8.14. From the figure, one may conclude that the rate of being involved in an RA at a junction with traffic lights is significantly higher than for the other junction control measures. As previously deduced from Figure 8.12(b), traffic lights are more common to highly

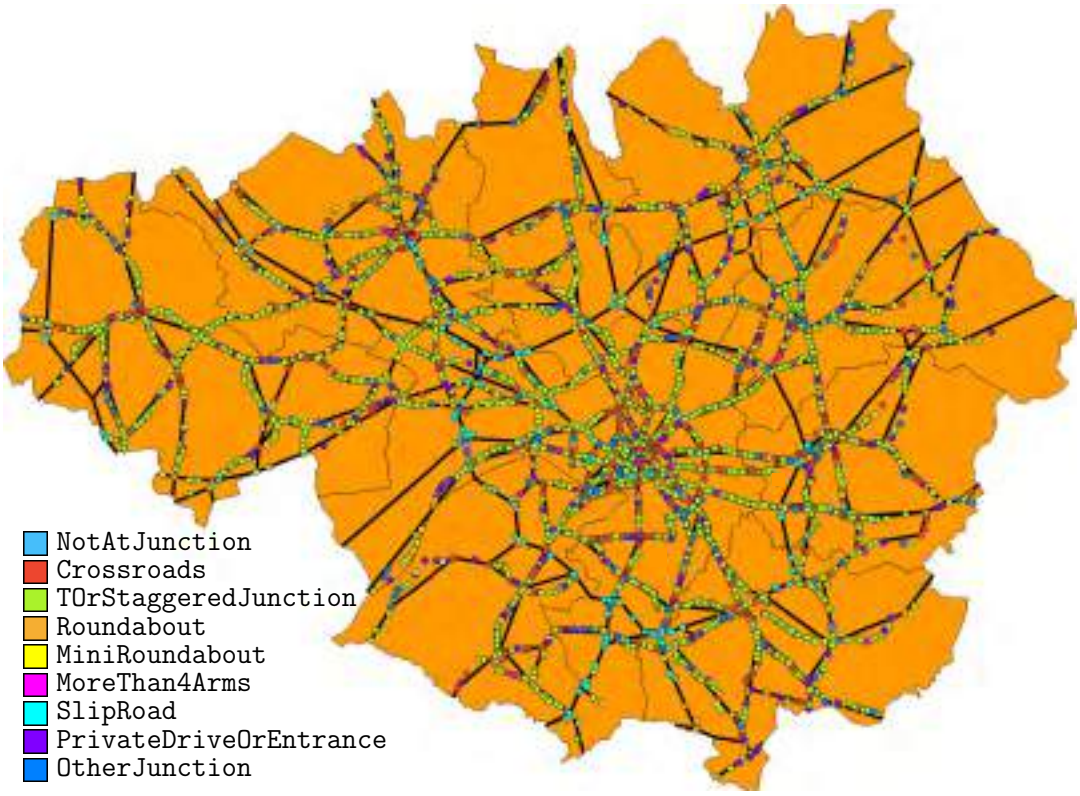


(a)

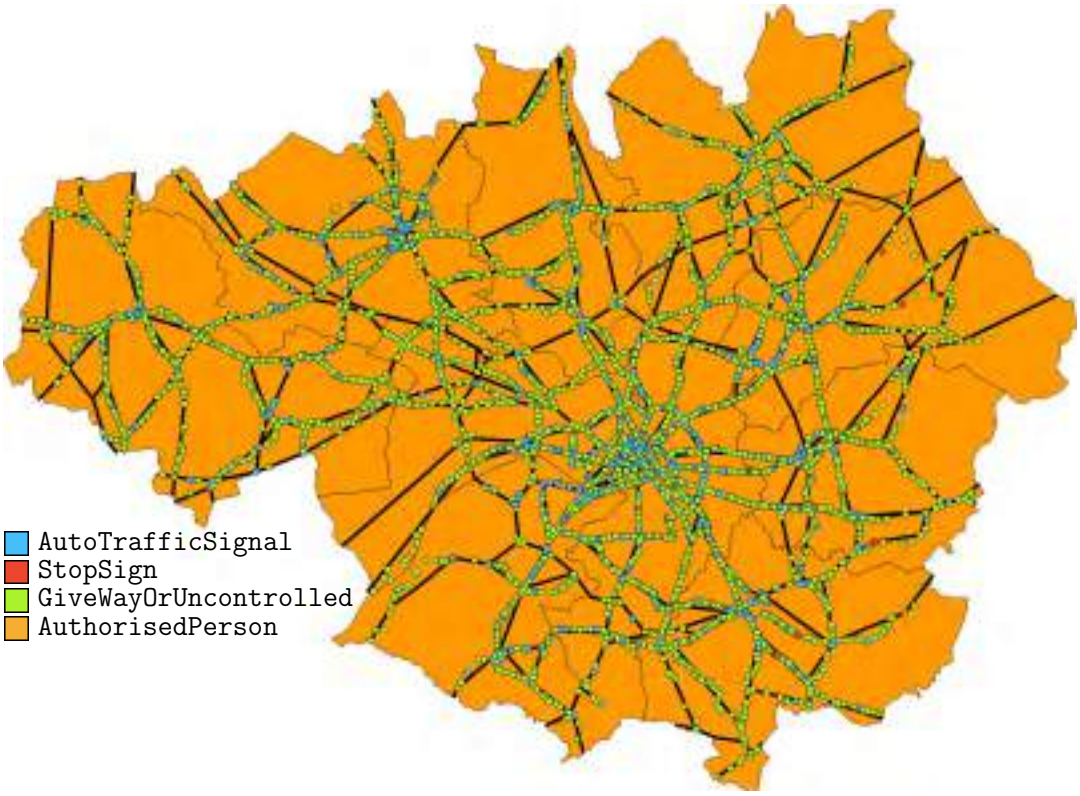


(b)

FIGURE 8.11: *A spatial representation of the Greater Manchester (a) accident rate and (b) accident severity measurements at road junctions.*

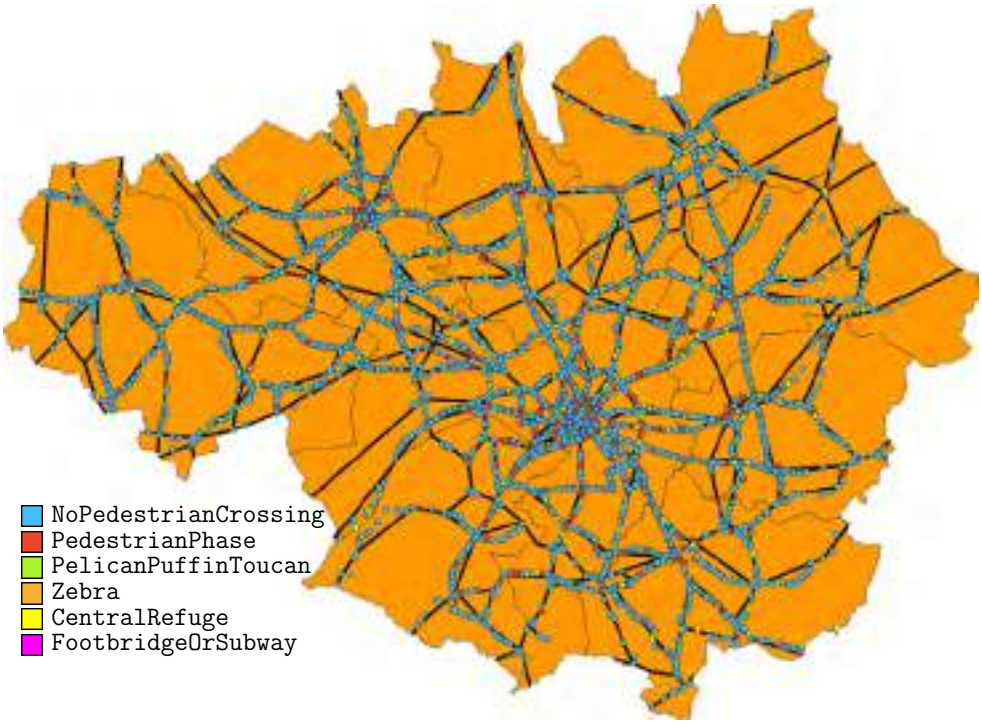


(a) `JunctionDetail`



(b) `JunctionControl`

FIGURE 8.12: *A spatial representation of different road junction attribute values in Greater Manchester.*



(c) `PhysicalPedestrianCrossing`

FIGURE 8.12: *A spatial representation of different road junction attribute values in Greater Manchester.*

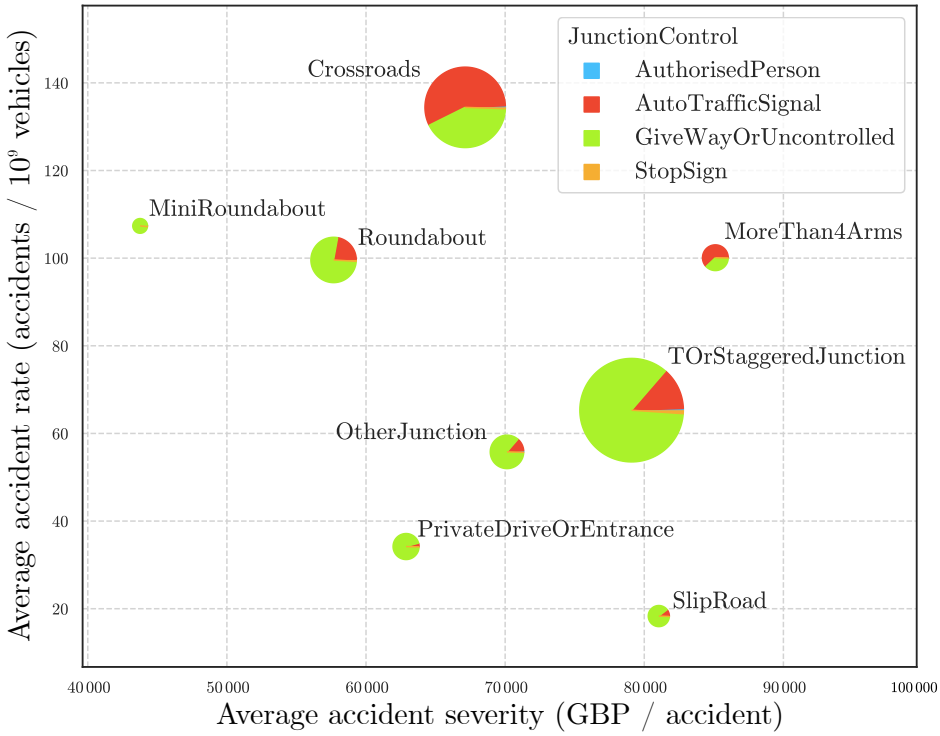


FIGURE 8.13: *An accident rate against accident severity scatter plot of the* `JunctionDetail` *attribute categories, where each point is plotted as a pie chart showing the road junction proportions coinciding with the* `JunctionControl` *attribute categories, and the size of a pie chart denotes accident frequency for the* `JunctionDetail` *attribute category.*

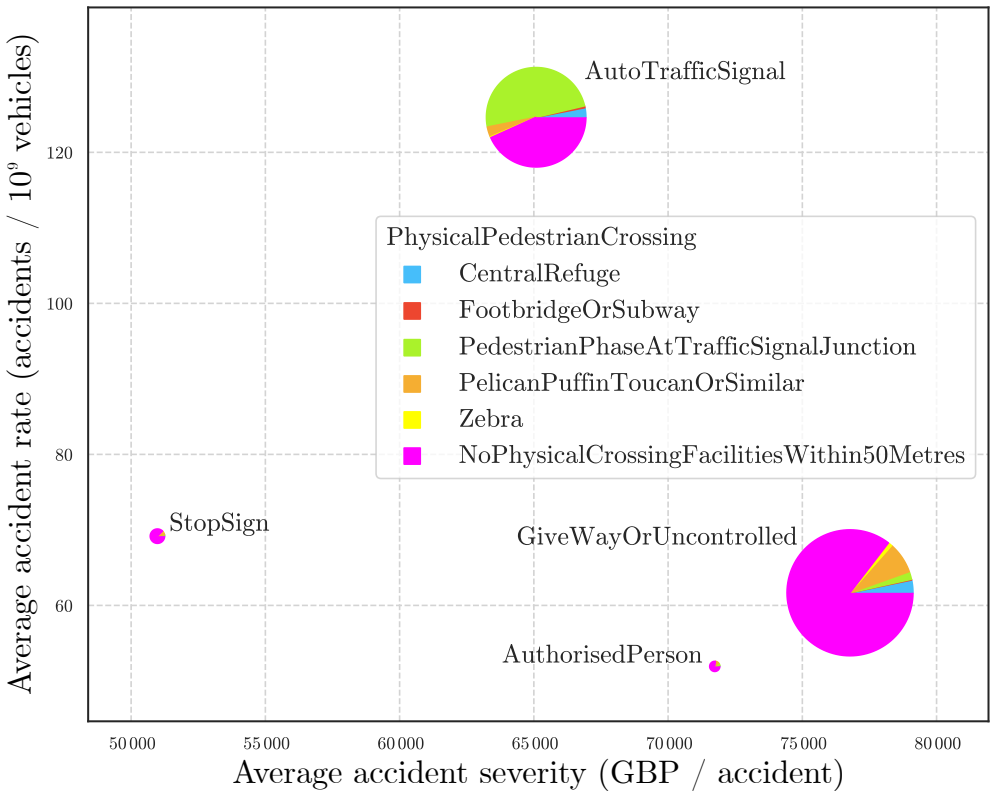


FIGURE 8.14: *An accident rate against accident severity scatter plot of the* `JunctionControl` *attribute categories, where each point is plotted as a pie chart showing the road junction proportions coinciding with the* `PhysicalPedestrianCrossing` *attribute categories, and the size of a pie chart denotes accident frequency for the* `JunctionControl` *attribute category.*

urbanised environments, which is confirmed by the high proportion of pedestrian phase crossing facilities at these junctions, as may be seen in Figure 8.14. In highly urbanised environments, there are more people and thus more vehicles per available road space, which may serve as a possible explanation for this observation. Uncontrolled junctions, on the other hand, experience the highest accident severity, on average, which is understandable since they account for slip roads exiting motorways as well as junctions in highly urbanised areas, as previously seen in Figure 8.12(b).

Figure 8.15 provides more insight into the accident rate and severity associated with each physical pedestrian crossing facility which may exist at a junction. As may be seen in the figure, junctions with pedestrian light phases have, on average, the highest accident rate. This is a product of what was previously observed in Figure 8.14, where junctions with traffic lights (the majority of which have pedestrian phases) experience the highest accident rate on average. This observation, therefore, highlights that multiple descriptive plots should be viewed by an analyst in order to arrive at a more informed interpretation of their results. Although junctions with a zebra crossing facility are relatively infrequent, they account for the highest average accident severity — a finding which suggests that more pedestrian-related accidents may occur at these junctions.

8.2.4 Parameter choices for predictive analysis of road junctions

As was the case in the road segment investigation, attention now shifts to demonstrating how Module 9.0 may be employed to provide a predictive analysis view of the results uncovered in

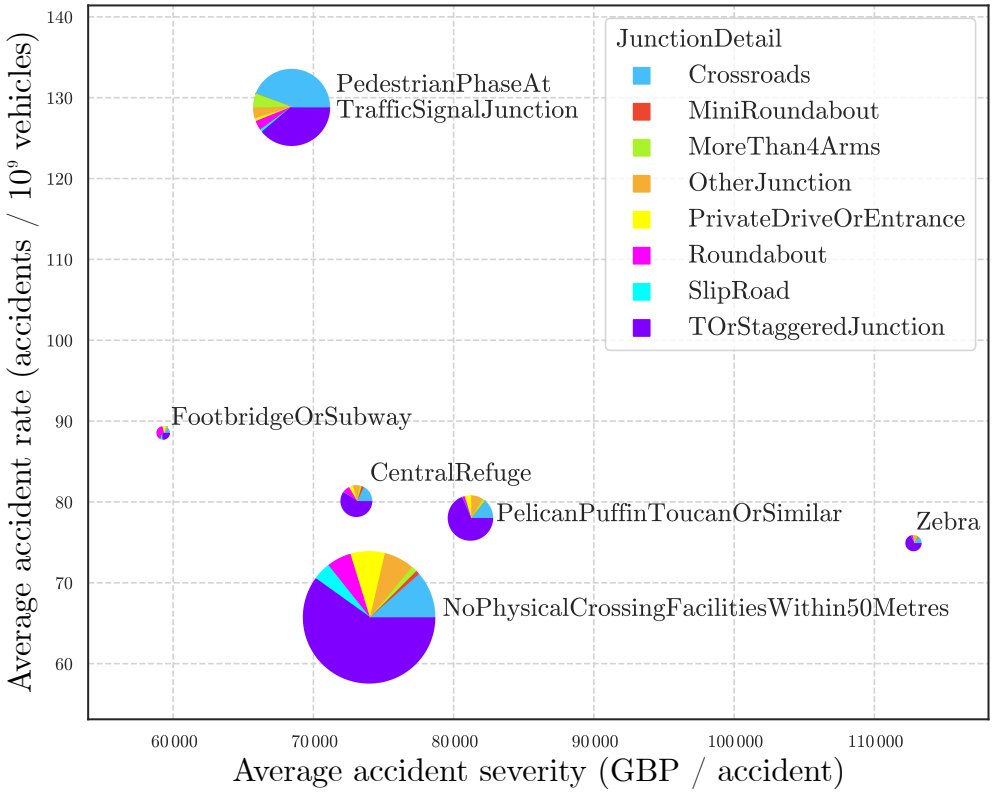


FIGURE 8.15: *An accident rate against accident severity scatter plot of the* `PhysicalPedestrianCrossing` *attribute categories, where each point is plotted as a pie chart showing the road junction proportions coinciding with the* `JunctionDetail` *attribute categories, and the size of a pie chart denotes accident frequency for the* `PhysicalPedestrianCrossing` *attribute category.*

the previous section. The procedure employed in this investigation is the same as that described in §8.1.4. The same six learning models, with the same hyper-parameter considerations, were employed in this investigation. A grid-based search technique was again employed in order to find (approximately) optimal parameters for each of the learning models, the results of which are as follows:

LR. For accident rate, the best LR learning model achieved an R^2 performance of 0.236, and had an L1-ratio of 0.98 and a regularisation penalty of 0.02. In the case of accident severity, the best LR learning model achieved an R^2 performance of 0.000, and had an L1-ratio of 0.49 and a regularisation strength of 0.2.

RT. For accident rate, the best RT learning model achieved an R^2 performance of 0.372 and employed the MSE splitting criterion with a minimum of nine samples per leaf. In the case of accident severity, the best RT learning model achieved an R^2 performance of -0.004 and employed the Friedman MSE splitting criterion with a minimum of 648 samples per leaf.

RF. For accident rate, the best RF learning model achieved an R^2 performance of 0.506 and employed 150 regression trees, each considering fifteen random attributes when growing the tree. For accident severity, the best RF learning model achieved an R^2 performance of -0.168 and employed 80 regression trees, each considering only one random attribute when growing the tree.

kNN. For accident rate, the best kNN learning model achieved an R^2 performance of 0.355

and considered the sixteen nearest neighbours based on an inversely weighted Manhattan distance. For accident severity, the best kNN learning model achieved an R^2 performance of -0.012 and considered the 85 nearest neighbours based on a uniformly weighted Chebyshev distance.

SVM. For accident rate, the best SVM learning model achieved an R^2 performance of 0.180 and employed a linear kernel with a cost penalty term of 10.0. For accident severity, the best SVM learning model achieved an R^2 performance of -0.068 and employed a linear kernel with a cost penalty term of 10.0.

MLP. For accident rate, the best structure achieved an R^2 performance of 0.308 and had two hidden layers, containing 30 and 33 neurons in the first and second hidden layers, respectively. In the case of accident severity, the best structure achieved an R^2 performance of -0.003 and had two hidden layers, containing 19 and 28 neurons in the first and second hidden layers, respectively.

8.2.5 Predictive analysis of road junctions

After (approximately) optimal parameter settings had been determined, each learning model was again evaluated according to the MSE and R^2 performance metrics in respect of the learning and testing set, as per the procedure described in §6.9. These results may be seen in Table 8.2 for both the cases in which accident rate and accident severity were selected as the target variables.

TABLE 8.2: *The MSE and R^2 performances achieved by the six learning models and the constant model in respect of the learning and testing sets when the accident rate and severity at a road junction are selected as the target variables.*

		Learning		Testing	
Target	Model	MSE	R^2	MSE	R^2
Rate	LR	4825.02	0.241	6846.46	0.209
	RT	2657.10	0.582	6476.91	0.252
	RF	1031.41	0.837	5399.67	0.376
	kNN	709.74	0.888	6444.10	0.256
	SVM	5224.87	0.179	7402.39	0.145
	MLP	4499.04	0.293	6727.50	0.223
	Const.	6369.82	0.000	8671.98	0.000
Severity	LR	2971.15	0.009	2868.17	0.001
	RT	2983.24	0.005	2872.37	0.000
	RF	2008.24	0.329	3212.55	−0.119
	kNN	2977.48	0.007	2872.21	0.000
	SVM	3170.32	−0.058	−0.051	−0.062
	MLP	2997.22	0.000	2874.54	−0.001
	Const.	3000.79	0.000	2879.39	0.000

These results echo many of the trends observed in Table 8.1. In the case when accident rate at road junctions was selected as the target variable, the RF learning model again achieved the best predictive performance when evaluated on the unbiased test set, while the kNN learning model again failed to generalise on the test set. The RT learning model also significantly overfitted the learning data — after achieving an R^2 performance of 0.582 on the learning data set, the model

only managed to obtain an R^2 performance of 0.252 on the test data set. In the case when accident severity at road junctions was selected as the target variable, it was again observed that all learning models exhibited a lack of ability to predict the accident severity measure — a finding which again begs the question whether non-spatial attributes are better predictors of accident severity than semi-spatial attributes.

The SHAP values corresponding the two MLP models trained to predict accident rate and severity at road junctions are employed to demonstrate how one may yield insight into the workings of the model, and are shown in Figures 8.16 and 8.17, respectively. The SHAP values corresponding to all other models may be viewed in Appendix I. As may be seen in Figure 8.16, the MLP trained to predict accident rate at road junctions identified `JunctionDetail=Crossroads` as the most important attribute. More specifically, a higher relative accident severity is observed when road observations are crossroads, which was previously observed in Figure 8.13. The same conclusion can be drawn in the case where the junction is controlled by a traffic light — a finding which is confirmed by the results previously shown in Figure 8.14.

8.3 Predictive analysis of road accident severity

In §8.1.5 and §8.2.5, it was demonstrated how a user may use the road segment and junction results of Modules 7.0 and 8.0, respectively, as a set of inputs to the supervised learning procedure contained in Module 9.0. In both these formulations, only the set of semi-spatial attributes was considered as explanatory variables since a majority consensus as to the value of the semi-attribute can be reached from all RAs occurring on the road entity (*i.e.* either along a road segment or at a junction). One could not, for example, consider the attribute `WeatherCondition` as an appropriate attribute of a road entity since the RAs which occurred on that road entity would likely not agree on the value of the weather, since the value of this attribute varies with time. If one wished to conduct a predictive investigation whereby the set of non-spatial attributes were selected as the set of explanatory variables, a road entity could no longer be defined as an observation. Rather, the RAs themselves must form the set of observations. One cannot, however, compare accidents which did not occur under the same (or similar) spatial conditions (*e.g.* accidents occurring on a motorway should not be compared with accidents that occurred at a crossroads since their road characteristics are fundamentally different). In order to conduct such an analysis, a subset of RAs occurring at the same (or a similar) spatial location on the road must, therefore, be defined.

In both §8.1.5 and §8.2.5, it was observed that semi-spatial attributes alone were not good predictors of accident severity. It was, therefore, hypothesised that accident severity may be more strongly linked to other non-spatial attributes inherent to the accident itself. This notion is addressed in this section by drawing from the combined data set, constructed in Chapter 7, and employing Module 9.0 to answer two investigative questions. The first investigation, conducted in §8.3.1, attempts to answer the question: *Which non-spatial factors result in a higher accident severity between two vehicles at crossroads controlled by traffic lights in Greater Manchester?* The second investigation, conducted in §8.3.2, serves to answer the question: *Which non-spatial factors result in a higher accident severity for single-vehicle accidents on motorways in Greater Manchester?* It should be noted that the set of observations are now no longer road segments or junctions, but rather the RAs themselves.

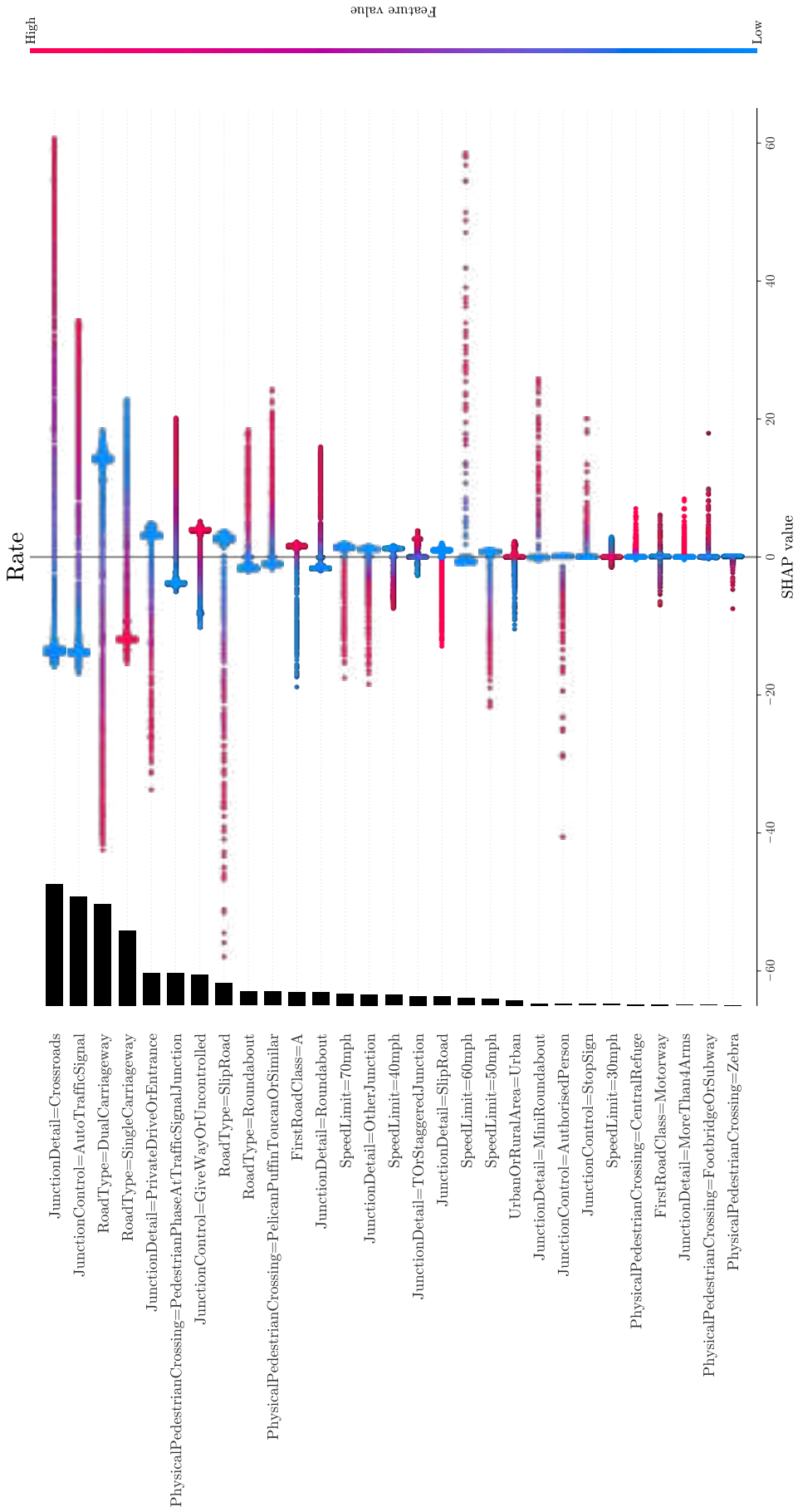


FIGURE 8.16: *SHAP value results for the MLP trained to predict accident rate, where each road junction observation is plotted according to the semi-spatial attribute categories.*

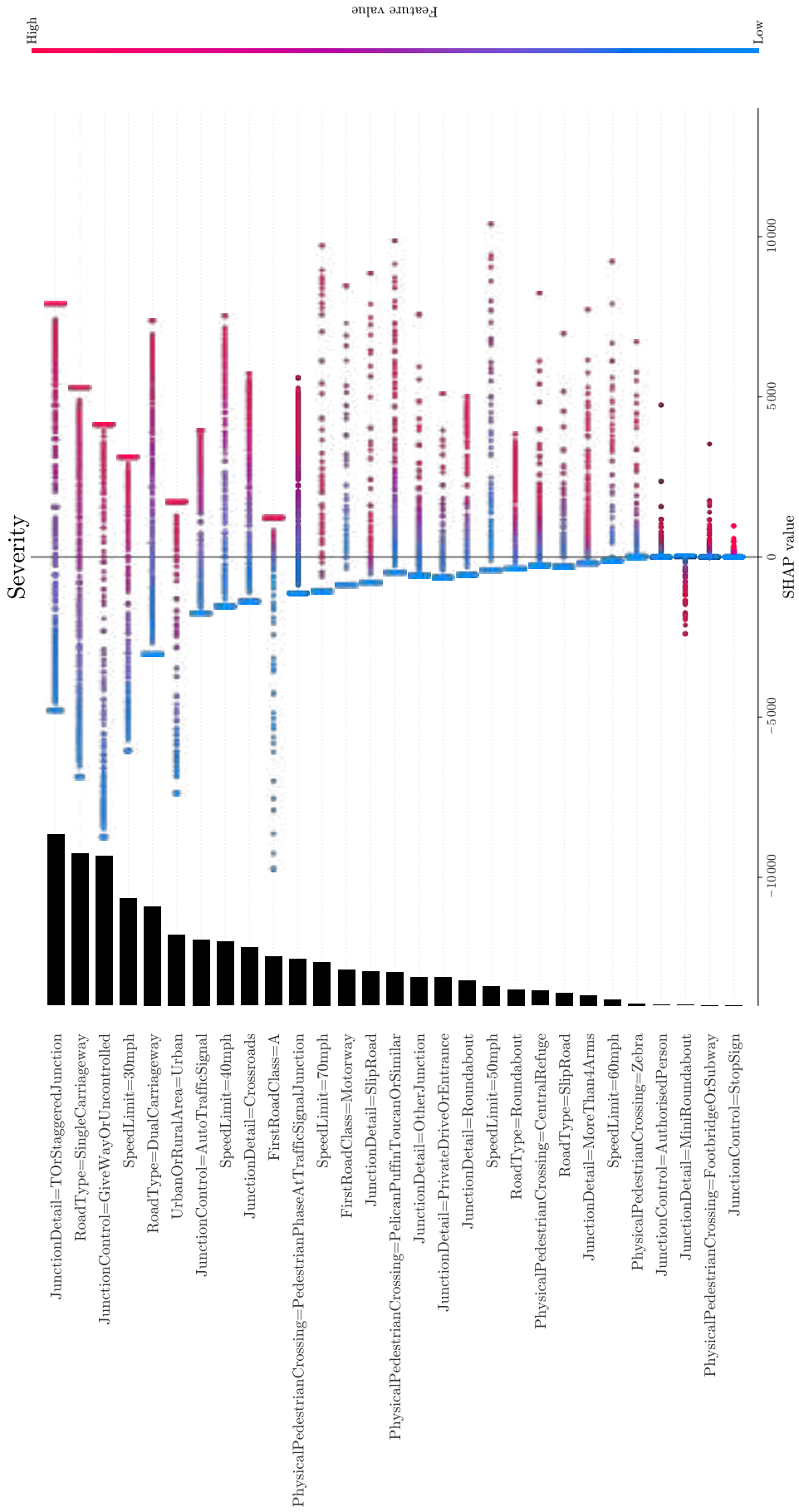


FIGURE 8.17: *SHAP value results for the MLP trained to predict accident severity, where each road junction observation is plotted according to the semi-spatial attribute categories.*

8.3.1 Road accident severity at crossroads controlled by traffic lights

For the purposes of this investigation, Sub-module 9.1 was employed to define a filtered subset of the combined RA data, defined as $\mathcal{X}_F = \mathcal{A}_{VR}\{\texttt{NumberOfVehicles} = 2, \texttt{JunctionDetail} = \texttt{Crossroads}, \texttt{JunctionControl} = \texttt{AutoTrafficSignal}, \hat{\mathcal{A}}_{VR} \in \hat{\mathcal{X}}_{NS}\}$, where only the RAs occurring at a crossroads with traffic lights are considered. This subset of filtered data contains a total of 4 001 accident observations. The categorical severity attribute `SevereAccident` (defined in §7.5) was then selected as the target attribute while all other non-spatial attributes subsequently define the set of explanatory attributes. These explanatory attributes were then one-hot encoded so as to proved more meaningful SHAP value interpretations. As described in 8.1.4, however, this leads to attributes that are highly correlated an so one accident attribute category was selected as a reference (as may be seen in Appendix B) and the corresponding accident attribute is removed.

A random 70% of the accident observations formed the learning set and the remaining observations the testing set. A total of seven classification learning models, described in Chapter 3, were selected for this classification task, namely the *Logistic Regression* (OR), *Naive Bayes* (NB), *Classification Tree* (CT), RF, kNN, SVM and MLP learning models. A constant model which always predicts the majority target variable class was also included to serve as a benchmark. A grid-based search technique was again employed to tune the hyper-parameters of each learning model.

The best OR learning model achieved an AUC performance of 0.617, employing L1 regularisation with a regularisation penalty of 170. The NB learning model has no hyper-parameter. The best CT learning model achieved an AUC performance of 0.597 and employed the MSE splitting criterion with a minimum of fifteen samples per leaf. The best RF learning model achieved an AUC performance of 0.664 and employed a hundred classification trees, each considering fifty random attributes when growing the tree. The best kNN learning model achieved an AUC performance of 0.521 and considered the seventeen nearest neighbours based on an inversely weighted Manhattan distance. The best SVM learning model achieved an AUC performance of 0.641 and employed an RBF kernel with a cost penalty term of 78. The best MLP structure achieved an AUC performance of 0.620 and had two hidden layers, containing four hundred neurons in both the first and second hidden layers.

After suitable parameter settings had been determined, each learning model was evaluated according to the AUC and F1 performance metrics, as previously described in §4.3.1, in respect of the learning and testing set, as per the procedure described in §6.9. These results are shown in Table 8.3. As may be seen in the table, the RF learning model achieved the best predictive performance when evaluated on the unbiased test set. All other models exhibited good predictive performance, which supports the hypothesis that non-spatial attributes are better predictors of accident severity than semi-spatial attributes. All possible classification thresholds are also illustrated graphically by the ROC plot in Figure 8.20. From these results, it may be concluded that all learning models are able to achieve a significantly better AUC performance than the constant classifier benchmark for all classification thresholds.

The SHAP values were then calculated for each learning model to provide insight into its working. By way of demonstration, the SHAP values corresponding the RF learning model trained to predict accident severity between two vehicles at crossroads with traffic lights is shown in Figure 8.19, while the SHAP values corresponding to all other learning models may be found in Appendix K. As may be seen in this figure, the RF learning model identified the `VehicleType=Car,PedalCycle`, `VehicleType=Car,MotorcycleOver500CC` and `VehicleType=Car,Motorcycle125CCAndUnder` as the first, second and fourth most important attributes, re-

TABLE 8.3: *The AUC and F1 performances achieved by the seven learning models and the constant model in respect of the learning and testing sets when the severity of an accident at a crossroad is selected as the target variable.*

	Learning		Testing	
Model	AUC	F1	AUC	F1
OR	0.914	0.922	0.711	0.882
NB	0.809	0.894	0.726	0.873
RT	0.959	0.940	0.622	0.876
RF	1.000	0.998	0.763	0.880
kNN	1.000	0.998	0.629	0.876
SVM	0.986	0.986	0.673	0.872
MLP	1.000	0.998	0.654	0.869
Const.	0.500	0.877	0.500	0.840

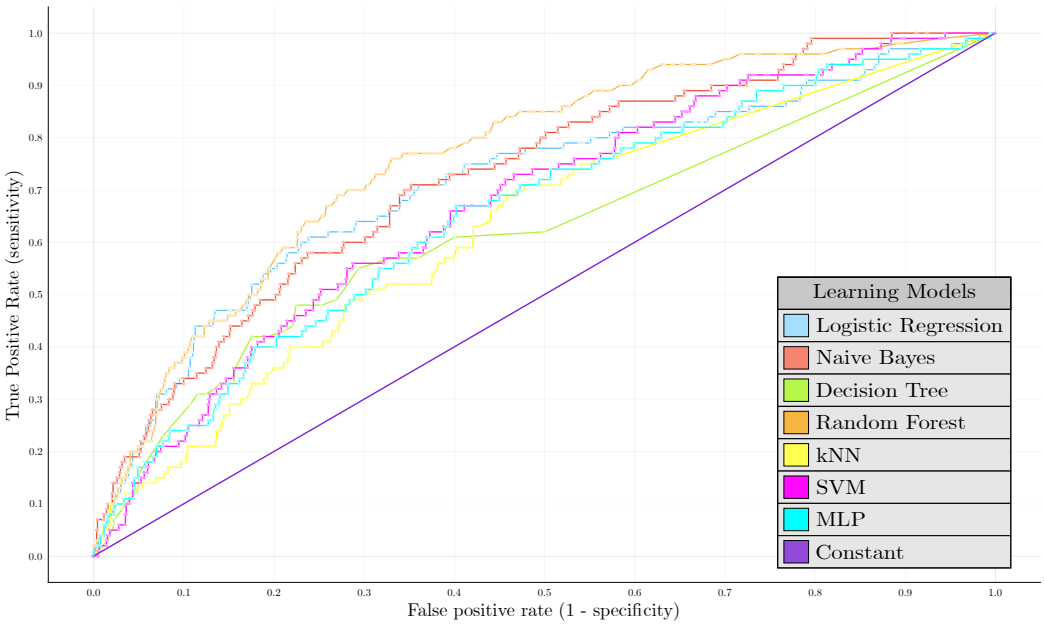


FIGURE 8.18: *An ROC plot showing the performance of the seven learning models and the constant model considered for predicting the severity of two-vehicle accidents at a crossroads controlled by traffic lights.*

spectively. More specifically, a higher relative accident severity is observed when these attribute values are true (indicated by the multiple red dots to the far right of the zero line for these attributes in Figure 8.19). This makes sense as one would expect a bicycle/motorcycle rider to sustain severe injuries if colliding with a car at a crossroads. The third most important attribute was `JunctionLocation=MidJunction,MidJunction` which suggests that accidents are more severe between two vehicles if they occurred in the centre of the crossroads. The fifth most important attribute was `VehicleManoeuvre=TurningRight,GoingAheadOther`. More specifically, a higher relative accident severity is observed when this attribute value is true. This is expected since, given that vehicles travel on the left-hand side of the road in the UK, vehicles turning right at a crossroads are likely to be in a more severe RA if colliding with a vehicle travelling in the opposite direction through the intersection. This RA occurrence may be due to vehicles travelling straight through the intersection, failing to stop during a traffic light's orange phase, which is typically when a vehicle turning right would execute its manoeuvre. The sixth most important attribute in Figure 8.19 further supports this conclusion.

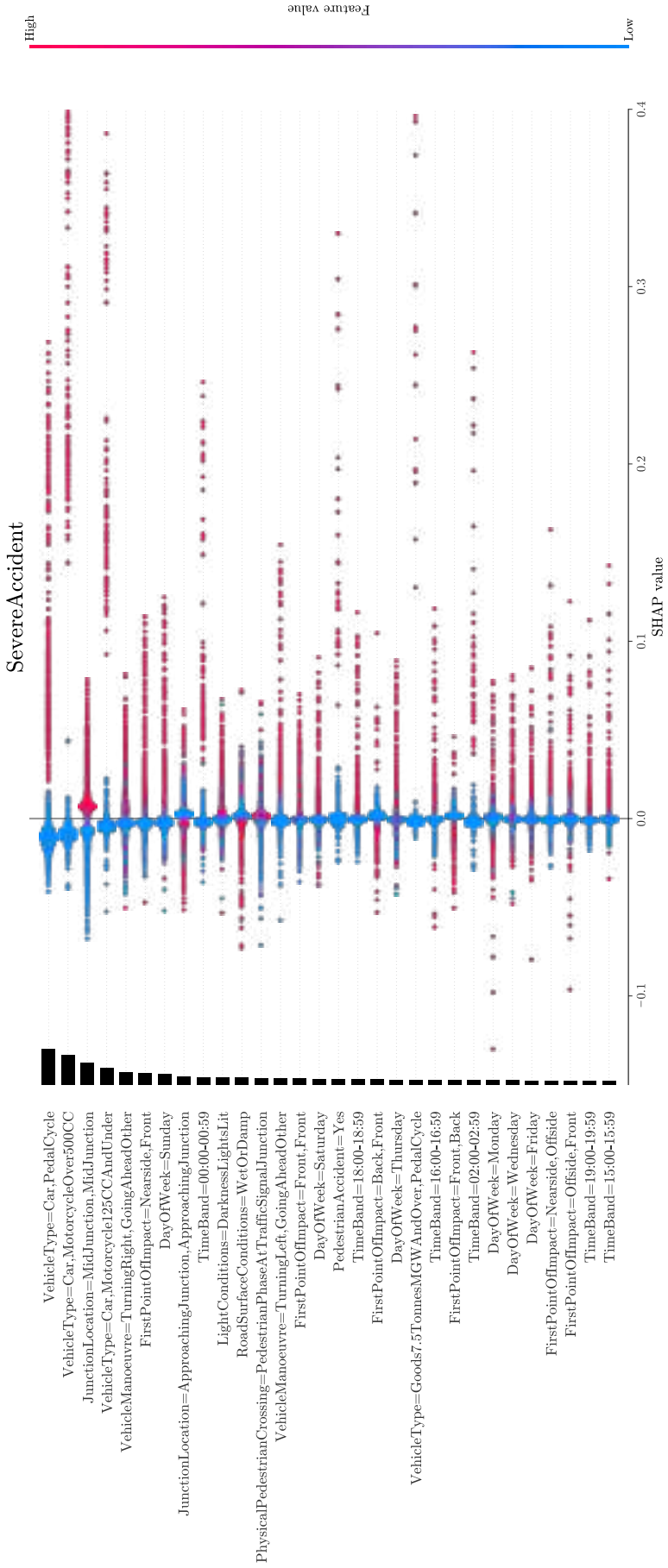


FIGURE 8.19: *SHAP value results for the RF learning model trained to predict accident severity, where each accident observation is plotted according to the non-spatial attribute categories.*

8.3.2 Road accident severity on a motorway

For the purposes of this second investigation, Sub-module 9.1 was employed to define a filtered subset of the combined RA data, defined as $\mathcal{X}_F = \mathcal{A}_{VR}\{\texttt{NumberOfVehicles} = 1, \texttt{FirstRoadClass} = \texttt{Motorway}, \texttt{JunctionControl} = \texttt{NotAtJunctionOrWithin50Metres}, \hat{\mathcal{A}}_{VR} \in \hat{\mathcal{X}}_{NS}\}$, where only the RAs occurring at a crossroads with traffic lights are considered. This subset of filtered data contains a total of 548 accident observations. The categorical severity attribute `SevereAccident` (defined in §7.5) was again selected as the target attribute while all other non-spatial attributes subsequently define the set of explanatory attributes.

The same seven learning models considered in the previous investigation were also investigated. The best OR learning model achieved an AUC performance of 0.672, employing L1 regularisation with a regularisation penalty of 20. The best CT learning model achieved an AUC performance of 0.635 and employed the MSE splitting criterion with a minimum of thirteen samples per leaf. The best RF learning model achieved an AUC performance of 0.702 and employed a hundred and eighty classification trees, each considering fifty random attributes when growing the tree. The best kNN learning model achieved an AUC performance of 0.654 and considered the thirty two nearest neighbours based on an inversely weighted Euclidean distance. The best SVM learning model achieved an AUC performance of 0.719 and employed an RBF kernel with a cost penalty term of 3.0. The best MLP structure achieved an AUC performance of 0.637 and had two hidden layers, containing thirty six neurons in both the first and second hidden layers.

After suitable parameter settings had been determined, each learning model was again evaluated according to the AUC and F1 performance metrics in respect of the learning and testing set, as per the procedure described in §6.9. These results are shown in Table 8.4. Again, the RF learning model achieved the best predictive performance when evaluated on the unbiased test set. All possible classification thresholds are illustrated by means of an ROC plot, shown in Figure 8.20, again illustrating that all learning models are generally able to achieve a significantly better AUC performance than the constant classifier benchmark for all classification thresholds.

TABLE 8.4: *The AUC and F1 performances achieved by the seven learning models and the constant model in respect of the learning and testing sets when the severity of an accident on a motorway is selected as the target variable.*

	Learning		Testing	
Model	AUC	F1	AUC	F1
LR	0.944	0.905	0.631	0.668
NB	0.839	0.823	0.662	0.714
RT	0.944	0.866	0.662	0.763
RF	1.000	1.000	0.705	0.763
kNN	1.000	1.000	0.642	0.705
SVM	1.000	0.960	0.690	0.740
MLP	1.000	1.000	0.679	0.755
Const.	0.500	0.668	0.500	0.668

As in the previous investigation, the SHAP values were then calculated for each learning model to provide insight into its working. The SHAP values corresponding the RF learning model trained to predict accident severity of a single-vehicle accident on a motorway is shown in Figure 8.21, while the SHAP values corresponding to all other learning models may be found in Appendix L. As may be seen in the figure, the RF learning model identified the `PedestrianAccident=Yes` as the most important attribute. Although infrequent, one would

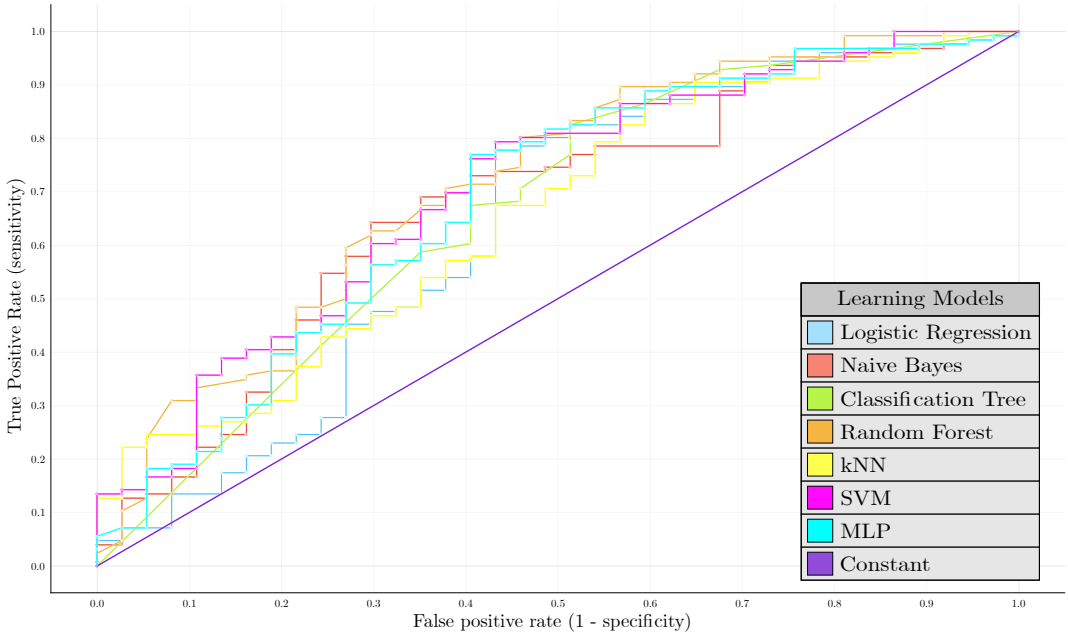


FIGURE 8.20: *An ROC plot showing the performance of the seven learning models and the constant model considered for predicting the severity of one-vehicle accidents on a motorway.*

likely expect an RA with a pedestrian on a motorway to be fatal. The second most important attribute was `VehicleType=MotorcycleOver500CC`. This result suggests that, if a motorcycle rider looses control of his or her vehicle on a motorway, the consequences are often severe. The third most important attribute was `SkiddingAndOverturning=SkiddingAndOverturned` which suggests that, if a vehicle rolls, the RA is often more severe. The fourth most important attribute is `HitObjectOffCarriageway=Tree`. One would expect a single-vehicle collision with a firmly fixed stationary object to be more severe, especially if that vehicle is travelling at motorway speeds. The fifth most important attribute is `RoadSurfaceConditions=WetOrDamp`. This result is interesting since it suggests that motorists are less likely to be involved in a serious accident when the motorway is wet. One may hypothesise that, when it is raining, vehicles typically travel at lower speeds since drivers perceive a higher risk of an RA, and, although a driver may be more likely to be in an RA when it is raining, the severity of such an accident is likely to be lower.

8.4 Chapter summary

The purpose of this chapter was to serve as a validation of Modules 7.0–9.0 in the RARE framework proposed in Part II by conducting multiple investigations into the context of the real-life Greater Manchester data prepared in Chapter 7. The chapter opened in §8.1 with a descriptive and predictive investigation into the road characteristics influencing the rate and severity of RAs along road segments so as to demonstrate how Module 7.0 may be employed in this context. The set of parameters required by Module 7.0 were first stated and substantiated in §8.1.1. Thereafter, a spatial analysis into the rate and severity along road segments ensued in §8.1.2, together with an investigation into the spatial distribution of various semi-spatial attribute values within the Greater Manchester road network. A demonstration of how a domain knowledge expert may use these results to identify and hypothesise why specific types of accidents occur at specific locations was also provided. This served as a basis for the non-spatial investigation, conducted in

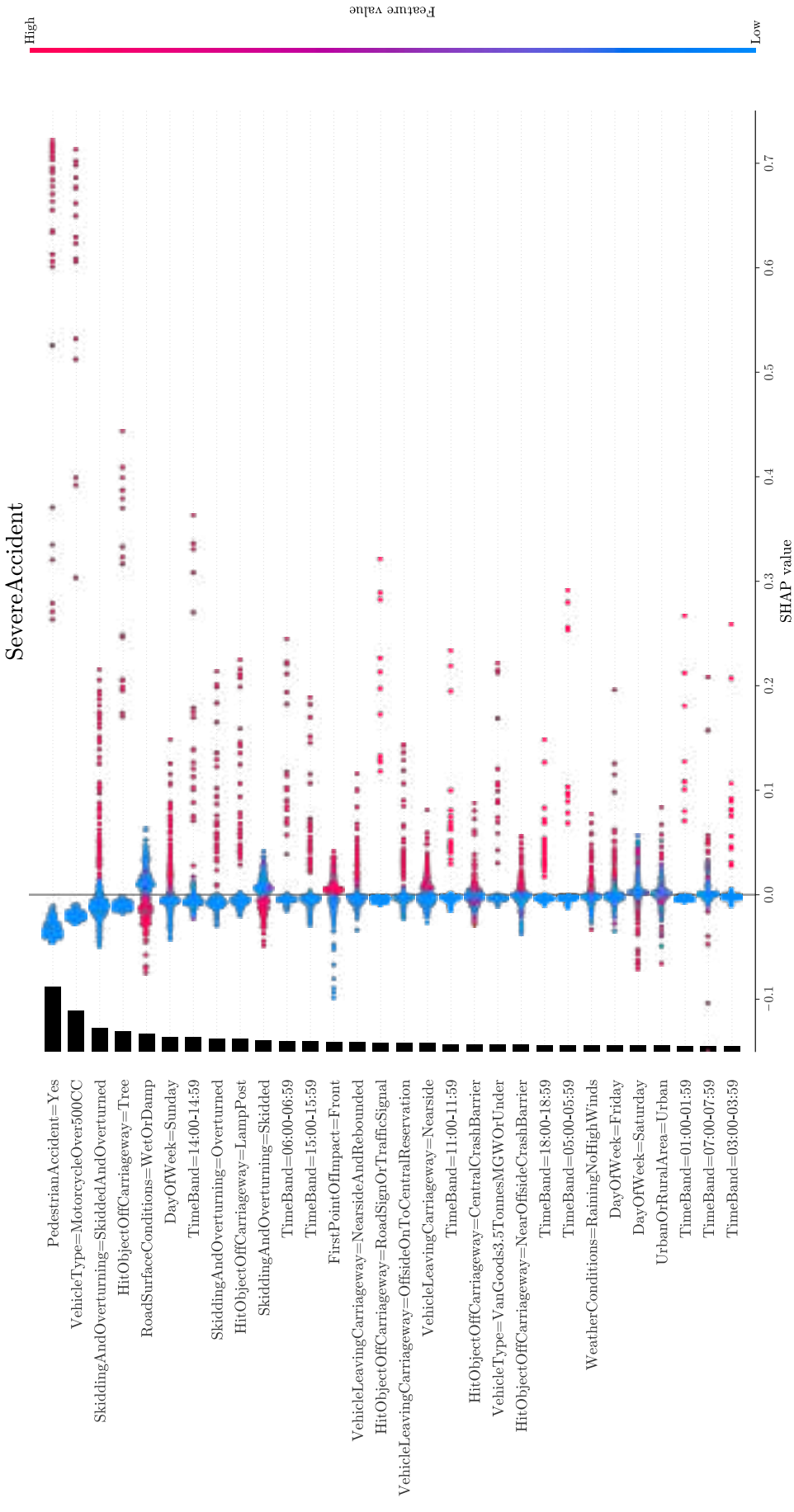


FIGURE 8.21: *SHAP value results for the RF learning model trained to predict accident severity, where each accident observation is plotted according to the non-spatial attribute categories.*

8.1.3, which employed multiple scatter plot representations of the average accident rate against the average accident severity corresponding to different road characteristics. Attention then shifted, in §8.1.4, to demonstrating how Module 9.0 may be employed to obtain a predictive analytic view of the results, which began by first describing the parameter choices associated with Module 9.0. This was followed, in §8.1.5, by an assessment of the predictive performance of six learning models, namely the LR, RT, RF, kNN, SVM and MLP learning models. An interpretation of the SHAP values associated with the MLP learning models trained to predict accident rate and severity along a road segments. A similar investigation into the accident rate and severity of accidents occurring at junctions then ensued in §8.2 and provided clarity to many of the observations made in §8.1. The chapter finally closed in §8.3 with two investigations concerning non-spatial attributes which are most telling of more severe accidents at defined spatial locations on the road.

Part IV

Conclusion

CHAPTER 9

Dissertation summary & appraisal

Contents

9.1 Dissertation summary 175
9.2 Appraisal of dissertation contributions 178

This chapter comprises two sections. In §9.1, a chapter-by-chapter overview of the research documented in this dissertation is provided, together with how the eight objectives presented in §1.3 were accomplished. This is followed in §9.2 by an appraisal of the novel contributions made in this dissertation.

9.1 Dissertation summary

Apart from Part IV, the current part, this dissertation comprised a further eight chapters, seven of which were partitioned into three parts. The introductory chapter, Chapter 1, was the only stand-alone chapter and served the purpose of formally establishing the problem considered in this dissertation. This chapter opened with a discussion on the origin and background of RAs and how research efforts into their prevention and severity reduction have greatly reduced the number of RA deaths over the years. The focus then shifted to the state of future research in this field. It was argued in the runup to the formal problem statement that there is a need for a data mining framework for conducting RARE as well as how leveraging the advances in ML could lead to the development of more powerful RA predictive models. This was followed by an elucidation of the eight research objectives pursued in this dissertation, together with a delimitation of the scope of the research undertaken. The chapter concluded with an outline of how the material presented was to be organised into the various chapters of this dissertation.

The first part of this dissertation was a literature review and consisted of three further chapters. In the first of these, Chapter 2, Objective I(b) was initially pursued by highlighting the most notable descriptive road safety approach towards modelling RAR. More specifically, the three main components of an RA — the human, vehicle and environment components — were identified and the results of five independent studies were then presented, all of which identified the human component as the leading cause of RAs, in fulfilment of Objective I(a). Thereafter, the three well-known dimensions of road safety were reviewed and elaborated upon. The Haddon matrix served to unify the three dimensions of road safety and the components of an RA into a single framework for categorising the various RAR factors. The focus then shifted to a review of various predictive models that have been applied in the road safety domain, together

with associated data and methodological issues that have to be considered during a predictive investigation, in fulfilment of Objective I(b). The chapter closed with a critical reflection on the road safety modelling approaches adopted in the literature with a view to provide direction for the modelling approach adopted in this dissertation.

The second chapter of Part I, Chapter 3, was devoted to a review of the methodologies employed in the field of GIS when attempting to model geographic features of the real world accurately. Object-based and field-based conceptual models were first introduced, together with their respective vector and raster data representations, in fulfilment of Objectives I(c). Thereafter, a description followed of techniques for performing spatial analyses exclusively in a network space, in pursuit of Objective I(d). More specifically, the reader was introduced to the notion of a homogeneous binomial point process and how it is used to derive the CSR hypothesis in statistical spatial analyses. This was followed by a description of how any non-uniform distribution can be transformed into one that is uniform by means of the uniform network transformation, and finally, that cognisance should be taken of edge biases when defining artificial network boundaries. The chapter finally closed with a discussion on two PPA methods, the first of which was the network NND method for testing whether points in a network are statistically close to one another or dispersed within the network space, while the second method was the network KDE method for estimating the density of points on a network.

The third and final chapter of Part I, Chapter 4, contained a review of the literature pertaining to ML in general. In pursuing Objective I(e), the reader was first introduced to the four learning paradigms, namely supervised, unsupervised, semi-supervised and reinforcement learning. This was followed by a review of the practical considerations associated with the various data formats and how data should be cleaned appropriately, together with a description of the data splitting methods of cross-validation and bootstrap re-sampling often employed during the evaluation of a learning model. Thereafter, additional considerations related to the learning model were described, which included an appropriate choice of performance evaluation measure, the notion of a learning model overfitting or underfitting the training data, and the practical considerations associated with model interpretability. The recently introduced model-agnostic SHAP method was identified as a promising approach for interpreting ML models that are inherently black-box. A description of six well-known ML methods then ensued, starting with a description of the notion of GLMs. This notion was then used to derive the well-known OLSR and logistic regression learning models. Attention then shifted to decision tree-based methods, where the reader was introduced to the CART, C4.5 and random forest learning models. This was followed by a detailed description of four further learning models, namely the naive Bayes classifier, the kNN classifier, SVM and MLP, in fulfilment of Objective I(f).

The second part was the heart of the dissertation and was concerned with the overall system design of the proposed RARE framework and consisted of two chapters. The first of these, Chapter 5, was devoted to a top-down description of the design and intended working of the generic RARE framework, in pursuit of Objective II. First, the reader was introduced to the various phases associated with the data mining process in terms of the CRISP-DM reference model. Thereafter, this reference model was used to formulate a high-level data mining framework which encompasses the phases concerned with configuring, enhancing and transforming the data into useful information. This data mining framework was subsequently used as the basis for designing a high-level architectural overview of the proposed RARE framework, which is modular in nature. After a detailed description of the required accident, vehicle and road data sets, the data flow between individual module components of the framework was described by means of appropriate DFDs.

In the second and final chapter of Part II, Chapter 6, the focus shifted to an elucidation of the working of each of the modular components of the proposed RARE framework, in fulfilment of Objective III. More specifically, a description was proffered of how Module 1.0 in the framework may be used to facilitate the categorisation of attributes according to their data types and assisting with the imputation of missing values, in pursuit of Objective III(a). Thereafter, a novel method for discerning so-called semi-spatial attributes (*i.e.* attributes that, although captured at the scene of an accident are, in fact, inherent properties of the road itself) from non-spatial attributes by means of the network NND method of Okabe *et al.* [194] was proposed and utilised in Module 2.0 for data categorisation of an accident attribute as being either spatial, semi-spatial or non-spatial. It was further described how data are formatted and categorised according to the Haddon matrix in this module. Thereafter, the algorithmic working of Modules 3.0 and 4.0 for respectively merging vehicle and road records with corresponding accident records were elucidated for the purpose of creating an all-encompassing combined data set, in fulfilment of Objective III(b). Thereafter, a discussion ensued on how a data set could further be enhanced through the creation of new information based on the data currently stored in the database, in pursuit of Objective III(c). This involved utilising Module 5.0 which allowed for the creation of accident severity attributes, as well as Module 6.0 which proposed a procedure for facilitating the construction of a new attribute by the user, based on one or more current attribute value(s). The focus then shifted to a description of Modules 7.0 and 8.0 which presented a novel approach for quantifying RAR along road segments (through the incorporation of the equal split network KDE method of Okabe *et al.* [193]) and at road junctions, respectively, in fulfilment of Objectives III(d) and III(e). It was finally described how these results are then taken as input to Module 9.0 where a supervised learning procedure and the SHAP method are employed so as to identify the most indicative factors which contribute to higher levels of RAR, in fulfilment of Objective III(f).

The third part of this dissertation was concerned with the application of the proposed RARE framework to a real-world case study and consisted of two chapters. The first of these, Chapter 7, opened with a description of how a concept demonstrator of the proposed RARE framework was implemented on a personal computer, as per Objective IV, in the Orange Data Mining software suite [70], where all purpose-built sub-modules were developed in the Python 3.7.0 programming language. This was followed by a detailed description of the Great Manchester accident, vehicle, road and victim data sets obtained for the purpose of conducting a real-world case study, in fulfilment of Objective V. The stipulations of Objective VI were then pursued in the remainder of the chapter by demonstrating how Modules 1.0–6.0 could be used to prepare the various data sets appropriately. An explanation was provided of how attributes were categorised, together with a description of how values missing from the data were imputed appropriately. This was followed by an investigation into the discovery of semi-spatial attributes present in the accident data set which resulted in the identification of seven semi-spatial attributes. Thereafter, the application and verification of the two proposed algorithms for merging vehicle and road records with their corresponding accident records was illustrated, allowing for the creation of an all-encompassing, combined data set. Finally, a description was provided of how additional attributes were created and appended to the current combined data set.

The second and final chapter of Part III, Chapter 8, was a demonstration of the intended outputs of Modules 7.0–9.0, consisting of multiple RAR investigations with respect to the data prepared in Chapter 7, in pursuit of Objective VII and the eight elements of the envisioned RARE framework outputs, as illustrated in Figure 1.4. The first investigation was concerned with quantifying the rate and severity of RAs along road segments in Greater Manchester and interpreting the spatial and attribute distribution representation of these results. From the spatial interpretation, it was concluded that higher rates of accidents tend to occur in highly urbanised settings likely due to the higher densities of road junctions in these areas, while mo-

torways tend to have lower accident rates. Using these results, the next step was to identify the road characteristics which lead to higher RA rates and severities. The SHAP value results of the random forest model confirmed that the road segments which formed part of junctions typically experienced higher rates of RAs than those which did not form part of a junction, as well as revealing that both single and dual carriageways typically experience higher accident severities. The next investigation was concerned with quantifying the rate and severity of RAs at road junctions in Greater Manchester. The interpretation of the spatial and attribute distribution results representation corroborated with the results obtained when analysing road entities as segments. The last investigation was concerned with attempting to determine which driver- and vehicle-related attributes lead to higher RA severities at multiple road locations for both one- and two-vehicle accidents.

9.2 Appraisal of dissertation contributions

The novel contributions of this dissertation are six-fold and are listed in this section in the order in which they first appear in the dissertation.

Contribution 1 *The design of a generic framework for conducting multiple risk evaluations of RAs which draws from the separate fields of traditional road safety analysis, network spatial analysis and ML.*

The overarching aim in this dissertation was to design and demonstrate the practical workability of a generic framework capable of configuring and enhancing RA data in order to facilitate multiple evaluations of RAR. This was achieved through the process of data mining by incorporating various ideas, techniques and methods from three separate fields of research, namely traditional RA analysis, network spatial analysis and ML, reviewed in Chapters 2, 3 and 4, respectively. A high-level architectural design of a modular RARE framework was formally presented and described in Chapter 5, and was supplemented in Chapter 6 with a documentation of the inner working of the individual modules comprising the proposed framework. The culmination of the ideas presented in this dissertation was realised in Chapters 7 and 8, where a concept demonstrator of the proposed RARE framework was implemented in the Orange Data Mining Software and all purpose-built sub-modules were subsequently developed in the Python 3.7.0 programming language. The overall working of this concept demonstrator was demonstrated through the application of the real-world case study, presented in Chapters 7 and 8.

Contribution 2 *The introduction of the notion of a semi-spatial attribute and a novel method by which such an attribute may be identified in the context of point and line geometry.*

In this dissertation, the notion that selected RA attributes contain inherent spatial information was advocated. In essence, this spatial information refers to attributes of an RA that, although captured at the scene of an accident, represent explicit characteristics of the road (and not of the accident itself). It was further argued that a distinction should be made between these attributes since, during an investigation, one may wish only to consider attributes that describe the actions of the driver and not attributes which capture the environmental circumstances of an RA. Building on the notions of spatial and non-spatial attributes in the GIS literature, it was proposed that a new class of attributes, called semi-spatial attributes, should be defined as attributes which implicitly imply a set of possible locations (or representations) that may be assumed by an object. In other words, although it may not be possible to define explicitly the location of an object based on the information conveyed by semi-spatial attribute (as opposed

the case of a spatial attribute), one may determine a subset of possible locations in which the object ought to be contained. This notion, therefore, serves as an addition to the hierarchical structure of the object-based geospatial model framework of Figure 3.2. During the design of Sub-module 2.3 of the proposed RARE framework in §6.6, a novel method for explicitly identifying semi-spatial attributes in the context of point and line geometry (*i.e.* accident points and road line segments) was also formally proposed. This was achieved by utilising the NND method of Okabe *et al.* [194] to construct a spatial score which serves as a quantitative measure of the amount of spatial information possessed by attributes in an RA data set. The effectiveness of this approach was formally demonstrated in §7.3 in the context of the real-world case study.

Contribution 3 *The design of a novel algorithm for merging two data sets exhibiting an inherent one-to-many relationship.*

During the design of Module 3.0, Algorithm 6.1 was proposed for ordering vehicle records in such a way as to nullify the arbitrary choice of vehicle references made by a data recorder. This allowed for the construction of a single representative vehicle record *via* a comma-delimited approach, while still preserving the integrity of the individual vehicle records — a technique that has, to the best of the author's knowledge, not been described formally in the literature. This technique is not only limited to the data sets described in this dissertation but may be applied to combine any two data sets exhibiting an inherent one-to-many relationship. The usefulness of this approach was formally demonstrated in §8.3 in the context of the real-world case study.

Contribution 4 *The design of a novel method for quantifying the rate and severity of RAs along road segments.*

During the design of Module 7.0 of the proposed RARE framework in §6.8, a novel method for defining and quantifying the rate and severity of RAs along road segments was proposed. In this formulation, the network KDE method of Okabe *et al.* [193] was employed to estimate the density of accidents along a road network. This was employed for the purpose of defining an RA rate and severity measure, expressed in (6.4) and (6.5), respectively, as well as a measure used to infer inherent road characteristics from RAs (*i.e.* semi-spatial attributes) to a specific road segment observation, expressed in (6.6). This approach also relied on a new formulation of accident severity proposed during the design of Module 5.0 for the case in which a categorical attribute deemed to represent a measure of RA severity must be converted to a continuous RA severity measure. This was achieved by introducing a scaling function which allows a user to easily adjust the quantitative distances between the qualitative categories of a categorical severity attribute as he or she sees fit. This method for quantifying and disaggregating RAR along road segments in terms of various road characteristics and viewing the rate and severity measures along road segments in unison has, to the best of the author's knowledge, not yet been employed in the RA analysis literature. The effectiveness of this approach was formally demonstrated in §8.1 in the context of the real-world case study.

Contribution 5 *The design of a novel method for grouping a set of accidents occurring at the same road junction together.*

During the design of Module 8.0 of the proposed RARE framework in §6.8, a novel method was proposed for grouping a set of accidents occurring at the same road junction together. This was achieved by proposing a junction dissimilarity score which allowed accidents which occurred geographically close to one another and possessed the same (or similar) semi-spatial attributes to be clustered together. This method for clustering accidents in order to construct road junction

observations and viewing the rate and severity measures at road junctions in unison has, to the best of the author's knowledge, not yet been employed in the RA analysis literature. The effectiveness of this approach was formally demonstrated in §8.2 in the context of the real-world case study.

Contribution 6 *The application of the SHAP method for interpreting the rule formulations of previously uninterpretable ML models in the RA analysis domain.*

As described in §2.3.2, the most popular predictive modelling methods in the RA analysis literature are based on the GLM framework, primarily due to their ability to produce interpretable coefficients for each attribute included in the model [297]. The adoption of ML models in the RA analysis literature has, however, been relatively limited compared to traditional statistical modelling approaches which may be primarily attributed to the black-box nature of many ML models which inhibits the appropriate interpretation of model results [176, 297]. In this dissertation, the newly proposed model-agnostic AFA method, called the SHAP method, was employed to combat this problem. The SHAP method has, to the best of the authors knowledge, not been utilised in the RA analysis literature for interpreting the results of black-box learning models employed in this dissertation (*i.e.* random forests, the kNN classifier, SVM and MLP). The effectiveness of this approach was formally demonstrated in §8.1, §8.2 and §8.3 in the context of the real-world case study. The hope is that this will encourage future application and exploration of ML models for predictive analysis in the RA research domain.

CHAPTER 10

Future work

Contents

10.1 Improvements to the current RARE framework 181
10.2 Additions to the RARE framework 182
10.3 Promoting the RARE framework 185

This final chapter details nine suggestions which may be pursued as follow-up work towards the improvement, enhancement and promotion of the RARE framework proposed in this dissertation, in fulfilment of Objective VIII of §1.3. In the case of each suggestion, a brief description and motivation is provided.

10.1 Improvements to the current RARE framework

This section contains three suggestions for future work related to possible improvements of the existing RARE framework proposed in this dissertation.

Suggestion 1 *Improve the algorithm for projecting points onto their closest line segment.*

During the design of Module 4.0 in §6.4, Algorithm 6.2 was proposed for the association of accident points with road segments. Although this algorithm proved effective (as described in §7.4), the application of the algorithm proved to be highly time-consuming. This is because the algorithm itself is naive in nature in that it considers the possibility of projecting each of the m_A accident points onto *all* of the m_R road segments, resulting in an algorithm with an $\mathcal{O}(m_A m_R)$ time complexity. Upon further investigation, there may be an opportunity to decrease the overall time complexity of the algorithm by utilising a quadtree data structure or exploring the use of Voronoi diagrams [66].

Suggestion 2 *Circumvent the functional form and shape choice of KDE base kernel by employing notions from multiple and infinite kernel learning.*

During the case study investigation of §8.1.1, it was necessary that an appropriate base kernel function K and associated bandwidth ω be specified in order to compute the rate and severity of RAs along a road segment. The choice of base kernel function and associated bandwidth is,

however, not entirely obvious or justifiable. It may, therefore, be worth investigating the notions employed in multiple and infinite kernel learning to circumvent the functional form and shape choice of KDE base kernel. In the *multiple kernel learning* (MKL) framework it is possible to learn weighted combinations of a predefined finite set of n_K kernels $\mathcal{K} = \{\kappa_1(\cdot), \ldots, \kappa_{n_K}(\cdot)\}$ in order to define an appropriate similarity measurement between the vectors $\boldsymbol{x}^{(i)}$ and $\boldsymbol{x}^{(j)}$, which takes the form

$$K_\beta(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}) = \sum_{i=1}^{n_K} \beta_i \kappa_i(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}) \tag{10.1}$$

where $\beta_i \geq 0$ for all $i \in \{1, \ldots, n_K\}$ and $\sum_{i=1}^{n_K} \beta_i = 1$. Özöğür-Akyüz and Weber [200] extended this notion by introducing *infinite kernel learning* (IKL) which introduces infinitely many kernel variations contained in the infinite set Ω, expressed mathematically as

$$K_\beta(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}) = \int_\Omega \kappa_i(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}, \tau) \, \mathrm{d}\beta(\tau), \tag{10.2}$$

where $\tau \in \Omega$ is a kernel parameter and $\beta(\cdot)$ is a monotonically increasing function which integrates to one [199].

In the context of the work proposed in this dissertation, the idea would, therefore, be to express the base kernel function K and associated parameters (including the function's bandwidth ω) in the form of (10.1) or (10.2) and solve an optimisation problem with the objective of maximising some defined fitness metric of the KDE, defined in (3.11). It is not clear, however, how this fitness metric should be defined in the context of quantifying RAR.

Suggestion 3 *Improve the algorithm for defining junction observations.*

During the design of Module 8.0 in §6.8, a novel method was proposed for grouping a set of accidents occurring at the same road junction together. Although this method proved effective during the case study of this dissertation, no formal metric was used to determine the junction dissimilarity threshold β required in Sub-module 8.2. Rather, this value was empirically determined using the spatial results produced in Sub-module 8.4, as shown in Appendix D. It may, therefore, be beneficial to construct a formal metric for measuring the quality of the geographical junction clusters. One possible suggestion may be to manually group a subset of accidents into correct junction clusters and assess the accuracy of different β-values and similarity measures according to this labelled subset. Furthermore, a more accurate junction dissimilarity score could be determined if the Euclidean distance between points (*i.e.* the quantity $d_\triangle(\boldsymbol{p}^{(i)}, \boldsymbol{p}^{(j)})$ in (6.7)) is substituted with the network distance between points.

10.2 Additions to the RARE framework

This section contains four suggestions for future work related to possible additions to the existing RARE framework proposed in this dissertation.

Suggestion 4 *Incorporate the notions of utility theory to construct a user-specific function of RAR.*

The notion of risk is, for the most part, a highly subjective concept and depends largely on the a specific context under consideration. During the design of Modules 7.0 and 8.0 in §6.4, risk

was defined, in the context of RAs, to be a function of two dimensions, namely: (1) The rate and (2) the severity of being involved in an RA. This formulation is arguably justified since it aligns with the two central risks associated with road transport, namely: (1) The damage/loss of property and (2) the threat to or loss of human life. It is, however, difficult to formulate a single accepted measure of RAR due to the subjectivity involved in these two facets of risk.

Consider, for example, the hypothetical case in which a driver must choose between taking one of two roads A or B, given that the outcome of this choice is governed by the two lotteries shown in Figures 10.1(a) and (b), respectively. If the driver chooses to travel along road A, there is a 90% chance that the driver will not experience an RA, while there is a 10% chance that the driver will indeed be involved in an accident. If, however, the driver is involved in an RA, it is known for certain that the driver will escape the accident uninjured. Contrast this to the choice of travelling along road B, where there is only a 0.1% chance that the driver will be involved in an RA. It is, however, known for certain that, if the driver is indeed involved in an RA on road B, the accident will be fatal. The question is which of these two roads the driver would prefer to take? A risk-averse driver would likely prefer travelling on road A since, although there is an abnormally high probability of being involved in an RA, there is a 0% chance of being killed. A risk-seeking driver, on the other hand, may prefer travelling on road B since there is an almost negligible chance of being involved in an RA, despite the fact that an accident on this road will be fatal. This notion of determining a decision maker's preference under uncertainty when presented with a choice is fundamental in the field of utility theory [289] which may be worth exploring for the purpose of constructing a single user-specific measure of RAR.

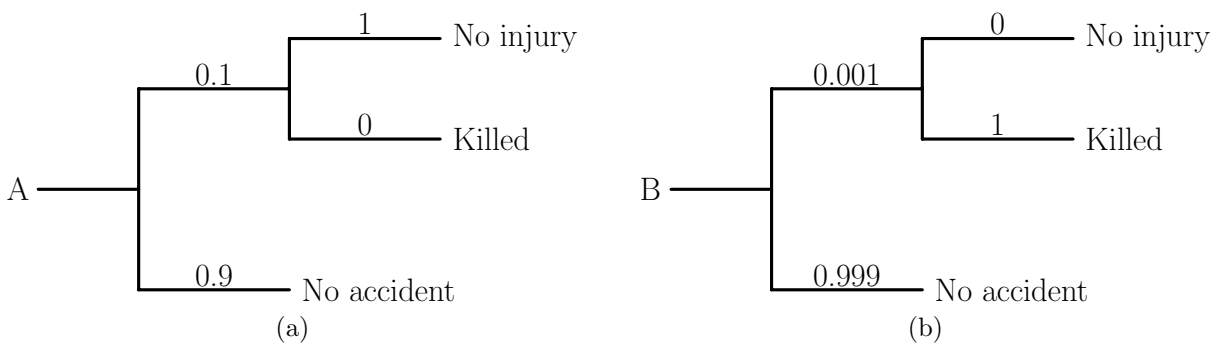


FIGURE 10.1: *Lotteries associated with choosing (a) road A or (b) road B, demonstrating the choice that a user must make in terms of the likelihood and consequence of being involved in an RA.*

Suggestion 5 *Formulate a routing strategy that minimises both the total rate and severity of an RA.*

A notion was introduced in §8.1 whereby one may use the results produced in Module 7.0 to find a route which minimises either the RA rate or severity experienced when travelling between two points in a network. Building on this idea, consider, for instance, the hypothetical network shown in Figure 10.2(a) composed of eight links (denoted by $\ell^{(1)}$–$\ell^{(8)}$), where road segment $\ell^{(k)}$ has an associated RA rate

$$R(\ell^{(k)}) = \sum_{j=1}^{g_k} R(\ell_j^{(k)})$$

and severity

$$S(\ell^{(k)}) = \sum_{j=1}^{g_k} S(\ell_j^{(k)}).$$

If one were to travel from node A to node B by traversing the links of this network, then there exists nine possible routes one may take if no single node is visited more than once. The total RA rate and severity of each route may, therefore, be plotted, as shown in Figure 10.2(b), where each point indicates the sequence of traversed links forming a route from node A to node B. From this plot, an optimal route when considering both of these risk measures would be any one of the four non-dominated routes lying on the Pareto front (indicated by the dotted line). Again, if Suggestion 4 is pursued, a single measure of RAR could be determined for each route allowing the bi-objective formulation (shown in Figure 10.2(b)) to be transformed into a single objective problem. This would thus allow one to determine a single user-specific route which minimises the total RAR experienced when travelling from a point A to a point B in a road network.

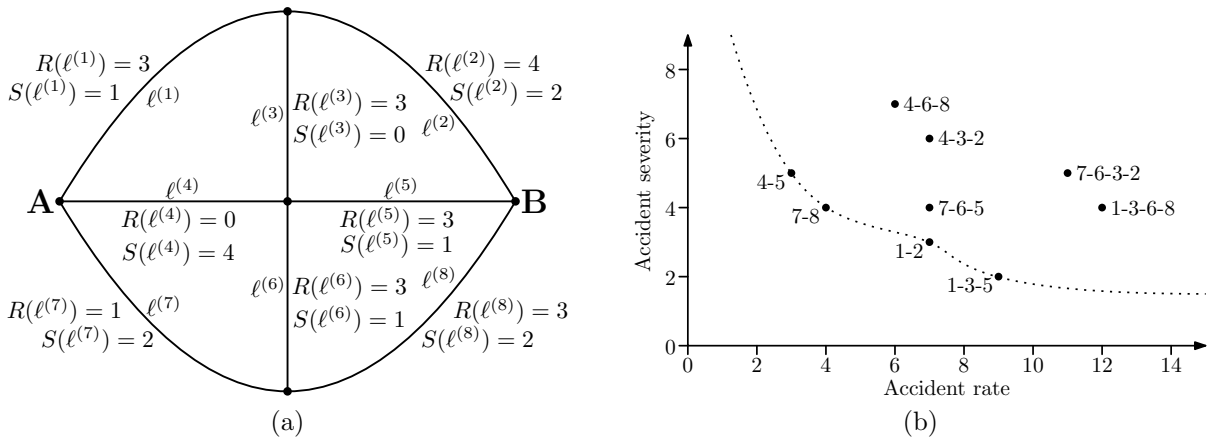


FIGURE 10.2: *(a) A hypothetical network composed of eight links and (b) a plot of the accident rate and severity corresponding to each of the nine possible routes.*

Suggestion 6 *Create new attributes which describe the surroundings of an RA by using a land cover map.*

In the case study conducted in Part III, the attribute `UrbanOrRuralArea` was used to describe the land surrounding the location of a road accident. Although this attribute proved useful in the case study, it conveys very limited information about other natural and man-made features present at the scene of an accident (*e.g.* trees, water, buildings). It would, therefore, be worthwhile to develop a new module in the proposed RARE framework (forming part of the Create class) which creates new RA attributes, each of which indicate the absence or presence of a specific land characteristic. This may be achieved by utilising a *land cover map* which classifies the individual cells of a raster representation of land according to specific categories, like that shown in Figure 10.3(a). Since the latitude and longitude position of an RA is typically known, this point would lie within one of the raster cells of the land cover map, as shown in Figure 10.3(b). A circle of raster cells surrounding the RA (indicated by bold cells in Figure 10.3(b)) could then be identified and used to indicate the presence or absence of specific land types.

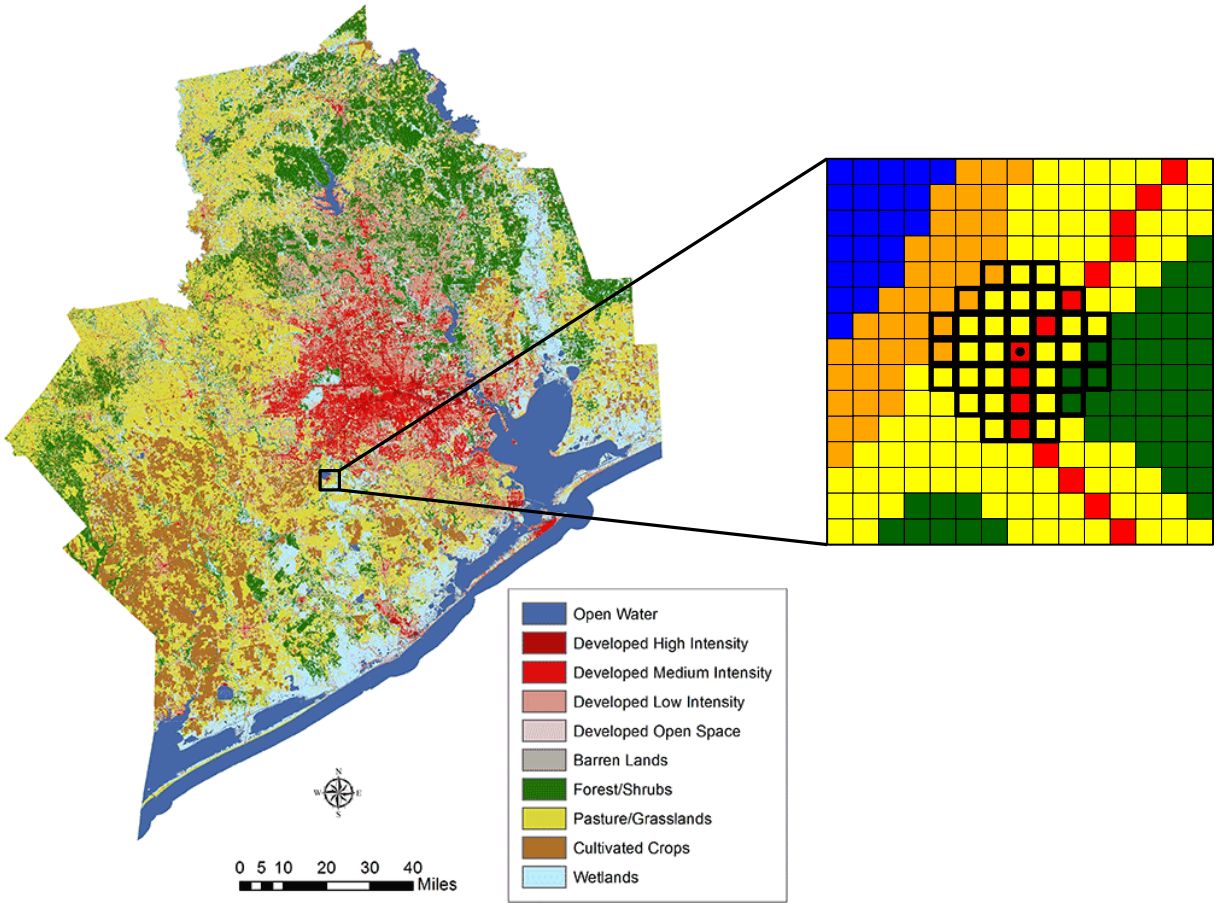


FIGURE 10.3: *A land cover map showing a zoomed-in view of a circle of raster cells surrounding an RA [110].*

Suggestion 7 *Investigate the performance of other ML models according to their ability to predict RA rate along road segments and at road junctions.*

During the case study part of this dissertation, six supervised learning algorithms were evaluated according to their ability to predict the newly proposed RA rate and severity measures in the cases where observations were road segments and road junctions. It was found, in the case of predicting RA rate, that the random forest learning model performed significantly better than the other learning models when only one attribute is used to construct the individual decision trees — a finding which begs the question whether other tree-based/ensemble learning models may also achieve similar or superior performance in this context. Consequently, it may be worthwhile investigating the predictive performance of gradient-boosted decision trees (or XGBoost) [59] and the Adaptive Boosting (or AdaBoost) [82] learning models in predicting RA rate along road segments and at road junctions. Other promising deep learning models include the newly proposed neural architecture search method [302] which implicitly determines a near-optimal neural network structure and weights in a single optimisation procedure.

10.3 Promoting the RARE framework

This section contains two suggestions for future work related to how the RARE framework proposed in this dissertation may be promoted for use in other research investigation.

Suggestion 8 *Conduct further case studies using RA data from multiple countries.*

In Part III of this dissertation, a case study was conducted using real-world RA data from Greater Manchester in the UK so as to demonstrate the intended functionality of the proposed RARE framework. To truly demonstrate the *generic* nature of the proposed framework, however, more case studies may be conducted in respect of real-world data from different countries. Such an undertaking would also facilitate interesting country RAR comparisons, in which Module 6.0 is first used to standardise the categories of specific categorical attributes in the different RA data sets (as described in §5.3.3) after which Modules 7.0, 8.0 and 9.0 are used to quantify and compare, for road segments and at road junctions, respectively, the rate and severity of the accidents experienced in the respective countries.

Suggestion 9 *Develop graphical user interfaces for the widgets developed in the Orange Data Mining Software.*

Since the RARE framework proposed in this dissertation is inherently modular, the concept demonstrator development employed the open-source Orange Data Mining API [70] as the framework basis (as described in §7.1), where all purpose-built sub-modules were developed in the Python 3.7.0 programming language. It would be beneficial if graphical user interfaces were to be designed for each of the purpose-built sub-module widgets developed in the Orange Data Mining Software. Such an endeavour would make the RARE framework sub-modules even more accessible to a much wider research audience and, as stated in §5.2, would also allow these models to be re-used or re-purposed in other problem application domains.

References

[1] Abdelwahab HT & Abdel-Aty MA, 2002, *Artificial neural networks and logit models for traffic safety analysis of toll plazas*, Transportation Research Record, **1784(1)**, pp. 115–125.

[2] Aguero-Valverde J & Jovanis PP, 2008, *Analysis of road crash frequency with spatial models*, Transportation Research Record, **2061(1)**, pp. 55–63.

[3] Aguinis H, Gottfredson RK & Joo H, 2013, *Best-practice recommendations for defining, identifying, and handling outliers*, Organizational Research Methods, **16(2)**, pp. 270–301.

[4] Analytics Vidhya Content Team, 2016, *A complete tutorial on tree based modeling from scratch (in R & Python)*, [Online], [Cited November 2018], Available from `https://www.analyticsvidhya.com/blog/2016/04/complete-tutorial-tree-based-modeling-scratch-in-python/`.

[5] Aptel I, Salmi LR, Masson F, Bourdé A, Henrion G & Erny P, 1999, *Road accident statistics: Discrepancies between police and hospital data in a French island*, Accident Analysis and Prevention, **31(1)**, pp. 101–108.

[6] Arries CS, Sewlal R & Mamabolo V, 2011, *The importance of exposure data for a comprehensive accident database*, Proceedings of the 30th Southern Africa Transport Conference (SATC 2011), Pretoria, pp. 254–262.

[7] Aven T & Renn O, 2009, *On risk defined as an event where the outcome is uncertain*, Journal of Risk Research, **12(1)**, pp. 1–11.

[8] Bach S, Binder A, Montavon G, Klauschen F, Müller KR & Same W, 2015, *On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation*, PLoS ONE, **10(7)**, pp. e0130140.

[9] Bailey TC & Gatrell AC, 1995, *Interactive spatial data analysis*, Longman Scientific & Technical, Harlow.

[10] Baltzan P & Phillips A, 2012, *Business driven information systems*, 3rd Edition, McGraw-Hill, New York (NY).

[11] Bartlett MS, 1975, *The statistical analysis of spatial pattern*, Chapman & Hall, London.

[12] El-Basyouny K & Sayed T, 2009, *Collision prediction models using multivariate Poisson-lognormal regression*, Accident Analysis and Prevention, **41(4)**, pp. 820–828.

[13] Bebis G & Georgiopoulos M, 1994, *Feed-forward neural networks*, IEEE Potentials, **13(4)**, pp. 27–31.

[14] Bernhardsen T, 2002, *Geographic information systems: An introduction*, 3rd Edition, John Wiley & Sons, Inc., New York (NY).

[15] Berry MJA & Linoff GS, 1997, *Data mining technique: For marketing, sales, and customer support*, Wiley Computer Publishing, New York (NY).

[16] Bijleveld F, 2005, *The covariance between the number of accidents and the number of victims in multivariate analysis of accident related outcomes*, Accident Analysis and Prevention, **37(4)**, pp. 591–600.

[17] Black WR, 1992, *Network autocorrelation in transport network and flow systems*, Geographical Analysis, **24(3)**, pp. 207–222.

[18] Blitz S, 2018, *What is exploratory data analysis?*, [Online], [Cited November 2018], Available from `https://www.sisense.com/blog/exploratory-data-analysis/`.

[19] Bonnefon JF, Shariff A & Rahwan I, 2016, *The social dilemma of autonomous vehicles*, Science, **352(6293)**, pp. 1573–1576.

[20] Bonneson JA & Pratt MP, 2008, *Procedure for developing accident modification factors from cross-sectional data*, Transportation Research Record, **2083(1)**, pp. 40–48.

[21] Boots BN & Getis A, 1988, *Point pattern analysis*, SAGE Publications, Newbury Park (CA).

[22] Boser BE, Guyon IM & Vapnik VN, 1992, *A training algorithm for optimal margin classifiers*, Proceedings of the Fifth Annual Workshop on Computational Learning Theory, Pittsburgh (PA), pp. 144–152.

[23] Bramer M, 2007, *Principles of data mining*, 1st Edition, Springer, London.

[24] Bratland D, 2018, *US motor vehicles: Deaths per VMT, deaths per capita, total deaths, VTM, and population*, [Online], [Cited November 2018], Available from `https://www-fars.nhtsa.dot.gov/Main/index.aspx`.

[25] Breiman L, 2001, *Random forests*, Machine Learning, **45(1)**, pp. 5–32.

[26] BrightPlanet, 2012, *Structured vs unstructured data*, [Online], [Cited November 2012], Available from `https://brightplanet.com/2012/06/structured-vs-unstructured-data/`.

[27] British Standards Institution, 1999, *OHSAS 18001: Occupational health and safety management system — Specification*, (Unpublished) Technical Report MD-19-02 Rev 4, London.

[28] Brown MS, 2014, *Data mining for dummies*, John Wiley & Sons, Inc., Hoboken (NJ).

[29] Brown MS, 2015, *CRISP-DM: The dominant process for data mining*, Proceedings of the PyData Conference, London.

[30] Brown MS, 2015, *What IT needs to know about the data mining process*, [Online], [Cited September 2018], Available from `https://www.forbes.com/sites/metabrown/2015/07/29/what-it-needs-to-know-about-the-data-mining-process/#3db7f2c4515f`.

[31] Brownlee J, 2017, *What is the difference between test and validation datasets?*, [Online], [Cited November 2018], Available from `https://machinelearningmastery.com/difference-test-validation-datasets/`.

[32] Bryan FR, 2014, *The birth of Ford Motor Company*, [Online], [Cited September 2018], Available from `http://hfha.org/the-ford-story/the-birth-of-ford-motor-company/`.

[33] Buckey DJ, 2013, *Vector and raster — Advantages and disadvantages*, [Online], [Cited February 2018], Available from `http://planet.botany.uwc.ac.za/nisl/gis/gis_primer/page_19.htm`.

[34] Bühlmann P & Geer SVD, 2011, *Statistics for high-dimensional data*, Springer, Berlin.

[35] Burrough PA & McDonnell RA, 1998, *Principles of geographical information systems*, Oxford University Press, New York (NY).

[36] Cameron AC & Trivedi PK, 2013, *Regression analysis of count data*, Cambridge University Press, Cambridge.

[37] Carson J & Mannering F, 2001, *The effect of ice warning signs on ice-accident frequencies and severities*, Accident Analysis and Prevention, **33(1)**, pp. 99–109.

[38] Çaydaş U & Ekici S, 2012, *Support vector machines models for surface roughness prediction in CNC turning of AISI 304 austenitic stainless steel*, Journal of Intelligent Manufacturing, **23(3)**, pp. 639–650.

[39] Central Intelligence Agency, 2016, *The world factbook: Country comparison of roadways*, [Online], [Cited November 2018], Available from `https://www.cia.gov/library/publications/the-world-factbook/rankorder/2085rank.html`.

[40] Chakrabarti S, Ester M, Fayyad U, Gehrke J, Han J, Morishita S, Piatetsky-Shapiro G & Wang W, 2006, *Data mining curriculum: A proposal (version 1.0)*, [Online], [Cited September 2018], Available from `http://www.kdd.org/exploration_files/CURMay06.pdf`.

[41] Chandola V, Boriah S & Kumar V, 2007, *Similarity measures for categorical data — A comparative study*, (Unpublished) Technical Report TR 07-022, Department of Computer Science and Engineering, University of Minnesota, Minneapolis (MN).

[42] Chang LY, 2005, *Analysis of freeway accident frequencies: Negative binomial regression versus artificial neural network*, Safety Science, **43(8)**, pp. 541–557.

[43] Chapman P, Clinton J, Kerber R, Khabaza T, Reinartz T, Shearer C & Wirth R, 2000, *CRISP-DM 1.0: Step-by-step data mining guide*, [Online], [Cited August 2018], Available from `https://www.the-modeling-agency.com/crisp-dm.pdf`.

[44] Chicco D, 2017, *Ten quick tips for machine learning in computational biology*, BioData Mining, **10(35)**, pp. 1–17.

[45] Christoffel T & Gallagher SS, 1999, *Injury prevention and public health*, Aspen Publishers, Gaithersburg (MD).

[46] Clark PJ & Evans FC, 1954, *Distance to nearest neighbor as a measure of spatial relationship in populations*, Ecology, **35(4)**, pp. 445–453.

[47] Clifton C, 2017, *Data mining*, [Online], [Cited September 2018], Available from `https://www.britannica.com/technology/data-mining`.

[48] Compare.com, 2016, *How car tracking programs are changing auto insurance*, [Online], [Cited February 2019], Available from `https://www.compare.com/auto-insurance/guides/how-insurance-telematics-work`.

[49] Conway RW & Maxwell WL, 1962, *A queuing model with state dependent service rates*, Journal of Industrial Engineering, **12(2)**, pp. 132–136.

[50] Cooke K, 2018, *How many people are left-handed?*, [Online], [Cited October 2018], Available from `https://uk.kantar.com/consumer/leisure/2018/how-many-people-are-left-handed/`.

[51] Copeland M, 2016, *What's the difference between artificial intelligence, machine learning, and deep learning?*, [Online], [Cited September 2018], Available from `https://blogs.nvidia.com/blog/2016/07/29/whats-difference-artificial-intelligence-machine-learning-deep-learning-ai/`.

[52] Copins M, 2010, *Successful people who had early setbacks*, [Online], [Cited September 2018], Available from `https://www.newsweek.com/photo/2010/10/27/successful-people-who-rebounded-from-early-failure.html`.

[53] Corporation IBM, 2017, *Descriptive, predictive, prescriptive: Transforming asset and facilities management with analytics*, (Unpublished) Technical Report, Somers (NY).

[54] Cortes C & Vapnik V, 1995, *Support vector networks*, Machine Learning, **20(3)**, pp. 273–297.

[55] Couclelis H, 1992, *People manipulate objects (but cultivate fields): Beyond the raster-vector debate in GIS*, pp. 65–77 in Frank AU, Campari I & Formentini U (Eds), *Theories and methods of spatio-temporal reasoning in geographic space*, Springer, Berlin.

[56] Cova TJ & Goodchild MF, 2002, *Extending geographical representation to include fields of spatial objects*, International Journal of Geographical Information Science, **16(6)**, pp. 509–532.

[57] Cressie N, 1991, *Statistics for spatial data*, John Wiley & Sons, Inc., New York (NY).

[58] Curry LA, Nembhard IM & Bradley EH, 2009, *Qualitative and mixed methods provide unique contributions to outcomes research*, Circulation, **119**, pp. 1442–1452.

[59] Cutler A, Cutler DR & Stevens JR, 2011, *Random forests*, pp. 157–176 in Zhang C & Ma YQ (Eds), *Ensemble Machine Learning*, Springer, New York (NY).

[60] Dacey MF, 1975, *Evaluation of the Poisson approximation to measures of the random pattern in the square*, Geographical Analysis, **12**, pp. 351–367.

[61] Dar P, 2018, *An autonomous car learned how to drive itself in 20 minutes using reinforcement learning*, [Online], [Cited November 2018], Available from `https://www.analyticsvidhya.com/blog/2018/07/autonomous-car-learnt-drive-itself-20-minutes-using-reinforcement-learning/`.

[62] Dataflair Team, 2018, *Kernel functions: Introduction to SVM kernel and examples*, [Online], [Cited November 2018], Available from `https://data-flair.training/blogs/svm-kernel-functions/`.

[63] Date CJ, 2003, *Introduction to database systems*, 8th Edition, Addison Wesley, Reading (MA).

[64] Datta A, Sen S & Zick Y, 2016, *Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems*, 2016 IEEE Symposium on Security and Privacy (SP), pp. 598–617.

[65] Davis MF, 2010, *Encyclopedia of research design*, SAGE Publications, Inc, Thousand Oaks (CA).

[66] Daws M, 2017, *Algorithm to project onto line segments*, [Online], [Cited November 2018], Available from `https://cs.stackexchange.com/questions/79314/algorithm-to-project-onto-line-segments`.

[67] Deeb AE, 2015, *The unreasonable effectiveness of random forests*, [Online], [Cited November 2018], Available from `https://medium.com/rants-on-machine-learning/the-unreasonable-effectiveness-of-random-forests-f33c3ce28883`.

[68] DeepAI, 2017, *Semi-supervised learning*, [Online], [Cited November 2018], Available from `https://deepai.org/machine-learning-glossary-and-terms/semi-supervised-learning`.

[69] DeMers MN, 2000, *Fundamentals of geographic information systems*, 2nd Edition, John Wiley & Sons, Ltd, New York (NY).

[70] Demšar J, Curk T, Erjavec A, Gorup Č, Hočevar T, Milutinovič M, Možina M, Polajnar M, Toplak M, Starič A, Štajdohar M, Umek L, Žagar L, Žbontar J, Žitnik M & Zupan B, 2013, *Orange: Data mining toolbox in python*, Journal of Machine Learning Research, **14**, pp. 2349–2353.

[71] Diggle PJ, 1983, *Statistical analysis of spatial point patterns*, Academic Press, London.

[72] Diggle PJ, 2003, *Statistical analysis of spatial point patterns*, 2nd Edition, Arnold, London.

[73] Doeden M, 2007, *Crazy cars*, Lerner Publications, Minneapolis (MN).

[74] Donnell ET & Mason JM, 2006, *Predicting the frequency of median barrier crashes on Pennsylvania interstate highways*, Accident Analysis and Prevention, **38(3)**, pp. 590–599.

[75] Esposito F, Malerba D & Semeraro G, 1997, *A comparative analysis of methods for pruning decision trees*, IEEE Transactions on Pattern Analysis and Machine Intelligence, **19(5)**, pp. 476–491.

[76] Fawcett T, 2006, *An introduction to ROC analysis*, Pattern Recognition Letters, **27(8)**, pp. 861–874.

[77] Fisher RA, 1932, *Statistical methods for research workers*, 4th Edition, Oliver & Boyd, Edinburgh.

[78] Fix E & Hodges JL, 1951, *Discriminatory analysis, nonparametric discrimination: Consistency properties*, (Unpublished) Technical Report Report No. 4, Project No. 21-49-004, USAF School of Aviation Medicine, Randolf Field (TX).

[79] Flahaut B, Mouchart M & Martin ES, 2003, *The local spatial autocorrelation and the kernel method for identifying black zones — A comparative approach*, Accident Analysis and Prevention, **35**, pp. 991–1004.

[80] Folkmann MN, 2011, *Encoding symbolism: Immateriality and possibility in design*, Design and Culture, **3(1)**, pp. 51–74.

[81] Ford Motor Company, 2018, *The Henry Ford story*, [Online], [Cited September 2018], Available from `https://www.ford.co.uk/experience-ford/history-and-heritage`.

[82] Freund Y & Schapire RE, 1999, *A short introduction to boosting*, Journal of the Japanese Society for Artificial Intelligence, **14(5)**, pp. 771–780.

[83] Fridstrøm L, Ifver J, Ingebrigtsen S, Kulmala R & Thomsen LK, 1995, *Measuring the contribution of randomness, exposure, weather, and daylight to the variation in road accident counts*, Accident Analysis and Prevention, **27(1)**, pp. 1–20.

[84] Galathiya AS, Ganatra AP & Bhensdadia CK, 2012, *Classification with an improved decision tree algorithm*, International Journal of Computer Applications, **46(23)**, pp. 1–6.

[85] Geedipally SR & Lord D, 2010, *Investigating the effect of modeling single-vehicle and multi-vehicle crashes separately on confidence intervals of Poisson-gamma models*, Accident Analysis and Prevention, **42(4)**, pp. 1273–1282.

[86] Gimond M, 2018, *Intro to GIS and spatial analysis*, [Online], [Cited February 2018], Available from `Availible%20at:%20https://mgimond.github.io/Spatial/index.html`.

[87] Goodchild MF, 1992, *Geographical data modeling*, Computers and Geosciences, **18(4)**, pp. 401–408.

[88] Goodfellow I, Bengio Y & Courville A, 2016, *Deep learning*, MIT Press, Cambridge (MA).

[89] Google, 2012, *Maps — Navigate Explore*, [Online], [Cited February 2019], Available from https://play.google.com/store/apps/details?id=com.google.android.apps.maps&hl=en.

[90] Google Maps, 2018, *78 A34 Manchester, England*, [Online], [Cited October 2018], Available from https://goo.gl/maps/zERtGrdpHmB2.

[91] Grace-Martin K, 2008, *Seven ways to make up data: Common methods to imputing missing data*, [Online], [Cited November 2008], Available from https://www.theanalysisfactor.com/seven-ways-to-make-up-data-common-methods-to-imputing-missing-data/.

[92] Griffith DA & Amrhein C, 1983, *An evaluation of correction techniques for boundary effects in spatial statistical analysis: Traditional methods*, Geographical Analysis, **15**, pp. 352–360.

[93] Gruver A, 2017, *Part VI: Objects vs. fields (and vector vs. raster)*, [Online], [Cited February 2018], Available from https://www.e-education.psu.edu/geog486/l1_p8.html.

[94] Guidoum AC, 2015, *Kernel estimator and bandwidth selection for density and its derivatives*, [Online], [Cited November 2018], Available from https://cran.r-project.org/web/packages/kedd/vignettes/kedd.pdf.

[95] Gupta P, 2017, *Decision trees in machine learning*, [Online], [Cited November 2018], Available from https://towardsdatascience.com/decision-trees-in-machine-learning-641b9c4e8052.

[96] Guyon IM, 2016, *Data mining history: The invention of support vector machines*, [Online], [Cited November 2018], Available from https://www.kdnuggets.com/2016/07/guyon-data-mining-history-svm-support-vector-machines.html.

[97] Haddon Jr W, 1980, *Advances in the epidemiology of injuries as a basis for public policy*, Public Health Reports, **95(5)**, pp. 411.

[98] Haimes YY, 2009, *On the complex definition of risk: A systems-based approach*, Risk Analysis: An International Journal, **29(12)**, pp. 1647–1654.

[99] Haining RP, 2003, *Spatial data analysis theory and practice*, Cambridge University Press, Cambridge.

[100] Han J & Kamber M, 2001, *Data mining: Concepts and techniques*, Morgan Kaufmann Publishers, Burlingtom (MA).

[101] Hand DJ, Mannila H & Smyth P, 2001, *Principles of data mining*, 1st Edition, MIT Press, Cambridge (MA).

[102] Hand DJ & Till RJ, 2001, *A simple generalization of the area under the ROC curve to multiple class classification problems*, Machine Learning, **45**, pp. 171–186.

[103] Harris E, 2002, *Information gain versus gain ratio: A study of split method biases*, Proceedings of the International Symposium on Artificial Intelligence and Mathematics, Fort Lauderdale (FL), pp. 2–4.

[104] Hastie T, Tibshirani R & Friedman J, 2009, *The elements of statistical learning*, 2nd Edition, Springer, New York (NY).

[105] Hauer E & Hakkert A, 1988, *Extent and some implications of incomplete accident reporting*, Transportation Research Record, **1185**, pp. 1–10.

[106] Hellerstein JM, 2008, *Quantitative data cleaning for large databases*, EECS Computer Science Division, University of California, Berkeley (CA).

[107] Hill T & Lewicki P, 2007, *Statistics — Methods and applications*, 1st Edition, StatSoft, Tusla (OK).

[108] Hiselius LW, 2004, *Estimating the relationship between accident frequency and homogeneous and inhomogeneous traffic flows*, Accident Analysis and Prevention, **36(6)**, pp. 985–992.

[109] History.com Staff, 2009, *Henry Ford*, [Online], [Cited September 2018], Available from `https://www.history.com/topics/henry-ford`.

[110] Houston-Galveston Area Council, 2013, *Land use & land cover data*, [Online], [Cited November 2018], Available from `http://www.h-gac.com/community/socioeconomic/land-use-data/default.aspx`.

[111] Huang J & Ling CX, 2005, *Using AUC and accuracy in evaluating learning algorithms*, IEEE Transactions on Knowledge Data Engineering, **17**, pp. 299–310.

[112] Illian J, Penttinen A, Stoyan H & Stoyan D, 2008, *Statistical analysis and modeling of spatial point patterns*, John Wiley & Sons, Inc., Chichester.

[113] Ingram DR, 1978, *An evaluation of procedures utilized in nearest-neighbor analysis*, Geografiska Annaler, **60**, pp. 65–70.

[114] James G, Witten D, Hastie T & Tibshirani R, 2013, *An introduction to statistical learning*, 6th Edition, Springer, New York (NY).

[115] James HF, 1991, *Under-reporting of road traffic accidents*, Traffic Engineering & Control, **32(12)**, pp. 574–583.

[116] Jensen K, 2012, *Process diagram showing the relationship between the different phases of CRISP-DM*, [Online], [Cited August 2018], Available from `ftp://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/18.0/en/ModelerCRISPDM.pdf`.

[117] Jiang MF, Tseng SS & Liao SY, 1999, *Data types generalization for data mining algorithm*, Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics, Tokyo, pp. 928–933.

[118] Jiang N & Okabe A, 2014, *Different ways of thinking about street networks and spatial analysis*, Geographic Analysis, **46(4)**, pp. 341–344.

[119] JP Research India, 2014, *Mumbai — Pune Expressway Road Accident Study*, (Unpublished) Technical Report, Maharashtra.

[120] Kaelbling LP, Littman ML & Moore AW, 1996, *Reinforcement learning: A survey*, Journal of Artificial Intelligence Research, **4**, pp. 237–285.

[121] Kearn M, 2016, *Machine learning is for muggles too!*, [Online], [Cited November 2018], Available from `https://blogs.msdn.microsoft.com/martinkearn/2016/03/01/machine-learning-is-for-muggles-too/`.

[122] Kecman V, 2005, *Support vector machines: Theory and applications*, pp. 1–48 in Wang L (Ed), Springer, New York (NY).

[123] Kendall KE & Kendall JE, 2014, *Systems analysis and design*, 9th Edition, Pearson, Upper Saddle River (NJ).

[124] Kendall MG, 1943, *Advanced theory of statistics — Volume I*, Charles Griffin, London.

[125] Kim E, 2013, *Everything you wanted to know about the kernel trick*, [Online], [Cited November 2018], Available from `http://www.eric-kim.net/eric-kim-net/posts/1/kernel_trick.html`.

[126] Kim H & Loh WY, 2001, *Classification trees with unbiased multiway splits*, Journal of the American Statistical Association, **96**, pp. 598–604.

[127] Kingma DP & Ba J, 2014, *Adam: A method for stochastic optimization*, arXiv preprint arXiv:1412.6980.

[128] Kreipke RC, 2007, *The Model T: A pictorial chronology of the most famous car in the world*, M.T. Publishing Company, Inc., Evansville (IN).

[129] Kulkarni M, 2017, *Decision trees for classification: A machine learning algorithm*, [Online], [Cited November 2018], Available from `https://www.xoriant.com/blog/product-engineering/decision-trees-machine-learning-algorithm.html`.

[130] Kumara S & Chin HC, 2005, *Application of Poisson underreporting model to examine crash frequencies at signalized three-legged intersections*, Transportation Research Record, **1908(1)**, pp. 46–50.

[131] Kumara S & Chin HC, 2003, *Modeling accident occurrence at signalized tee intersections with special emphasis on excess zeros*, Traffic Injury Prevention, **4(1)**, pp. 53–57.

[132] Lambert D, 1992, *Zero-inflated Poisson regression, with an application to defects in manufacturing*, Technometrics, **34(1)**, pp. 1–14.

[133] Laurini R & Thompson D, 1992, *Fundamentals of spatial information systems*, Ademic Press, London.

[134] Lee DT, 1982, *On k-nearest neighbor Voronoi diagrams in the plane*, IEEE Transactions on Computers, **31(6)**, pp. 478–487.

[135] Lee J & Mannering F, 2002, *Impact of roadside features on the frequency and severity of run-off-roadway accidents: An empirical analysis*, Accident Analysis and Prevention, **34(2)**, pp. 149–161.

[136] Lewis C, 2017, *AI & machine learning black boxes: The need for transparency and accountability*, [Online], [Cited November 2018], Available from `https://www.kdnuggets.com/2017/04/ai-machine-learning-black-boxes-transparency-accountability.html`.

[137] Lewis RJ, 2000, *An introduction to classification and regression tree (CART) analysis*, Proceedings of the 310th Annual Meeting of the Society for Academic Emergency Medicine, San Francisco (CA), p. 14.

[138] Li X, Lord D, Zhang Y & Xie Y, 2008, *Predicting motor vehicle crashes using support vector machine models*, Accident Analysis and Prevention, **40(4)**, pp. 1611–1618.

[139] Liang F, 2005, *Bayesian neural networks for nonlinear time series forecasting*, Statistics and Computing, **15(1)**, pp. 13–29.

[140] Lipovetsky S & Conklin M, 2001, *Analysis of regression in game theory approach*, Applied Stochastic Models in Business and Industry, **17(4)**, pp. 319–330.

[141] Litman T, 2015, *Autonomous vehicle implementation predictions: Implications for transport planning*, (Unpublished) Technical Report, Victoria Transport Policy Institute, Victoria.

[142] Longley PA, Goodchild MF, Maguire DJ & Rhind DW, 2005, *Geographic information systems and science*, 2nd Edition, John Wiley & Sons, Inc., Chichester.

[143] Lord D & Bonneson J, 2005, *Calibration of predicive models for estimating the safety of ramp design configurations*, Transportation Research Record, **1908**, pp. 88–95.

[144] Lord D & Mannering F, 2010, *The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives*, Transportation Research, Part A: Policy and Practice, **44(5)**, pp. 291–305.

[145] Lord D, Washington S & Ivan JN, 2005, *Poisson, Poisson-gamma and zero-inflated regression models of motor vehicle crashes: Balancing statistical fit and theory*, Accident Analysis and Prevention, **37(1)**, pp. 35–46.

[146] Lord D, 2006, *Modeling motor vehicle crashes using Poisson-gamma models: Examining the effects of low sample mean values and small sample size on the estimation of the fixed dispersion parameter*, Accident Analysis and Prevention, **38(4)**, pp. 751–766.

[147] Lord D & Miranda-Moreno LF, 2008, *Effects of low sample mean values and small sample size on the estimation of the fixed dispersion parameter of Poisson-gamma models for modeling motor vehicle crashes: A Bayesian perspective*, Safety Science, **46(5)**, pp. 751–770.

[148] Lord D & Persaud BN, 2000, *Accident prediction models with and without trend: Application of the generalized estimating equations procedure*, Transportation Research Record, **1717(1)**, pp. 102–108.

[149] Lord D, Washington S & Ivan JN, 2007, *Further notes on the application of zero-inflated models in highway safety*, Accident Analysis and Prevention, **39(1)**, pp. 53–57.

[150] Lundberg SM & Lee SI, 2017, *A unified approach to interpreting model predictions*, Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach (CA), pp. 4765–4774.

[151] Ma J, 2009, *Bayesian analysis of underreporting Poisson regression model with an application to traffic crashes on two-lane highways*, Paper #09-3192. Presented at the 88th Annual Meeting of the Transportation Research Board, Washington (DC).

[152] Ma J & Kockelman KM, 2006, *Bayesian multivariate Poisson regression for models of injury count, by severity*, Transportation Research Record, **1950(1)**, pp. 24–34.

[153] Ma J, Kockelman KM & Damien P, 2008, *A multivariate Poisson-lognormal regression model for prediction of crash counts by severity, using Bayesian methods*, Accident Analysis and Prevention, **40**, pp. 964–975.

[154] Maher MJ & Mountain LJ, 1988, *The identification of accident blackspots: A comparison of current methods*, Accident Analysis and Prevention, **20(2)**, pp. 143–151.

[155] Maher MJ & Summersgill I, 1996, *A comprehensive methodology for the fitting of predictive accident models*, Accident Analysis and Prevention, **28(3)**, pp. 281–296.

[156] Manning CD, Raghavan P & Schütze H, 2009, *An introduction to information retrieval*, Cambridge University Press, Cambridge.

[157] Marr B, 2016, *The top 10 AI and machine learning use cases everyone should know about*, [Online], [Cited November 2018], Available from `https://www.forbes.com/sites/bernardmarr/2016/09/30/what-are-the-top-10-use-cases-for-machine-learning-and-ai/#27a2e2b694c9`.

[158] Marsland S, 2009, *Machine learning: An algorithmic perspective*, 1st Edition, Chapman & Hall, London.

[159] Matern B, 1960, *Spatial variation stochastic models and their application to some problems in forest surveys and other sampling investigations*, Meddelanden Fran Statens Skogsforskningstitut, **49(5)**, pp. 1–144.

[160] MathWorks, 2016, *Support vector machines for binary classification*, [Online], [Cited November 2018], Available from `http://www.mathworks.com/help/stats/support-vector-machines-for-binary-classification.html?requestedDomain=www.mathworks.com`.

[161] Maycock G & Hall R, 1984, *Accidents at 4-arm roundabouts*, (Unpublished) Technical Report TRRL Report 1120, Transportation and Road Research Laboratory, Crowthorne.

[162] McCullagh P & Nelder JA, 1989, *Generalised linear models*, 2nd Edition, Chapman & Hall, London.

[163] McCulloch WS & Pitts W, 1943, *A logical calculus of the ideas immanent in nervous activity*, Bulletin of Mathematical Biophysics, **5(4)**, pp. 115–133.

[164] Meyer MH & Webb PH, 2005, *Modular, layered architecture: The necessary foundation for effective mass customisation in software*, International Journal of Mass Customisation, **1(1)**, pp. 14–36.

[165] Miaou SP, 1994, *The relationship between truck accidents and geometric design of road sections: Poisson versus negative binomial regressions*, Accident Analysis and Prevention, **26(4)**, pp. 471–482.

[166] Miaou SP & Lord D, 2003, *Modeling traffic crash-flow relationships for intersections: Dispersion parameter, functional form, and Bayes versus empirical Bayes*, Transportation Research Record, **1840**, pp. 31–40.

[167] Miaou SP & Lum H, 1993, *Modeling vehicle accidents and highway geometric design relationships*, Accident Analysis and Prevention, **25(6)**, pp. 689–709.

[168] Miaou SP & Song JJ, 2005, *Bayesian ranking of sites for engineering safety improvements: Decision parameter, treatability concept, statistical criterion, and spatial dependence*, Accident Analysis and Prevention, **37(4)**, pp. 699–720.

[169] Mikkonen V & Peltola H, 1997, *Road safety principles and models: Review of descriptive, predictive, risk and accident consequence models*, (Unpublished) Technical Report IRRD No. 892483, Organization for Economic Cooperation and Development, Paris.

[170] Miladinovic B, 2008, *Kernel density estimation of reliability with applications to extreme value distribution*, PhD Dissertation, University of South Florida, Tampa (FL).

[171] Milton J & Mannering F, 1998, *The relationship among highway geometrics, traffic-related elements and motor-vehicle accident frequencies*, Transportation, **25(4)**, pp. 395–413.

[172] Mingers J, 1989, *An empirical comparison of selection measures for decision-tree induction*, Machine Learning, **3(4)**, pp. 319–342.

[173] Mohri M, Rostamizadeh A & Talwalkar A, 2012, *Foundations of machine learning*, MIT Press, Cambridge (MA).

[174] Moisen GG, 2008, *Classification and regression trees*, Encyclopedia of Ecology, **1**, pp. 582–588.

[175] Møller J & Waagepetersen RP, 2002, *Statistical inference and simulation for spatial point processes*, Chapman & Hall, London.

[176] Molnar C, 2018, *Interpretable machine learning: A guide for making black box models explainable*, [Online], [Cited November 2018], Available from `https://christophm.github.io/interpretable-ml-book/`.

[177] Moore D, McCabe G & Bruce C, 2009, *Introduction to the practice of statistics*, 6th Edition, W.H. Freeman and Company, New York (NY).

[178] Morgan R & Epstein J, 2002, *Road safety audit*, (Unpublished) Technical Report D3124, Austroads, Sydney.

[179] Mountain L, Maher M & Fawaz B, 1998, *The influence of trend on estimates of accidents at junctions*, Accident Analysis and Prevention, **30(5)**, pp. 641–649.

[180] Nair V & Hinton GE, 2010, *Rectified linear units improve restricted boltzmann machines*, Proceedings of the 27th International Conference on Machine Learning (ICML-10), Haifa, pp. 807–814.

[181] National Academies of Science, Engineering and Medicine, 2016, *Commercial motor vehicle driver fatigue, long-term health, and highway safety: Research needs*, National Academies Press, Washington (DC).

[182] National Aeronautics and Space Administration, 2017, *How far away is the moon?*, [Online], [Cited November 2018], Available from `https://spaceplace.nasa.gov/moon-distance/en/`.

[183] Nelder JA & Wedderburn RWM, 1972, *Generalized linear models*, Journal of the Royal Statistical Society, Series A (General), **135(3)**, pp. 370–384.

[184] Ng A, 2011, *Machine learning and AI via brain simulations*, [Online], [Cited February 2019], Available from `http://ai.stanford.edu/ang/slides/DeepLearning-Mar2013.pptx`.

[185] Ng A, 2012, *CS229 — Machine Learning*, Stanford University, [Lecture notes], Available from `https://see.stanford.edu/Course/CS229`.

[186] Nilsson G, 2004, *Traffic safety dimensions and the power model to describe the effect of speed on safety*, PhD Dissertation, Lund University, Lund.

[187] Nisbet R, Elder J & Miner G, 2009, *Handbook of statistical analysis and data mining applications*, 1st Edition, Elsevier, London.

[188] Oh J, Washington S & Choi K, 2004, *Development of accident prediction models for rural highway intersections*, Transportation Research Record, **1897**, pp. 18–27.

[189] Oh J, Washington S & Nam D, 2006, *Accident prediction model for railway highway interfaces*, Accident Analysis and Prevention, **38(2)**, pp. 346–356.

[190] Okabe A, Boots B, Sugihara K & Chiu SN, 2000, *Spatial tessellations: Concepts and applications of Voronoi diagrams*, 2nd Edition, John Wiley & Sons, Inc., Chichester.

[191] Okabe A, Okunuki K & SANET Team, 2015, *A spatial analysis along networks (Standalone Ver.1.0)*, [Online], [Cited August 2017], Available from `http://sanet.csis.u-tokyo.ac.jp/sub_en/standalone.html`.

[192] Okabe A & Satoh T, 2005, *Uniform network transformation for points pattern analysis on a nonuniform network*, Journal of Geographical Systems, **8**, pp. 25–27.

[193] Okabe A, Satoh T & Sugihara K, 2009, *A kernal density estimation method for networks, its computational method and a GIS-based tool*, International Journal of Geographic Information Science, **23**, pp. 7–23.

[194] Okabe A & Sugihara K, 2012, *Spatial analysis along networks: Statistical and computational methods*, John Wiley & Sons, Inc., Hoboken (NJ).

[195] Okabe A, Yamada H & Kitamura M, 1995, *Statistical analysis of the distribution of points on a network*, Geographical Analysis, **27(2)**, pp. 152–175.

[196] Okabe A & Yamanda I, 2001, *The K-function method on a network and its computational implementation*, Geographical Analysis, **33(3)**, pp. 271–290.

[197] O'Sullivan D & Unwin DJ, 2002, *Geographic information analysis*, John Wiley & Sons, Inc., Hoboken (NJ).

[198] O'Sullivan D & Wong DWS, 2007, *A surface-based approach to measuring spatial segregation*, Geographic Analysis, **39(2)**, pp. 147–168.

[199] Özöğür-Akyüz S, Üstünkar G & Weber GW, 2016, *Adapted infinite kernel learning by multi-local algorithm*, International Journal of Pattern Recognition and Artificial Intelligence, **30(4)**, pp. 1651004-2–1651004-21.

[200] Özöğür-Akyüz S & Weber GW, 2008, *Learning with infinitely many kernels via semi-infinite programming*, Proceedings of the Continuous Optimization and Knowledge Based Technologies, 20th EURO Mini Conference, Neringa, pp. 342–348.

[201] Park ES & Lord D, 2007, *Multivariate Poisson-lognormal models for jointly modeling crash frequency by severity*, Transportation Research Record, **2019(1)**, pp. 1–6.

[202] Peden M, Scurfield R, Sleet D, Mohan D, Hyder AA, Jarawan E & Mathers C, 2004, *World report on road traffic injury prevention*, (Unpublished) Technical Report, World Health Organisation, Geneva.

[203] Persaud BN, Lord D & Palminaso J, 2002, *Issues of calibration and transferability in developing accident prediction models for urban intersections*, Transportation Research Record, **1784**, pp. 57–64.

[204] Piegorsch WW, 1990, *Maximum likelihood estimation for the negative binomial dispersion parameter*, Biometrics, **46(3)**, pp. 863–867.

[205] Plisson F, 2017, *Finding the right k in k-nearest neighbors*, [Online], [Cited November 2018], Available from `https://www.fabienplisson.com/finding-the-right-k-in-k-nearest-neighbors/`.

[206] Plunkett Research, 2010, *Automobile industry introduction*, [Online], [Cited September 2018], Available from `http://www.plunkettresearch.com:80/automobiles%5C%20trucks%5C%20market%5C%20research/industry%5C%20overview`.

[207] Press G, 2016, *Cleaning big data: Most time-consuming, least enjoyable data science task, survey says*, [Online], [Cited November 2018], Available from `https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/#346e81816f63`.

[208] Pyle D, 1999, *Data preparation for data mining*, Morgan Kaufmann Publishers, Los Altos (CA).

[209] Pyle D & José CS, 2015, *An executive's guide to machine learning*, [Online], [Cited September 2018], Available from `https://www.mckinsey.com/industries/high-tech/our-insights/an-executives-guide-to-machine-learning`.

[210] Quinlan JR, 1993, *C4.5: Programs for machine learning*, Morgan Kaufmann Publishers, San Francisco (CA).

[211] Quinlan JR, 1986, *Induction of decision trees*, Machine Learning, **1**, pp. 81–106.

[212] Quinlan JR, 1987, *Simplifying decision trees*, International Journal of Man-Machine Studies, **27(8)**, pp. 221–234.

[213] Rashidi S, Ranjitkar P & Hadas Y, 2014, *Modelling bus dwell time with decision tree-based methods*, Journal of the Transportation Research Board, **2418(1)**, pp. 74–83.

[214] Rennie JD, Shih L, Teevan J & Karger DR, 2003, *Tackling the poor assumptions of naive Bayes text classifiers*, Proceedings of the 20th International Conference on Machine Learning (ICML-03), Washington (DC), pp. 616–623.

[215] Ribeiro MT, Singh S & Guestrin C, 2016, *Why should I trust you?: Explaining the predictions of any classifier*, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco (CA), pp. 1135–1144.

[216] Rigaux P, Scholl M & Voisard A, 2002, *Spatial databases with applications to GIS*, Morgan Kaufmann Publishers, San Francisco (CA).

[217] Ripley BD, 1979, *Tests of 'randomness' for spatial point patterns*, Journal of the Royal Statistical Society, Series B, **41(3)**, pp. 368–374.

[218] Ripley BD, 1981, *Spatial statistics*, John Wiley & Sons, Inc., New York (NY).

[219] Ripley BD, 1988, *Spatial inference for spatial processes*, Cambridge University Press, Cambridge.

[220] RiskSolutions, 2012, *Contributory factors analysis for road traffic collisions: A report for the National Roads Authority of Ireland*, (Unpublished) Technical Report, Warrington.

[221] Riviere C, Lauret P, Ramsamy JM & Page Y, 2006, *A Bayesian neural network approach to estimating the energy equivalent speed*, Accident Analysis and Prevention, **38(2)**, pp. 248–259.

[222] Rokach L & Maimon O, 2008, *Data mining with decision trees: Theory and applications*, World Scientific Publishing, River Edge (NJ).

[223] Rosenblatt F, 1958, *The perceptron: A probabilistic model for information storage and organization in the brain*, Psychological Review, **65(6)**, pp. 386.

[224] Rosset S, 2004, *Model selection via AUC*, Proceedings of the 21st International Conference on Machine Learning (ICML 2004), New York (NY), p. 89.

[225] Rouse M, 2012, *How to solve your TMI problem: Data science analytics to the rescue*, [Online], [Cited November 2018], Available from `http://searchsoftwarequality.techtarget.com/essentialguide/How-to-solve-your-TMI-problem-Data-science-analyticsto-the-rescue`.

[226] Rouse M & Stedman C, 2008, *What is data analytics?*, [Online], [Cited August 2018], Available from `https://searchdatamanagement.techtarget.com/definition/data-analytics`.

[227] Ruder S, 2016, *An overview of gradient descent optimization algorithms*, arXiv preprint arXiv:1609.04747.

[228] Ruggieri S, 2002, *Efficient C4.5 (classification algorithm)*, IEEE Transactions on Knowledge and Data Engineering, **14(2)**, pp. 438–444.

[229] Rumar K, 1999, *Transport safety visions, targets and strategies: Beyond 2000*, (Unpublished) Technical Report, European Transport Safety Council, Brussels.

[230] Rustagi N, Kumar A, Norbu L & Vyas D, 2017, *Applying Haddon matrix for evaluation of road crash victims in Delhi, India*, Indian Journal of Surgery, **80(5)**, pp. 479–487.

[231] Sabey BE & Staughton GC, 1975, *Interacting roles of road environment, vehicle and road user in accidents*, Proceedings of the 5th International Conference of the International Association for Accident and Traffic Medicine, London, pp. 1–17.

[232] Samuel AL, 1959, *Some studies in machine learning using the game of checkers*, IBM Journal of Research and Development, **4(1.2)**, pp. 206–226.

[233] Sayad S, 2010, *Naive Bayesian*, [Online], [Cited November 2018], Available from `https://www.saedsayad.com/naive_bayesian.htm`.

[234] Schabenberger O & Gotway CA, 2005, *Statistical methods for spatial data analysis*, Chapman & Hall, Boca Raton (FL).

[235] Schmandt M, 2013, *Chapter 1: Introduction*, [Online], [Cited February 2018], Available from `http://giscommons.org/introduction-concepts/`.

[236] Scikit Learn, 2013, *Model evaluation: Quantifying the quality of predictions*, [Online], [Cited November 2018], Available from `https://scikit-learn.org/stable/modules/model_evaluation.html`.

[237] Senn S, 2003, *A conversation with John Nelder*, Statistical Science, **18(1)**, pp. 118–131.

[238] Shah T, 2017, *About train, validation and test sets in machine learning*, [Online], [Cited November 2018], Available from `https://towardsdatascience.com/train-validation-and-test-sets-72cb40cba9e7`.

[239] Shankar V, Mannering F & Barfield W, 1995, *Effect of roadway geometrics and environmental factors on rural freeway accident frequencies*, Accident Analysis and Prevention, **27(3)**, pp. 371–389.

[240] Shankar VN, Albin RB, Milton JC & Mannering FL, 1998, *Evaluating median crossover likelihoods with clustered accident counts: An empirical inquiry using the random effects negative binomial model*, Transportation Research Record, **1635(1)**, pp. 44–48.

[241] Shankar VN, Ulfarsson GF, Pendyala RM & Nebergall MB, 2003, *Modeling crashes involving pedestrians and motorized traffic*, Safety Science, **41(7)**, pp. 627–640.

[242] Shankar V, Milton J & Mannering F, 1997, *Modeling accident frequencies as zero-altered probability processes: An empirical inquiry*, Accident Analysis and Prevention, **29(6)**, pp. 829–837.

[243] Shannon CE & Weaver W, 1949, *The mathematical theory of communication*, 1st Edition, University of Illinois Press, Urbana (IL).

[244] Shapley LS, 1953, *A value for n-person games*, Contributions to the Theory of Games, **2(28)**, pp. 307–317.

[245] Shearer C, 2000, *The CRISP-DM model: The new blueprint for data mining*, Journal of Data Warehousing, **5(4)**, pp. 13–22.

[246] Shrikumar A, Greenside P, Shcherbina A & Kundaje A, 2016, *Not just a black box: Learning important features through propagating activation differences*, arXiv preprint arXiv:1605.01713.

[247] Silverman BW, 1986, *Density estimation for statistics and data analysis*, Chapman & Hall, London.

[248] Sittikariya S, Shankar V & Venkataraman N, 2009, *Modeling heterogeneity: Traffic accidents*, VDM Publishing, Riga.

[249] Song JJ, Ghosh M, Miaou S & Mallick B, 2006, *Bayesian multivariate spatial models for roadway traffic crash mapping*, Journal of Multivariate Analysis, **97(1)**, pp. 246–273.

[250] Sönnichsen N, 2019, *Average cost of casualties and accidents on the road in Great Britain in 2017 (in GBP)*, [Online], [Cited September 2019], Available from `https://www.statista.com/statistics/322862/average-cost-of-road-accidents-and-%5C%20casualties-in-great-britain-uk/`.

[251] Statista, 2018, *Number of cars sold worldwide from 1990 to 2018 (in million units)*, [Online], [Cited October 2018], Available from `https://www.statista.com/statistics/200002/international-car-sales-since-1990/`.

[252] Statista, 2018, *Number of passenger cars and commercial vehicles in use worldwide from 2006 to 2015 in (1,000 units)*, [Online], [Cited September 2018], Available from `https://www.statista.com/statistics/281134/number-of-vehicles-in-use-worldwide/`.

[253] Statista, 2018, *Revenue of the leading automotive manufacturers worldwide in FY 2017 (in billion U.S. dollars)*, [Online], [Cited October 2018], Available from `https://www.statista.com/statistics/232958/revenue-of-the-leading-car-manufacturers-worldwide/`.

[254] StatSoft Inc., 2016, *Popular decision tree: Classification and regression trees (CART)*, [Online], [Cited November 2018], Available from `http://www.statsoft.com/Textbook/Classification-and-Regression-Trees`.

[255] Steenberghen T, Dufays T, Thomas I & Flahaut B, 2004, *Intra-urban location and clustering of road accidents using GIS: A Belgian example*, International Journal of Geographical Information Science, **18(2)**, pp. 169–181.

[256] Steynberg R, 2016, *A framework for identifying the most likely successful underprivileged tertiary bursary applicants*, MEng Thesis, Stellenbosch University, Stellenbosch.

[257] Stoyan D, Kendall WS & Mecke J, 1995, *Stochastic geometry and its applications*, 2nd Edition, John Wiley & Sons, Inc., Chichester.

[258] Strang G, 2009, *Introduction to linear algebra*, 4th Edition, Cambridge Press, Wellesley (MA).

[259] Štrumbelj E & Kononenko I, 2014, *Explaining prediction models and individual predictions with feature contributions*, Knowledge and Information Systems, **41(3)**, pp. 647–665.

[260] Swingler K, 1996, *Applying neural networks: A practical guide*, Academic Press, San Diego (CA).

[261] Tape TG, 2001, *The area under an ROC curve*, [Online], [Cited November 2018], Available from `http://gim.unmc.edu/dxtests/roc3.htm`.

[262] Techechelons, 2018, *10 Buzzwords of machine learning in 2018*, [Online], [Cited October 2018], Available from `https://www.techechelons.com/blog/10-buzzwords-of-machine-learning-in-2018`.

[263] Techopedia, 2011, *Data mining*, [Online], [Cited September 2018], Available from `https://www.techopedia.com/definition/1181/data-mining`.

[264] Thakur D, Markandaiah N & Raj DS, 2010, *Re optimization of ID3 and C4.5 decision tree*, Proceedings of the 2010 International Conference on Computer and Communication Technology (ICCCT), Allahabad, pp. 448–450.

[265] The Henry Ford, 2018, *Henry Ford: Changing the production model*, [Online], [Cited October 2018], Available from https://www.thehenryford.org/collections-and-research/.

[266] The Road Safety Foundation, 2014, *Risk rating of Britain's motorways and A roads (risk bands 2020)*, [Online], [Cited October 2018], Available from https://www.eurorap.org/partner-countries/great-britain/.

[267] Theodoridis S & Koutroumbas K, 2008, *Pattern recognition*, 4th Edition, Elsevier, Amsterdam.

[268] Thulin H, 1987, *Trafikolyckor och trafikskadade enligt polis, sjukvård och försäkringsbolag*, (Unpublished) Technical Report, Statens väg- och Trafikinstitut, Linköping.

[269] Tieleman T & Hinton GE, 2012, *Lecture 6.5 — RmsProp: Divide the gradient by a running average of its recent magnitude*, COURSERA: Neural Networks for Machine Learning, URL: https://class.coursera.org/neuralnets-2012-001/lecture.

[270] Timofeev R, 2004, *Classification and regression trees (CART) — Theory and applications*, PhD Dissertation, Humboldt University, Berlin.

[271] Tonidandel S, Lebreton JM & Johnson JW, 2009, *Determining the statistical significance of relative weights*, Psychological Methods, **14(4)**, pp. 387–399.

[272] Toroyan T, 2015, *Global status report on road safety 2015*, (Unpublished) Technical Report ISBN 9789241565066, World Health Organisation, Geneva.

[273] Treat JR, 1980, *A study of precrash factors involved in traffic accidents*, HSRI Research Review, **10(6)**, pp. 1–35.

[274] UK Department for Transport, 2015, *Facts on road fatalities*, [Online], [Cited November 2018], Available from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/448037/road-fatalities-2013-data.pdf.

[275] UK Department for Transport, 2016, *GB road traffic counts*, [Online], [Cited April 2017], Available from https://data.gov.uk/dataset/208c0e7b-353f-4e2d-8b7a-1a7118467acc/gb-road-traffic-counts.

[276] UK Department for Transport, 2018, *Reported road casualties in Great Britain: Notes, definitions, symbols and conventions — 2017*, [Online], [Cited October 2019], Available from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/743853/reported-road-casualties-gb-notes-definitions.pdf.

[277] UK Department for Transport, 2016, *Road safety data*, [Online], [Cited April 2017], Available from https://data.gov.uk/dataset/cb7ae6f0-4be6-4935-9277-47e5ce24a11f/road-safety-data.

[278] UK Department for Transport, 2005, *STATS19 report form*, [Online], [Cited October 2018], Available from http://docs.adrn.ac.uk/888043/mrdoc/pdf/888043_stats19-road-accident-injury-statistics-report-form.pdf.

[279] UK Department for Transport, 2011, *STATS20: Instructions for the completion of road accident reports from non-CRASH sources*, (Unpublished) Technical Report, London.

[280] Ulfarsson GF & Shankar VN, 2003, *Accident count model based on multiyear cross-sectional roadway data with serial correlation*, Transportation Research Record, **1840(1)**, pp. 193–197.

[281] Vasudev R, 2017, *What is one hot encoding? Why and when do you have to use it?*, [Online], [Cited November 2018], Available from `https://hackernoon.com/what-is-one-hot-encoding-why-and-when-do-you-have-to-use-it-e3c6186d008f`.

[282] Vella M, 2016, *Why you shouldn't be allowed to drive*, [Online], [Cited November 2018], Available from `http://time.com/4236980/against-human-driving/`.

[283] Voronoi G, 1907, *Nouvelles applications des paramètres continus à la théorie des formes quadratiques*, Journal für die Reine und Angewandte Mathematik, **133**, pp. 97–178.

[284] Washington State Department of Transportation, 2009, *Roundabout benefits*, [Online], [Cited October 2018], Available from `http://www.wsdot.wa.gov/Safety/roundabouts/benefits.htm`.

[285] Washington SP, Karlaftis MG & Mannering F, 2010, *Statistical and econometric methods for transportation data analysis*, 2nd Edition, Chapman & Hall, London.

[286] Welling M, 2010, *Support vector machines*, (Unpublished) Technical Report, University of Toronto, Toronto.

[287] Wilks SS, 1962, *Mathematical statistics*, John Wiley & Sons, Inc., New York (NY).

[288] Williams G, Baxter R, He H, Hawkins S & Gu L, 2002, *A comparative study of RNN for outlier detection in data mining*, Proceedings of the IEEE International Conference on Data-mining, Maebashi City, pp. 709–712.

[289] Winston WL, 2004, *Operations research: Applications and algorithms*, 4th Edition, Curt Hinrichs, Belmont (CA).

[290] Wirth R & Hipp J, 2000, *CRISP-DM: Towards a standard process model for data mining*, Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining, Manchester, pp. 29–39.

[291] Wolpert DH & Macready WG, 2005, *Coevolutionary free lunches*, IEEE Transactions on Evolutionary Computation, **9(6)**, pp. 721–735.

[292] Wood G, 2002, *Generalised linear accident models and goodness of fit testing*, Accident Analysis and Prevention, **34(4)**, pp. 417–427.

[293] Worboys MF & Duckham M, 2004, *GIS: A computing perspective*, 2nd Edition, CRC Press, Boca Raton (FL).

[294] World Health Organisation, 2009, *Risk factors for road traffic injuries*, (Unpublished) Technical Report, World Health Organisation, Geneva.

[295] World Health Organisation, 2018, *Global health estimates 2016: Deaths by cause, age, sex, by country and by region*, (Unpublished) Technical Report, World Health Organisation, Geneva.

[296] World Health Organisation, 2018, *Global Health Observatory (GHO) data: Road safety*, [Online], [Cited October 2018], Available from `http://www.who.int/gho/road_safety/en/`.

[297] Xie Y, Lord D & Zhang Y, 2007, *Predicting motor vehicle collisions using Bayesian neural network models: An empirical analysis*, Accident Analysis and Prevention, **39(5)**, pp. 922–933.

[298] Xu Z & Sui DZ, 2007, *Small-worked characteristics on transportation networks: A perspective from network autocorrelation*, Journal of Geographical Systems, **9**, pp. 189–205.

[299] Yamada I, 2009, *Edge effects*, International Encyclopedia of Human Geography, pp. 381–388.

[300] Yannis G, Papadimitriou E & Evgenikos P, 2011, *Effectiveness of road safety measures at junctions*, Proceedings of the 1st International Conference on Access Management, Athens, pp. 143–155.

[301] Zhu X, 2010, *Advanced natural language processing: Support vector machines*, (Unpublished) Technical Report, University of Wisconsin-Madison, Madison (WI).

[302] Zoph B & Le QV, 2016, *Neural architecture search with reinforcement learning*, [Online], [Cited September 2019], Available from `https://arxiv.org/abs/1611.01578`.

APPENDIX A

GitHub repository contents

This appendix contains a brief description of the contents that have been made available on a online GitHub code repository. The repository contains seventeen Python scripts corresponding to Sub-modules 2.1, 2.2, 2.3, 2.4, 3.1, 4.1, 5.0, 7.1, 7.2, 7.3, 8.1, 8.3, 8.5, 9.3, 9.4 and 9.5, as described in Chapter 6, the data used in the case study investigation of Part III, and an electronic copy of the dissertation, where Chapter 6 serves as a form of documentation for the various sub-modules. The concept demonstrator was created in the Orange Data Mining Software (version 3.19) and may be built according to the workflow illustrated in Figure 7.1. The link to the GitHub repository is as follows: `http://github.com/shanevh/RARE_framework.git`.

APPENDIX B

Accident, vehicle and victim categories

The various categories of the accident, vehicle and victim data sets used in the case study of Part III are summarised in Tables B.1, B.2 and B.3, respectively. The last column contains the proportion of records with a specific categorical attribute value. Attribute categories marked with an asterisk denote the road segment/junction attributes which served as the reference attribute during the demonstration of Sub-module 9.0 in §8.1.5 and §8.2.5.

Table B.1: *The categories and corresponding data proportions of the various accident attributes.*

Accidents			
No.	**Attribute**	**Categories**	**%**
1	AccidentIndex	—	—
2	Longitude	—	—
3	Latitude	—	—
4	AccidentSeverity	Fatal	1.56%
		Serious	15.21%
		Slight	83.23%
5	NumberOfVehicles	—	—
6	Date	—	—
7	DayOfWeek	Sunday	11.35%
		Monday	14.29%
		Tuesday*	14.53%
		Wednesday	15.18%
		Thursday	15.05%
		Friday	16.51%
		Saturday	13.09%
8	Time	—	—
9	FirstRoadClass	Motorway	11.99%
		AM*	0.91%
		A	87.10%
10	RoadType	Roundabout	5.75%
		OneWayStreet*	1.66%
		DualCarriageway	31.17%
		SingleCarriageway	59.69%

Continued on next page

Table B.1 — *Continued from previous page*

No.	Attribute	Categories	%
10	RoadType	SlipRoad	1.73%
11	SpeedLimit	20mph*	0.49%
		30mph	72.79%
		40mph	10.60%
		50mph	2.86%
		60mph	1.26%
		70mph	11.97%
12	JunctionDetail	NotAtJunctionOrWithin20Metres*	27.33%
		Roundabout	6.69%
		MiniRoundabout	0.70%
		TOrStaggeredJunction	34.83%
		SlipRoad	1.45%
		Crossroads	21.02%
		MoreThan4Arms	2.15%
		PrivateDriveOrEntrance	2.20%
		OtherJunction	3.64%
13	JunctionControl	NotAtJunctionOrWithin20Metres*	27.33%
		AuthorisedPerson	0.28%
		AutoTrafficSignal	27.57%
		StopSign	0.57%
		GiveWayOrUncontrolled	44.24%
14	PedestrianCrossing-HumanControl	NoneWithin50Metres*	99.09%
		ControlBySchoolCrossingPatrol	0.60%
		ControlByOtherAuthorisedPerson	0.32%
15	PedestrianCrossing-PhysicalFacilities	NoPhysicalCrossingFacilitiesWithin50Metres*	72.52%
		Zebra	0.78%
		PelicanPuffinToucanOrSimilarNonJunctionPedestrianLightCrossing	6.95%
		PedestrianPhaseAtTrafficSignalJunction	16.25%
		FootbridgeOrSubway	0.49%
		CentralRefuge	3.00%
16	LightConditions	Daylight*	69.37%
		DarknessLightsLit	27.72%
		DarknessLightsUnlit	0.34%
		DarknessNoLighting	1.90%
		DarknessLightingUnknown	0.67%
17	WeatherConditions	FineNoHighWinds*	78.27%
		RainingNoHighWinds	15.71%
		SnowingNoHighWinds	0.59%
		FineHighWinds	1.18%
		RainingHighWinds	2.10%
		SnowingHighWinds	0.15%
		FogOrMist	0.31%
		Other	1.69%

Continued on next page

Table B.1 — *Continued from previous page*

No.	Attribute	Categories	%
18	RoadSurfaceConditions	Dry*	65.17%
		WetOrDamp	33.13%
		Snow	0.40%
		FrostOrIce	1.19%
		FloodOver3cmDeep	0.11%
19	SpecialConditionsAtSite	None*	97.36%
		AutoTrafficSignalOut	0.24%
		AutoSignalPartDefective	0.07%
		RoadSignOrMarkingDefectiveOrObscured	0.11%
		Roadworks	1.76%
		RoadSurfaceDefective	0.09%
		OilOrDiesel	0.34%
		Mud	0.03%
20	CarriagewayHazards	None*	98.92%
		VehicleLoadOnRoad	0.08%
		OtherObjectOnRoad	0.55%
		PreviousAccident	0.15%
		DogOnRoad	0.00%
		PedestrianInCarriagewayNotInjured	0.16%
		AnyAnimalInCarriageway	0.14%
21	UrbanOrRuralArea	Urban	84.20%
		Rural*	15.80%

Table B.2: *The categories of the various vehicle attributes.*

Vehicles			
No.	**Attribute**	**Categories**	**%**
1	AccidentIndex	—	—
2	VehicleReference	—	—
3	VehicleType	PedalCycle	5.35%
		Motorcycle50CCAndUnder	0.64%
		Motorcycle125CCAndUnder	1.69%
		Motorcycle125CCAndUpTo500CC	0.70%
		MotorcycleOver500CC	2.10%
		TaxiPrivateHireCar	2.67%
		Car*	75.28%
		Minibus8-16PassengerSeats	0.44%
		BusOrCoach17OrMorePassengerSeats	2.54%
		RiddenHorse	0.01%
		AgriculturalVehicle	0.05%
		Tram	0.06%
		VanGoods3.5TonnesMGWOrUnder	3.99%
		GoodsOver3.5TonnesAndUnder7.5Tonnes	0.76%

Continued on next page

Table B.2 – *Continued from previous page*

No.	Attribute	Categories	%
3	VehicleType	Goods7.5TonnesMGWAndOver	2.59%
		MobilityScooter	0.00%
		ElectricMotorcycle	0.00%
		OtherVehicle	1.13%
		MotorcycleUnknownCC	0.00%
		GoodsVehicleUnknownWeight	0.00%
4	TowingAndArticulation	NoTowArticulation*	98.05%
		ArticulatedVehicle	1.49%
		DoubleOrMultipleTrailer	0.04%
		Caravan	0.05%
		SingleTrailer	0.26%
		OtherTow	0.11%
5	VehicleManoeuvre	Reversing	0.56%
		Parked	2.46%
		WaitingToGoHeldUp	8.00%
		SlowingOrStopping	10.19%
		MovingOff	3.17%
		UTurn	0.84%
		TurningLeft	3.03%
		WaitingToTurnLeft	0.46%
		TurningRight	11.49%
		WaitingToTurnRight	1.71%
		ChangingLaneToLeft	1.38%
		ChangingLaneToRight	1.36%
		OvertakingMovingVehicleOffside	1.15%
		OvertakingStaticVehicleOffside	0.93%
		OvertakingNearside	0.71%
		GoingAheadLeftHandBend	1.73%
		GoingAheadRightHandBend	1.76%
		GoingAheadOther*	53.94%
6	VehicleLocation	OnMainCarriagewayNotInRestrictedLane*	97.51%
		TramLightRailTrack	0.18%
		BusLane	0.50%
		BuswayIncludingGuidedBusway	0.08%
		CycleLaneOnMainCarriageway	0.61%
		CyclewayOrSharedUseFootway NotPartOfMainCarriageway	0.04%
		OnLayByOrHardShoulder	0.26%
		EnteringLayByOrHardShoulder	0.05%
		LeavingLayByOrHardShoulder	0.12%
		FootwayPavement	0.66%
		NotOnCarriageway	0.00%
7	JunctionLocation	NotAtOrWithinTwentyMetresOfJunction*	28.56%
		ApproachingJunctionOrWaiting-ParkedAtJunctionApproach	29.17%

Continued on next page

Table B.2 – *Continued from previous page*

No.	Attribute	Categories	%
7	JunctionLocation	ClearedJunctionOrWaiting-ParkedAtJunctionExit	6.17%
		LeavingRoundabout	1.01%
		EnteringRoundabout	2.39%
		LeavingMainRoad	2.29%
		EnteringMainRoad	4.20%
		EnteringFromSlipRoad	0.50%
		MidJunctionOnRoundaboutOrOnMainRoad	25.71%
8	SkiddingAndOverturning	None*	89.79%
		Skidded	7.95%
		SkiddedAndOverturned	1.13%
		Jackknifed	0.03%
		JackknifedAndOverturned	0.01%
		Overturned	1.09%
9	HitObjectInCarriageway	None*	97.61%
		PreviousAccident	0.03%
		RoadWorks	0.04%
		ParkedVehicle	0.70%
		BridgeRoof	0.00%
		BridgeSide	0.04%
		BollardOrRefuge	0.74%
		OpenDoorOfVehicle	0.01%
		CentralIslandOfRoundabout	0.00%
		Kerb	0.80%
		OtherObject	0.01%
		AnyAnimalExceptRiddenHorse	0.01%
10	VehicleLeavingCarriageway	DidNotLeaveCarriageway*	93.06%
		Nearside	3.44%
		NearsideAndRebounded	0.70%
		StraightAheadAtJunction	0.21%
		OffsideOnToCentralReservation	0.55%
		OffsideOnToCentralReservationAndRebounded	0.39%
		OffsideCrossedCentralReservation	0.11%
		Offside	1.28%
		OffsideAndRebounded	0.26%
11	HitObjectOffCarriageway	None*	93.73%
		RoadSignOrTrafficSignal	0.97%
		LampPost	0.67%
		TelegraphOrElectricityPole	0.08%
		Tree	0.35%
		BusStopOrBusShelter	0.09%
		CentralCrashBarrier	1.06%
		NearOffsideCrashBarrier	1.27%
		SubmergedInWater	0.00%
		EnteredDitch	0.10%
		OtherPermanentObject	1.67%

Continued on next page

Table B.2 – *Continued from previous page*

No.	Attribute	Categories	%
11	HitObjectOffCarriageway	WallOrFence	0.00%
12	FirstPointOfImpact	DidNotImpact*	6.67%
		Front	47.42%
		Back	19.98%
		Offside	12.92%
		Nearside	13.00%
13	WasVehicleLeftHandDrive	No	99.65%
		Yes	0.35%
14	JourneyPurposeOfDriver	JourneyAsPartOfWork	60.46%
		CommutingToFromWork	36.89%
		TakingPupilToFromSchool	2.02%
		PupilRidingToFromSchool	0.62%
		Other	0.00%
15	SexOfDriver	Male	70.63%
		Female	29.37%
16	AgeOfDriver	—	—
17	EngineCapacityCC	—	—
18	PropulsionCode	Petrol	65.51%
		HeavyOil	34.05%
		Electric	0.02%
		Steam	0.00%
		Gas	0.03%
		PetrolGasLPG	0.05%
		GasBiFuel	0.10%
		HybridElectric	0.26%
		NewFuelTechnology	0.00%
		FuelCells	0.00%
		ElectricDiesel	0.01%
19	AgeOfVehicle	—	—
20	DriverIMDDecile	MostDeprived10%	21.46%
		MoreDeprived10-20%	14.81%
		MoreDeprived20-30%	11.13%
		MoreDeprived30-40%	10.38%
		MoreDeprived40-50%	7.94%
		LessDeprived40-50%	7.90%
		LessDeprived30-40%	7.69%
		LessDeprived20-30%	7.18%
		LessDeprived10-20%	6.83%
		LeastDeprived10%	4.69%
21	DriverHomeAreaType	UrbanArea	94.05%
		SmallTown	2.93%
		Rural	13.02%

Table B.3: *The categories of the various victim attributes.*

Victims			
No.	**Attribute**	**Categories**	**%**
1	AccidentIndex	—	—
2	VehicleReference	—	—
3	VictimReference	—	—
4	VictimSeverity	Fatal	0.89%
		Serious	8.83%
		Slight	90.28%
4	VictimClass	DriverOrRider	58.71%
		Passenger	28.85%
		Pedestrian	12.44%

APPENDIX C

Road accident data collection form

The STATS19 report form used for the collection of RA data in Great Britain may be found in Figures C.1–C.4. Detailed instructions for the appropriate completion of this form are stipulated in the STATS20 manual [279].

Sept. 2004

MG NSRF/A

ACCIDENT STATISTICS

Incident URN

Other ref.

1.3 ACCIDENT REFERENCE

*FATAL / SERIOUS / SLIGHT

1.9 TIME H H M M DAY* Su M T W Th F S 1.7 DATE D D M M 2 0 Y Y

1st Road Class & No. or (Unclassified - UC) (Not Known - NK) 1st Road Name

Outside House No. or Name or Marker Post No. at junction with / or metres N S E W * of

2nd Road Class & No. or (Unclassified - UC) (Not Known - NK) 2nd Road Name

Town Sector /Beat No.

County or Borough

Parish No. or Name 1.10 Local Auth No. (if known)

1.11 Grid Reference E → N ↑

REPORTING OFFICER Name Number

BCU/Stn 1.2 Force Tel Number

1.5	Number of vehicles		
1.6	Number of casualties		

1.14 ROAD TYPE		✗
Roundabout	1	
One way street	2	
Dual carriageway	3	
Single carriageway	6	
Slip road	7	
Unknown	9	

1.15	Speed Limit (Permanent)		

1.16 JUNCTION DETAIL		✗
Not at or within 20 metres of junction	00	
Roundabout	01	
Mini roundabout	02	
T or staggered junction	03	
Slip road	05	
Crossroads	06	
Multiple junction	07	
Using private drive or entrance	08	
Other junction	09	

JUNCTION ACCIDENTS ONLY

1.17 JUNCTION CONTROL		✗
Authorised person	1	
Automatic traffic signal	2	
Stop sign	3	
Give way or uncontrolled	4	

1.20a PEDESTRIAN CROSSING - HUMAN CONTROL		✗
None within 50 metres	0	
Control by school crossing patrol	1	
Control by other authorised person	2	

1.20b PEDESTRIAN CROSSING - PHYSICAL FACILITIES		✗
No physical crossing facility within 50m	0	
Zebra crossing	1	
Pelican, puffin, toucan or similar non-junction pedestrian light crossing	4	
Pedestrian phase at traffic signal junction	5	
Footbridge or subway	7	
Central refuge — no other controls	8	

1.22 WEATHER		✗
Fine without high winds	1	
Raining without high winds	2	
Snowing without high winds	3	
Fine with high winds	4	
Raining with high winds	5	
Snowing with high winds	6	
Fog or mist — if hazard	7	
Other	8	
Unknown	9	

1.23 ROAD SURFACE CONDITION		✗
Dry	1	
Wet / Damp	2	
Snow	3	
Frost / Ice	4	
Flood (surface water over 3cm deep)	5	

1.21 LIGHT CONDITIONS		✗
Daylight: street lights present	1	
Daylight: no street lighting	2	
Daylight: street lighting unknown	3	
Darkness: street lights present and lit	4	
Darkness: street lights present but unlit	5	
Darkness: no street lighting	6	
Darkness: street lighting unknown	7	

1.24 SPECIAL CONDITIONS AT SITE		✗
None	0	
Auto traffic signal out	1	
Auto traffic signal partially defective	2	
Permanent road signing or marking defective or obscured	3	
Roadworks	4	
Road surface defective	5	
Oil or diesel	6	
Mud	7	

1.25 CARRIAGEWAY HAZARDS		✗
None	0	
Dislodged vehicle load in carriageway	1	
Other object in carriageway	2	
Involvement with previous accident	3	
Pedestrian in carriageway - not injured	6	
Any animal in carriageway (except ridden horse)	7	

1.26 Did a police officer attend the scene and obtain the details for this report?		✗
Yes	1	
No	2	

Subject to local directions, boxes with a grey background need not be completed if already recorded

*** Circle as appropriate**

UNCLASSIFIED

FIGURE C.1: *Page 1 of the STATS19 accident report form [278].*

MG NSRF/B

VEHICLE RECORD

Sept. 2004

2.26 VEHICLE REGISTRATION MARK	
Vehicle 001	
Vehicle 002	
Vehicle 003	
Vehicle 004	

2.28 FOREIGN REGISTERED VEHICLE ✗		VEHICLE 1	2	3	4
Not foreign registered vehicle	0				
Foreign registered vehicle LHD	1				
Foreign registered vehicle RHD	2				
Foreign reg' vehicle-two wheeler	3				

2.5 TYPE OF VEHICLE ✗					
Pedal cycle	01				
M/cycle 50cc and under	02				
M/cycle over 50cc and up to 125cc	03				
M/cycle over 125cc and up to 500cc	04				
Motorcycle over 500cc	05				
Taxi / Private hire car	08				
Car	09				
Minibus (8-16 passenger seats)	10				
Bus or coach (17 or more passenger seats)	11				
Other motor vehicle	14				
Other non-motor vehicle	15				
Ridden horse	16				
Agricultural vehicle (include diggers etc)	17				
Tram / Light rail	18				
Goods vehicle 3.5 tonnes mgw and under	19				
Goods vehicle over 3.5 tonnes mgw and under 7.5 tonnes mgw	20				
Goods vehicle 7.5 tonnes mgw and over	21				

2.6 TOWING AND ARTICULATION ✗					
No tow or articulation	0				
Articulated vehicle	1				
Double or multiple trailer	2				
Caravan	3				
Single trailer	4				
Other tow	5				

2.21 SEX OF DRIVER ✗					
Male	1				
Female	2				
Driver not traced	3				

2.22 AGE OF DRIVER (Estimate if necessary)

Vehicle 001		Vehicle 002	
Vehicle 003		Vehicle 004	

2.27 DRIVER HOME POSTCODE
or Code: 1- Unknown 2- Non UK Resident 3 - Parked & unattended

Vehicle 001		
Vehicle 002		
Vehicle 003		
Vehicle 004		

2.23 BREATH TEST ✗		VEHICLE 1	2	3	4
Not applicable	0				
Positive	1				
Negative	2				
Not requested	3				
Refused to provide	4				
Driver not contacted at time of acc'	5				
Not provided (medical reasons)	6				

2.24 HIT AND RUN ✗					
Not hit and run	0				
Hit and run	1				
Non-stop vehicle, not hit	2				

2.29 JOURNEY PURPOSE OF DRIVER/RIDER ✗					
Journey as part of work	1				
Commuting to / from work	2				
Taking school pupil to/from school	3				
Pupil riding to / from school	4				
Other/Not known	5				

2.9 VEHICLE LOCATION AT TIME OF ACCIDENT RESTRICTED LANE/AWAY FROM MAIN C'WAY ✗					
On main carriageway not in restricted lane	00				
Tram / Light rail track	01				
Bus lane	02				
Busway (inc. guided busway)	03				
Cycle lane (on main carriageway)	04				
Cycleway or shared use footway (not part of main carriageway)	05				
On lay-by / hard shoulder	06				
Entering lay-by/ hard shoulder	07				
Leaving lay-by / hard shoulder	08				
Footway (pavement)	09				

2.10 JUNCTION LOCATION OF VEHICLE ✗					
Not at or within 20m of junction	0				
Approaching junction or waiting /parked at junction approach	1				
Cleared junction or waiting/ parked at junction exit	2				
Leaving roundabout	3				
Entering roundabout	4				
Leaving main road	5				
Entering main road	6				
Entering from slip road	7				
Mid junction– on roundabout or on main road	8				

2.7 MANOEUVRES ✗					
Reversing	01				
Parked	02				
Waiting to go ahead but held up	03				
Slowing or stopping	04				
Moving off	05				
U turn	06				
Turning left	07				
Waiting to turn left	08				
Turning right	09				
Waiting to turn right	10				
Changing lane to left	11				
Changing lane to right	12				
O'taking moving veh on its offside	13				
O'taking stationary veh on its offside	14				
Overtaking on nearside	15				
Going ahead left hand bend	16				
Going ahead right hand bend	17				
Going ahead other	18				

2.11 SKIDDING AND OVERTURNING ✗		VEHICLE 1	2	3	4
No skidding, jack-knifing or overturning	0				
Skidded	1				
Skidded and overturned	2				
Jack-knifed	3				
Jack-knifed and overturned	4				
Overturned	5				

2.12 HIT OBJECT IN CARRIAGEWAY ✗					
None	00				
Previous accident	01				
Roadworks	02				
Parked vehicle	04				
Bridge-roof	05				
Bridge-side	06				
Bollard / Refuge	07				
Open door of vehicle	08				
Central island of roundabout	09				
Kerb	10				
Other object	11				
Any animal (except ridden horse)	12				

2.13 VEHICLE LEAVING CARRIAGEWAY ✗					
Did not leave carriageway	0				
Left carriageway nearside	1				
Left carriageway nearside and rebounded	2				
Left carriageway straight ahead at junction	3				
Left carriageway offside onto central reservation	4				
Left carriageway offside onto central reserve and rebounded	5				
Left carriageway offside and crossed central reservation	6				
Left carriageway offside	7				
Left carriageway offside and rebounded	8				

2.14 FIRST OBJECT HIT OFF CARRIAGEWAY ✗					
None	00				
Road sign / Traffic signal	01				
Lamp post	02				
Telegraph pole / Electricity pole	03				
Tree	04				
Bus stop / Bus shelter	05				
Central crash barrier	06				
Nearside or offside crash barrier	07				
Submerged in water (completely)	08				
Entered ditch	09				
Other permanent object	10				

2.16 FIRST POINT OF IMPACT ✗					
Did not impact	0				
Front	1				
Back	2				
Offside	3				
Nearside	4				

2.17 FIRST CONTACT BETWEEN EACH VEHICLE
Example: In a 3 car collision vehicle 1 collides with the rear of vehicle 2 pushing it into vehicle 3.

Example Code:			
Vehicle 001 first collides with vehicle 002	0	0	2
Vehicle 002 first collides with vehicle 001	0	0	1
Vehicle 003 first collides with vehicle 002	0	0	2

Vehicle 001	0	Vehicle 002	0
Vehicle 003	0	Vehicle 004	0

Subject to local directions, boxes with a grey background need not be completed if already recorded

UNCLASSIFIED

FIGURE C.2: *Page 2 of the STATS19 accident report form [278].*

MG NSRF/C Sept. 2004

2.8 DIRECTION OF VEHICLE TRAVEL

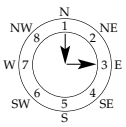
1. Using the Example shown complete the FROM and TO boxes for the vehicles concerned, indicating direction of travel FROM and TO
2. If PARKED enter '00'

Vehicle 001		Vehicle 002	
FROM	TO	FROM	TO

Vehicle 003		Vehicle 004	
FROM	TO	FROM	TO

EXAMPLE

FROM	TO
1	3



CASUALTY RECORD

3.4 VEHICLE REFERENCE NUMBER
Enter VEH No. which CASUALTY occupied (for pedestrians, code vehicle that struck them) e.g. 001,002 etc.

Casualty 001	0	Casualty 002	0
Casualty 003	0	Casualty 004	0
Casualty 005	0	Casualty 006	0

3.18 CASUALTY HOME POSTCODE
or Code: 1- Unknown
2- Non UK Resident

Casualty 001	
Casualty 002	
Casualty 003	
Casualty 004	
Casualty 005	
Casualty 006	

3.7 SEX OF CASUALTY ✗

		CASUALTY 1	2	3	4	5	6
Male	1						
Female	2						

3.8 AGE OF CASUALTY (Estimate if necessary)
For children less than a year enter 00

Casualty 001		Casualty 002	
Casualty 003		Casualty 004	
Casualty 005		Casualty 006	

3.6 CASUALTY CLASS ✗

		1	2	3	4	5	6
Driver/Rider	1						
Veh./pillion Passenger	2						
Pedestrian	3						

3.9 SEVERITY OF CASUALTY ✗

		1	2	3	4	5	6
Fatal	1						
Serious	2						
Slight	3						

3.13 SCHOOL PUPIL CASUALTY ✗

		CASUALTY 1	2	3	4	5	6
School pupil on journey to or from school	1						
Other	0						

3.15 CAR PASSENGER (not driver) ✗

Not a car passenger	0						
Front seat passenger	1						
Rear seat passenger	2						

3.16 BUS OR COACH PASSENGER ✗
(17 passenger seats or more)

Not a bus or coach passenger	0						
Boarding	1						
Alighting	2						
Standing passenger	3						
Seated passenger	4						

PEDESTRIAN CASUALTIES ONLY

3.10 PEDESTRIAN LOCATION ✗

		CASUALTY 1	2	3	4	5	6
In carriageway, crossing on pedestrian crossing facility	01						
In carriageway, crossing within zig-zag lines at crossing approach	02						
In carriageway, crossing within zig-zag lines at crossing exit	03						
In carriageway, crossing elsewhere within 50m of pedestrian crossing	04						
In carriageway, crossing elsewhere	05						
On footway or verge	06						
On refuge, central island or central reservation	07						
In centre of carriageway, not on refuge, island or central reservation	08						
In carriageway, not crossing	09						
Unknown or other	10						

3.11 PEDESTRIAN MOVEMENT ✗

		CASUALTY 1	2	3	4	5	6
Crossing from driver's nearside	1						
Crossing from driver's nearside-masked by parked or stationary veh'	2						
Crossing from driver's offside	3						
Crossing from driver's offside-masked by parked or stationary veh'	4						
In carriageway, stationary - not crossing (standing or playing)	5						
In carriageway, stationary -not crossing (standing or playing), masked by parked or stationary veh'	6						
Walking along in carriageway-facing traffic	7						
Walking along in carriageway-back to traffic	8						
Unknown or other	9						

3.12 PEDESTRIAN DIRECTION ✗

		CASUALTY 1	2	3	4	5	6
Standing still	0						
Northbound	1						
Northeast bound	2						
Eastbound	3						
Southeast bound	4						
Southbound	5						
Southwest bound	6						
Westbound	7						
Northwest bound	8						
Unknown	9						

3.19 PEDESTRIAN INJURED IN THE COURSE OF 'On The Road' WORK
Work actively carried out on public road (e.g. delivery services, road maintenance, postal delivery, traffic control etc.) ✗

No	0						
Yes	1						
Not known	2						

LOCAL STATISTICS

Subject to local directions, boxes with a grey background need not be completed if already recorded

UNCLASSIFIED

FIGURE C.3: *Page 3 of the STATS19 accident report form [278].*

MG NSRF/D **CONTRIBUTORY FACTORS** Sept. 2004

1. Select up to six factors from the grid, relevant to the accident.
2. Factors may be shown in any order, but an indication must be given of whether each factor is **very likely (A)** or **possible (B).**
3. Only include factors that you consider contributed to the accident. (i.e. do NOT include "Poor road surface" unless relevant).
4. More than one factor may, if appropriate, be related to the same road user.
5. The same factor may be related to more than one road user.
6. The participant should be identified by the relevant vehicle or casualty ref no. (e.g. 001, 002 etc.), preceded by "V" if the factor applies to a vehicle, driver/rider or the road environment (e.g. V002), or "C" if the factor relates to a pedestrian or passenger casualty (e.g. C001).
7. Enter U000 if the factor relates to an uninjured pedestrian.

Road Environment Contributed	**101**	**102**	**103**	**104**	**105**	**106**	**107**	**108**	**109**	
	Poor or defective road surface	Deposit on road (e.g. oil, mud, chippings)	Slippery road (due to weather)	Inadequate or masked signs or road markings	Defective traffic signals	Traffic calming (e.g. speed cushions, road humps, chicanes)	Temporary road layout (e.g. contraflow)	Road layout (e.g. bend, hill, narrow carriageway)	Animal or object in carriageway	
Vehicle Defects	**201**	**202**	**203**	**204**	**205**	**206**				
	Tyres illegal, defective or under-inflated	Defective lights or indicators	Defective brakes	Defective steering or suspension	Defective or missing mirrors	Overloaded or poorly loaded vehicle or trailer				
Driver/Rider Only (Includes Pedal Cycles and Horse Riders) – Injudicious Action	**301**	**302**	**303**	**304**	**305**	**306**	**307**	**308**	**309**	**310**
	Disobeyed automatic traffic signal	Disobeyed 'Give Way' or 'Stop' sign or markings	Disobeyed double white lines	Disobeyed pedestrian crossing facility	Illegal turn or direction of travel	Exceeding speed limit	Travelling too fast for conditions	Following too close	Vehicle travelling along pavement	Cyclist entering road from pavement
Driver/Rider Error or Reaction	**401**	**402**	**403**	**404**	**405**	**406**	**407**	**408**	**409**	**410**
	Junction overshoot	Junction restart (moving off at junction)	Poor turn or manoeuvre	Failed to signal or misleading signal	Failed to look properly	Failed to judge other person's path or speed	Passing too close to cyclist, horse rider or pedestrian	Sudden braking	Swerved	Loss of control
Impairment or Distraction	**501**	**502**	**503**	**504**	**505**	**506**	**507**	**508**	**509**	**510**
	Impaired by alcohol	Impaired by drugs (illicit or medicinal)	Fatigue	Uncorrected, defective eyesight	Illness or disability, mental or physical	Not displaying lights at night or in poor visibility	Cyclist wearing dark clothing at night	Driver using mobile phone	Distraction in vehicle	Distraction outside vehicle
Behaviour or Inexperience	**601**	**602**	**603**	**604**	**605**	**606**	**607**			
	Aggressive driving	Careless, reckless or in a hurry	Nervous, uncertain or panic	Driving too slow for conditions or slow vehicle (e.g. tractor)	Learner or inexperienced driver/rider	Inexperience of driving on the left	Unfamiliar with model of vehicle			
Vision Affected by	**701**	**702**	**703**	**704**	**705**	**706**	**707**	**708**	**709**	**710**
	Stationary or parked vehicle(s)	Vegetation	Road layout (e.g. bend, winding road, hill crest)	Buildings, road signs, street furniture	Dazzling headlights	Dazzling sun	Rain, sleet, snow or fog	Spray from other vehicles	Visor or windscreen dirty or scratched	Vehicle blind spot
Pedestrian Only (Casualty or Uninjured)	**801**	**802**	**803**	**804**	**805**	**806**	**807**	**808**	**809**	**810**
	Crossing road masked by stationary or parked vehicle	Failed to look properly	Failed to judge vehicle's path or speed	Wrong use of pedestrian crossing facility	Dangerous action in carriageway (e.g. playing)	Impaired by alcohol	Impaired by drugs (illicit or medicinal)	Careless, reckless or in a hurry	Pedestrian wearing dark clothing at night	Disability or illness, mental or physical
Special Codes	**901**	**902**	**903**	**904**						***999**
	Stolen vehicle	Vehicle in course of crime	Emergency vehicle on a call	Vehicle door opened or closed negligently						Other – Please specify below

	1st	2nd	3rd	4th	5th	6th
Factor in the accident						
Which participant? (e.g. V001, C001, U000)						
Very likely (A) or Possible (B)						

* If 999 Other, give brief details .....

(Note: Only use if another factor contributed to the accident **and include it in the text description of how the accident occurred**)

These factors reflect the reporting officer's opinion at the time of reporting and may not be the result of extensive investigation

UNCLASSIFIED

FIGURE C.4: *Page 4 of the STATS19 accident report form [278].*

APPENDIX D

UK Department for Transport terminology

The terms and definitions shown in Table D.1 have been taken from the 2017 Reported Road Casualties in Great Britain technical report [276], which provide insights into the terminology adopted by the UK Department for Transport in the context of recording the RA data used in the case study of Part III.

Table D.1: *Terminology adopted by the UK Department for Transport.*

Terminology	
Term	**Definition**
Accident	Involves personal injury occurring on the public highway (including footways) in which at least one road vehicle or a vehicle in collision with a pedestrian is involved and which becomes known to the police within 30 days of its occurrence. One accident may give rise to several casualties. "Damage-only" accidents are not included in this publication.
Adults	Persons aged 16 years and over (except where otherwise stated).
Agricultural vehicles	Mainly comprises agricultural tractors (whether or not towing) but also includes mobile excavators and front dumpers.
Animal in carriageway	"Animal in carriageway" includes led animals, but not ridden horses which are recorded separately on the accident statistics report.
Built-up roads	Accidents on "built-up roads" are those which occur on roads with speed limits (ignoring temporary limits) of 40 mph or less. "Non built-up roads" refer to speed limits over 40mph. Motorway accidents are shown separately and are excluded from the totals for built-up and non built-up roads.
Buses and coaches	Buses or coaches equipped to carry 17 or more passengers, regardless of use.
Carriageway hazards	Carriageway hazards are recorded whether or not an animal or object concerned was hit and whether or not its presence is known to have contributed to the accident.
Cars	Includes taxis, estate cars, three and four wheel cars and minibuses except where otherwise stated.

Continued on next page

Table D.1 — *Continued from previous page*

Term	Definition
Casualty	A person killed or injured in an accident. Casualties are subdivided into killed, seriously injured and slightly injured.
Children	Persons under 16 years of age (except where otherwise stated).
Crossroads	"Crossroads" includes only junctions where the alignments of both of the roads are uninterrupted, whatever the angle of the crossing (*i.e.* the arms are not staggered). If there is more than one junction within 20 metres of the accident, the nearest is coded.
Darkness	From half an hour after sunset to half an hour before sunrise.
Daylight	All times other than darkness.
Drivers	Persons in control of vehicles other than pedal cycles, motorcycles and ridden animals (see definition for "riders"). Other occupants of vehicles are passengers.
Fatal accident	An accident in which at least one person is killed.
Fog	"Fog" does not include light mist if it does not constitute a driving hazard on the road where the accident occurred.
Goods vehicles	These are divided into two groups according to vehicle weight. They include tankers, tractor units without their semi-trailers, trailers, articulated vehicles and pick-up trucks.
Heavy goods vehicles	Goods vehicles over 3.5 tonnes maximum permissible gross vehicle weight.
Human control	Crossings with "human control" are those controlled by school crossing (or "lollipop") patrols and other authorised persons (*e.g.* police, traffic wardens).
Injury accident	An accident involving human injury or death.
Junction accident	An accident is considered to be at a junction if it is within 20 metres of an intersection or *roundabout*. Grade separated crossings (by bridge or underpass) are not junctions.
Killed	Human casualties who sustained injuries which caused death less than 30 days after the *accident*. Confirmed suicides are excluded.
KSI	Killed or seriously injured.
Light controlled junction	"Light controlled junction" (with pedestrian phase) is any crossing with traffic lights at a junction, with a "green man phase" or other indicator light for pedestrians, this does not include normal traffic signals with pedestrian stud crossing points but no special indicator lights for pedestrians.
(Light) goods vehicles	Goods vehicles, mainly vans (including car derived vans), not over 3.5 tonnes maximum permissible gross vehicle weight.
Motorcycles	Two-wheel motor vehicles, including mopeds, motor scooters and motorcycle combinations.
Motorways	Motorway and A(M) roads.

Continued on next page

Table D.1 — *Continued from previous page*

Term	Definition
Mobility scooter	A powered wheelchair or scooter with a maximum unladen weight of 150kg and a maximum speed of 8mph.
Other object in carriageway	"Other object in carriageway" comprises those not expected to be found in the carriageway; it does not include permanent features such as a bollard or pedestrian refuge.
Other roads	All B, C and unclassified roads, unless otherwise noted.
Other vehicles	Other vehicles include ambulances, fire engines, trams, refuse vehicles, road rollers, agricultural vehicles, excavators, mobile cranes, mobility scooters and motorised wheelchairs etc, except where otherwise stated. Also included are non motorised vehicles which include those drawn by an animal, ridden horse, wheelchairs without a motor, street barrows *etc.*, except where otherwise stated. In certain tables "other vehicles" may also include buses and coaches and/or goods vehicles.
Passengers	Occupants of vehicles, other than the person in control (the driver or rider). Includes pillion passengers.
Pedal cycles	Includes tandems, tricycles and toy cycles ridden on the carriageway.
Pedal cyclists	Riders of pedal cycles, including any passengers.
Pedestrians	Includes children riding toy cycles on the footway, persons pushing bicycles, pushing or pulling other vehicles or operating pedestrian-controlled vehicles, those leading or herding animals, children in prams or buggies, and people who alight safely from vehicles and are subsequently injured.
Pelican or puffin crossing	A "pelican" or "puffin" crossing has lights controlling the traffic including a flashing amber phase, and lights controlling pedestrians (or pedestrians and cyclist/horse riders) including a flashing "green man" phase. This category also includes any crossing with traffic lights which is not a pelican/puffin/toucan crossing but which has an indicator light for pedestrians only.
Raining	"Raining" includes drizzle, hail and sleet not tending to build up a deposit.
Riders	Persons in control of pedal cycles, motorcycles or ridden animals. Other occupants of these vehicles are passengers.
Road users	Pedestrians and vehicle riders, drivers and passengers.
Roundabout	"Roundabout" includes mini-roundabout junctions.
Rural roads	Major roads and minor roads outside urban areas and having a population of less than ten thousand.
Serious accident	One in which at least one person is seriously injured but no person (other than a confirmed suicide) is killed.

Continued on next page

Table D.1 — *Continued from previous page*

Term	Definition
Serious injury	An injury for which a person is detained in hospital as an in-patient, or any of the following injuries whether or not they are detained in hospital: fractures, concussion, internal injuries, crushings, burns (excluding friction burns), severe cuts, severe general shock requiring medical treatment and injuries causing death 30 or more days after the accident. An injured casualty is recorded as seriously or slightly injured by the police on the basis of information available within a short time of the accident. This generally will not reflect the results of a medical examination, but may be influenced according to whether the casualty is hospitalised or not. Hospitalisation procedures will vary regionally.
Severity	Of an accident; the severity of the most severely injured casualty (either fatal, serious or slight). Of a casualty; killed, seriously injured or slightly injured.
Skidded	"Skidded" does not include vehicles which also jack-knifed. A vehicle which, as a result of the accident, was at any time on its roof, side, front or rear is recorded as having overturned, even though it may have come to rest on its wheels.
Slight accident	One in which at least one person is slightly injured but no person is killed or seriously injured.
Slight injury	An injury of a minor character such as a sprain (including neck whiplash injury), bruise or cut which are not judged to be severe, or slight shock requiring roadside attention. This definition includes injuries not requiring medical treatment.
Snowing	"Snowing" includes sleet building up a deposit.
Speed limits	Permanent speed limits applicable to the roadway.
Taxi	Any vehicle operating as a hackney carriage, regardless of construction, and bearing the appropriate district council or local authority hackney carriage plates. Also includes private hire cars.
T-junction	"T junction" includes slip roads joining dual carriageways.
Users of a vehicle	All occupants (*i.e.* driver or rider and passengers) including persons injured while boarding or alighting from the vehicle.
Urban / rural roads	Urban roads are those within an area of population of 10 000 or more. Tables produced for years prior to 2017 are based on the 2001 Communities and Local Government definition of Urban Settlements. Tables produced for 2017 are based on the 2011 census data that uses a revised 2001 Communities and Local Government classification. Roads outside these areas will be classified as rural.
Van	Vans mainly include vehicles of the van type constructed on a car chassis. These are defined as those vehicles not over 3.5 tonnes maximum permissible gross vehicle weight.

Continued on next page

Table D.1 — *Continued from previous page*

Term	Definition
Vehicles	Vehicles (except taxis) are classified according to their structural type and not according to their employment or category of licence at the time of an accident.
Vehicles involved in accidents	Vehicles whose drivers or passengers are injured, which hit and injure a pedestrian or another vehicle whose driver or passengers are injured, or which contributes to the accident. Vehicles which collide, after the initial accident which caused injury, are not included unless they aggravate the degree of injury or lead to further casualties. Includes pedal cycles ridden on the footway.
Vehicle manoeuvre	In all cases the manoeuvres are those being performed immediately before the accident.
Zebra crossing	A "zebra" crossing has broad black and white stripes on the road and orange flashing beacons.

APPENDIX E

Category spatial scores

The spatial scores associated with the various attribute categories obtained during the semi-spatial attribute testing procedure of Chapter 7 are shown in Table E.1.

Table E.1: *The spatial scores associated with the various attribute categories obtained during the semi-spatial attribute testing procedure.*

Accident attributes			
No.	**Attribute**	**Categories**	$s_c^{(j)}$
7	DayOfWeek	Sunday	0.323
		Monday	0.321
		Tuesday	0.300
		Wednesday	0.344
		Thursday	0.347
		Friday	0.343
		Saturday	0.365
9	FirstRoadClass	Motorway	0.654
		AM	0.905
		A	0.503
10	RoadType	Roundabout	0.902
		OneWayStreet	0.588
		DualCarriageway	0.427
		SingleCarriageway	0.575
		SlipRoad	0.437
11	SpeedLimit	20mph	0.651
		30mph	0.538
		40mph	0.788
		50mph	0.807
		60mph	0.000
		70mph	0.631

Continued on next page

Table E.1 — *Continued from previous page*

No.	Attribute	Categories	$s_c^{(j)}$
12	JunctionDetail	NotAtJunctionOrWithin20Metres	0.185
		Roundabout	0.888
		MiniRoundabout	0.382
		TOrStaggeredJunction	0.597
		SlipRoad	0.489
		Crossroads	0.779
		MoreThan4Arms	0.527
		PrivateDriveOrEntrance	0.258
		OtherJunction	0.283
13	JunctionControl	NotAtJunctionOrWithin20Metres	0.185
		AuthorisedPerson	0.120
		AutoTrafficSignal	0.793
		StopSign	0.084
		GiveWayOrUncontrolled	0.512
14	PedestrianCrossing-HumanControl	NoneWithin50Metres	0.399
		ControlBySchoolCrossingPatrol	0.035
		ControlByOtherAuthorisedPerson	0.121
15	PedestrianCrossing-PhysicalFacilities	NoPhysicalCrossingFacilities-Within50Metres	0.370
		Zebra	0.235
		PelicanPuffinToucanOrSimilar-NonJunctionPedestrianLightCrossing	0.625
		PedestrianPhaseAtTrafficSignalJunction	0.802
		FootbridgeOrSubway	0.664
		CentralRefuge	0.389
16	LightConditions	Daylight	0.415
		DarknessLightsLit	0.390
		DarknessLightsUnlit	0.007
		DarknessNoLighting	0.000
		DarknessLightingUnknown	0.029
17	WeatherConditions	FineNoHighWinds	0.400
		RainingNoHighWinds	0.370
		SnowingNoHighWinds	0.072
		FineHighWinds	0.007
		RainingHighWinds	0.082
		SnowingHighWinds	0.000
		FogOrMist	0.008
		Other	0.149
18	RoadSurfaceConditions	Dry	0.414
		WetOrDamp	0.395
		Snow	0.044
		FrostOrIce	0.019
		FloodOver3cmDeep	0.000

Continued on next page

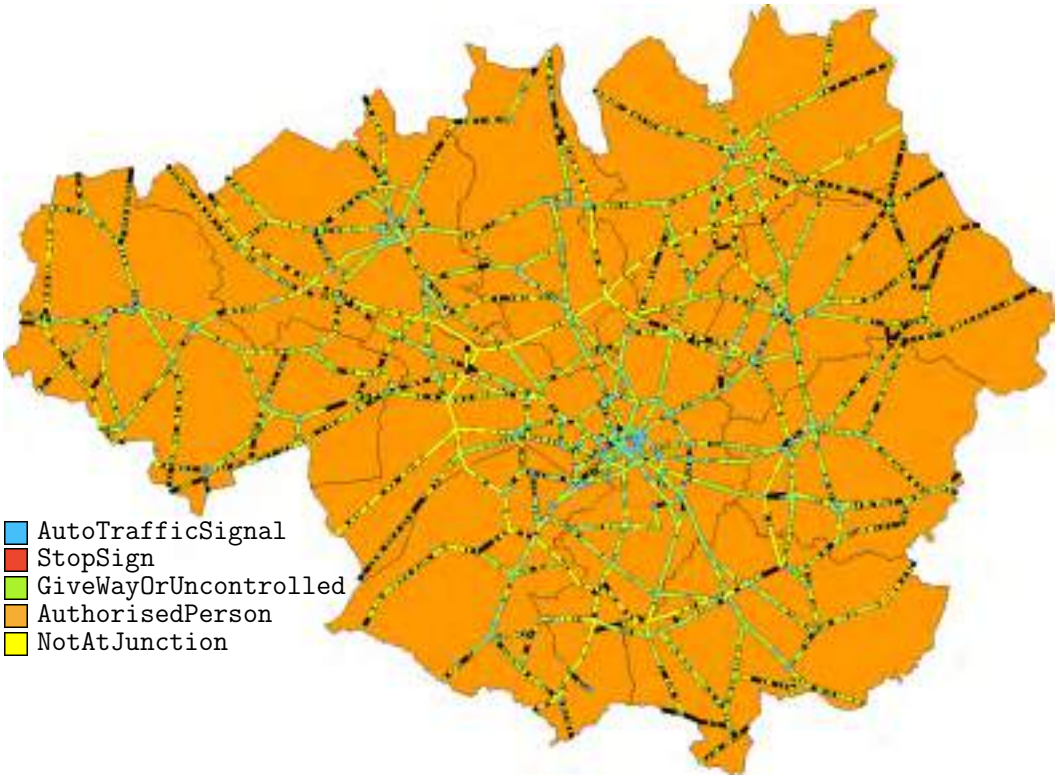
Table E.1 — *Continued from previous page*

No.	Attribute	Categories	$s_c^{(j)}$
19	SpecialConditionsAtSite	None	0.4
		AutoTrafficSignalOut	0.151
		AutoSignalPartDefective	0.022
		RoadSignOrMarking-DefectiveOrObscured	0.121
		Roadworks	0.282
		RoadSurfaceDefective	0.000
		OilOrDiesel	0.000
		Mud	0.000
20	CarriagewayHazards	None	0.399
		VehicleLoadOnRoad	0.000
		OtherObjectOnRoad	0.032
		PreviousAccident	0.000
		DogOnRoad	0.000
		PedestrianInCarriagewayNotInjured	0.095
		AnyAnimalInCarriageway	0.000
21	UrbanOrRuralArea	Urban	0.473
		Rural	0.694

APPENDIX F

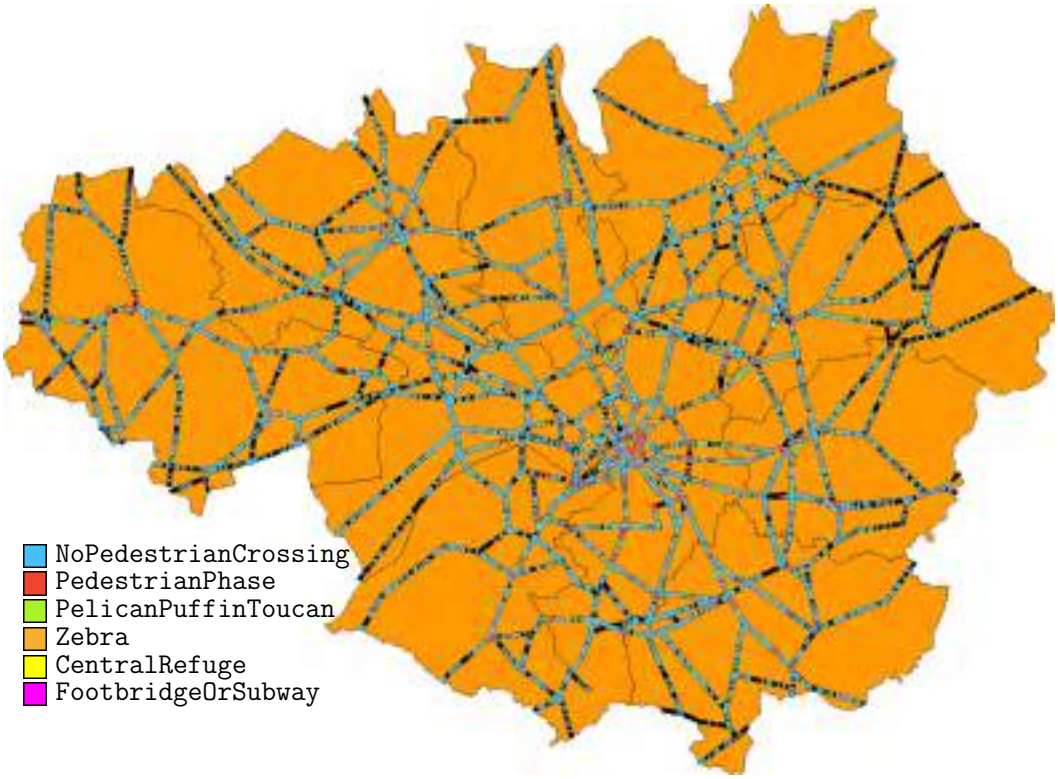
Road segment spatial representation

The spatial results corresponding to the `JunctionControl` and `PhysicalPedestrianCrossing` semi-spatial attributes determined during the demonstration of Sub-module 7.2 in §8.1.2 are shown in Figure F.1.



(a) `JunctionControl`

FIGURE F.1: *A spatial representation of different road segment attribute values in Greater Manchester.*



(b) `PhysicalPedestrianCrossing`

FIGURE F.1: *A spatial representation of different road segment attribute values in Greater Manchester.*

APPENDIX G

Road segment SHAP values

The SHAP values corresponding to the LR, DT, RF, kNN and SVM learning models trained to predict accident rate and severity along a road segment during the demonstration of Submodule 9.5 in §8.1.5 are contained in this appendix. The learning models trained to predict accident rate are shown in Figures G.1–G.5, while those that were trained to predict accident severity are shown in Figures G.6–G.10, respectively.

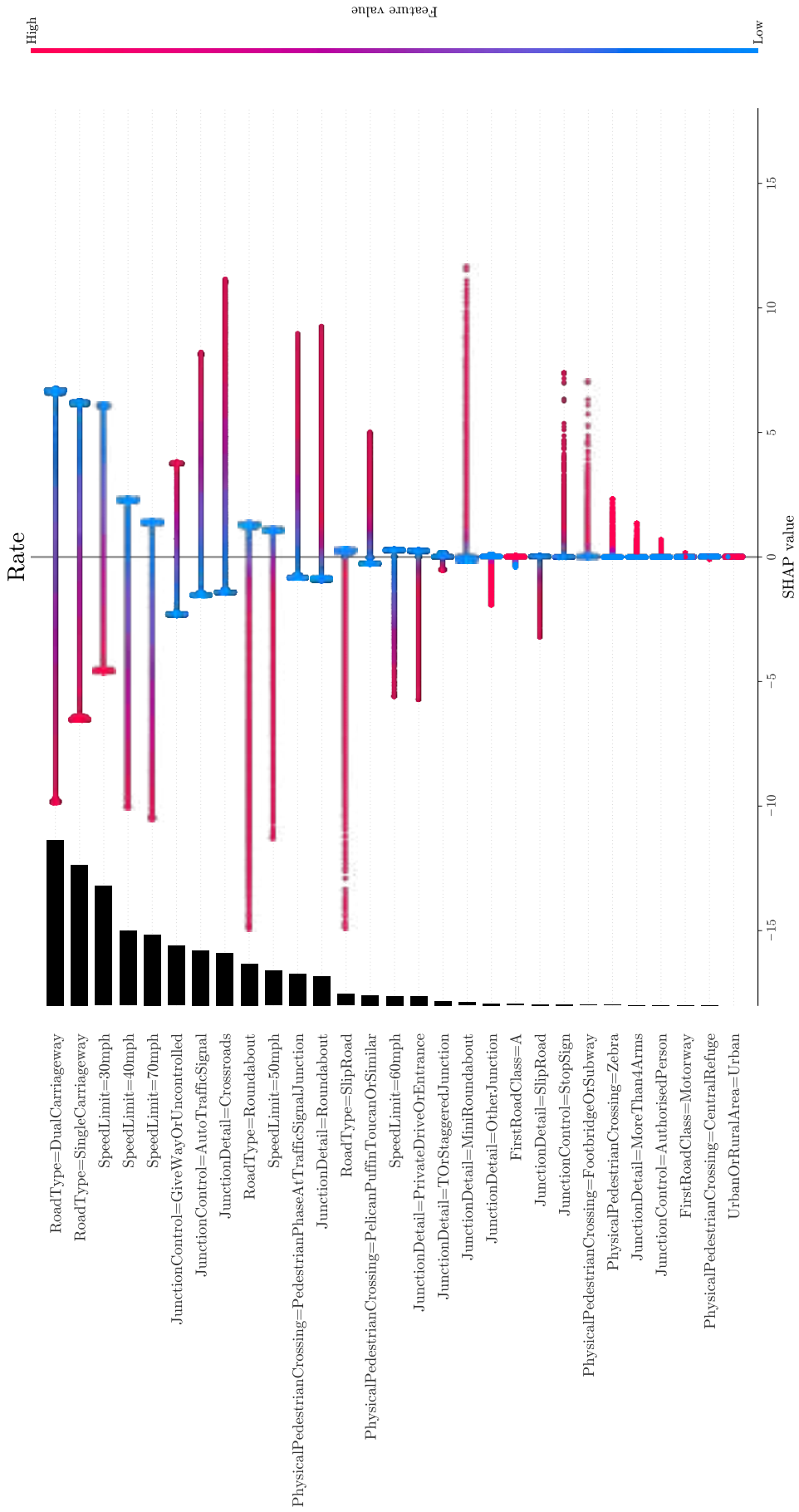


FIGURE G.1: *SHAP value results for the LR learning model trained to predict accident rate, where each road segment observation is plotted according to each semi-spatial attribute.*

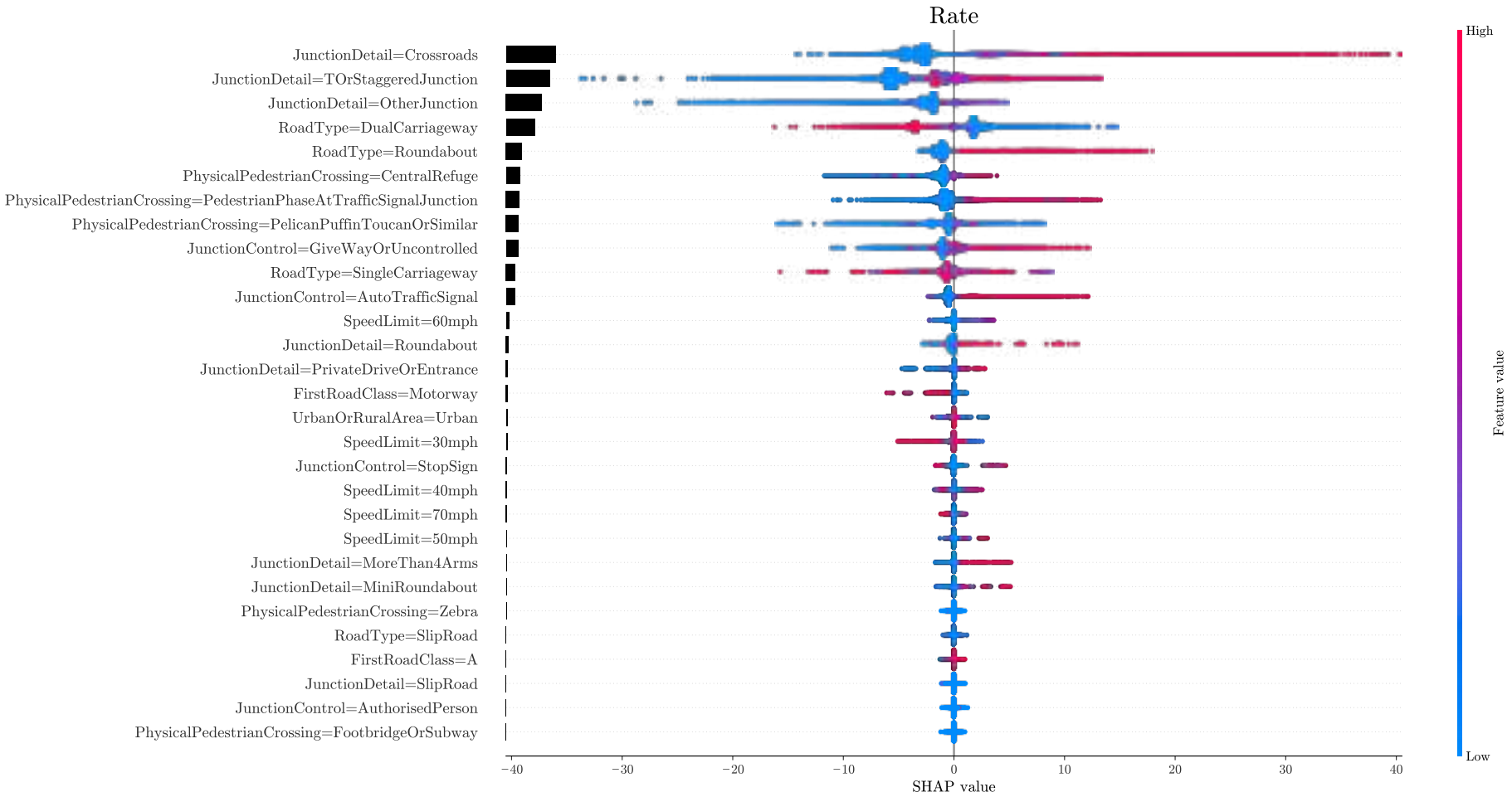


FIGURE G.2: *SHAP value results for the DT learning model trained to predict accident rate, where each road segment observation is plotted according to each semi-spatial attribute.*

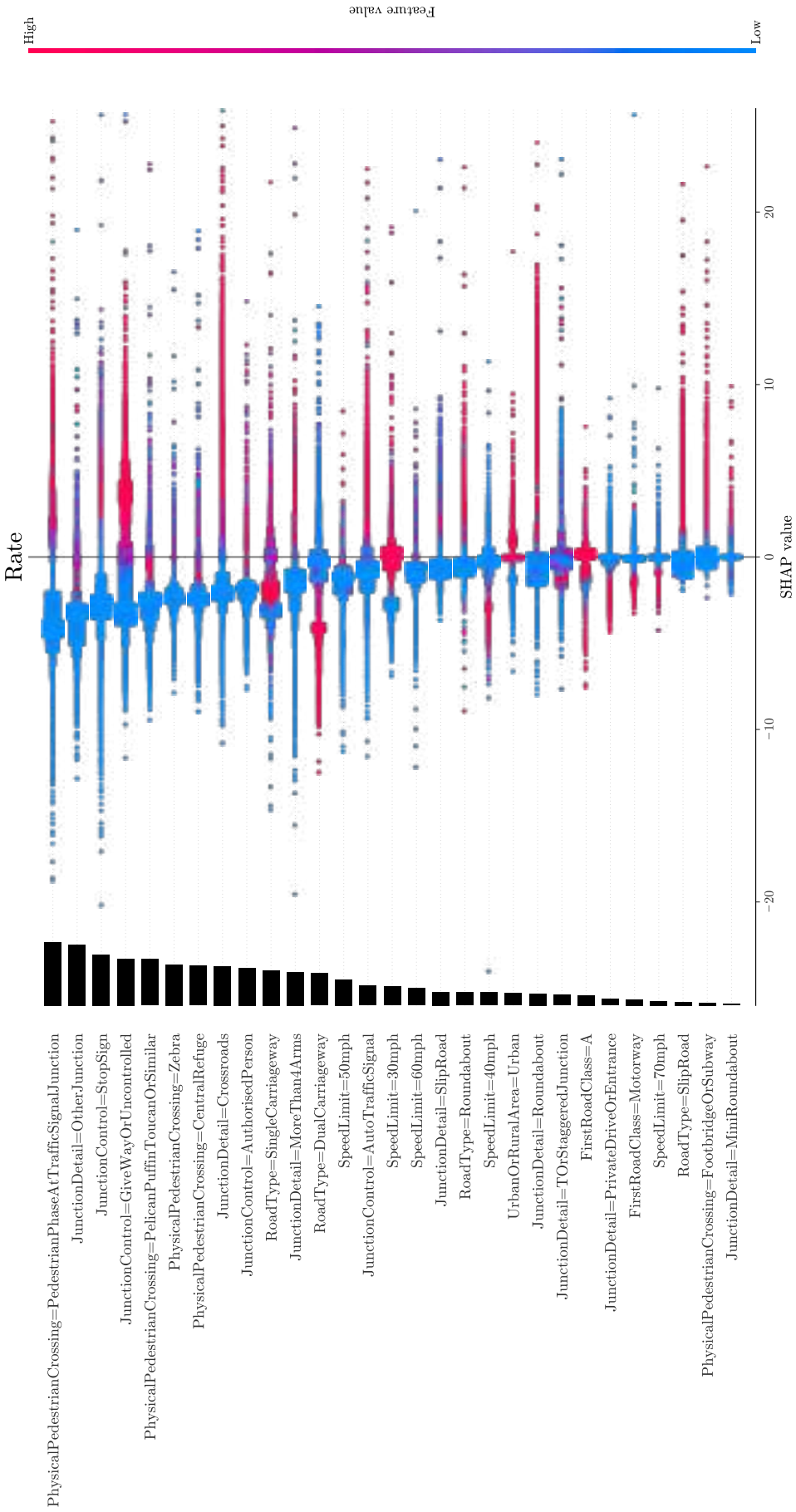


Figure G.3: *SHAP value results for the RF learning model trained to predict accident rate, where each road segment observation is plotted according to each semi-spatial attribute.*

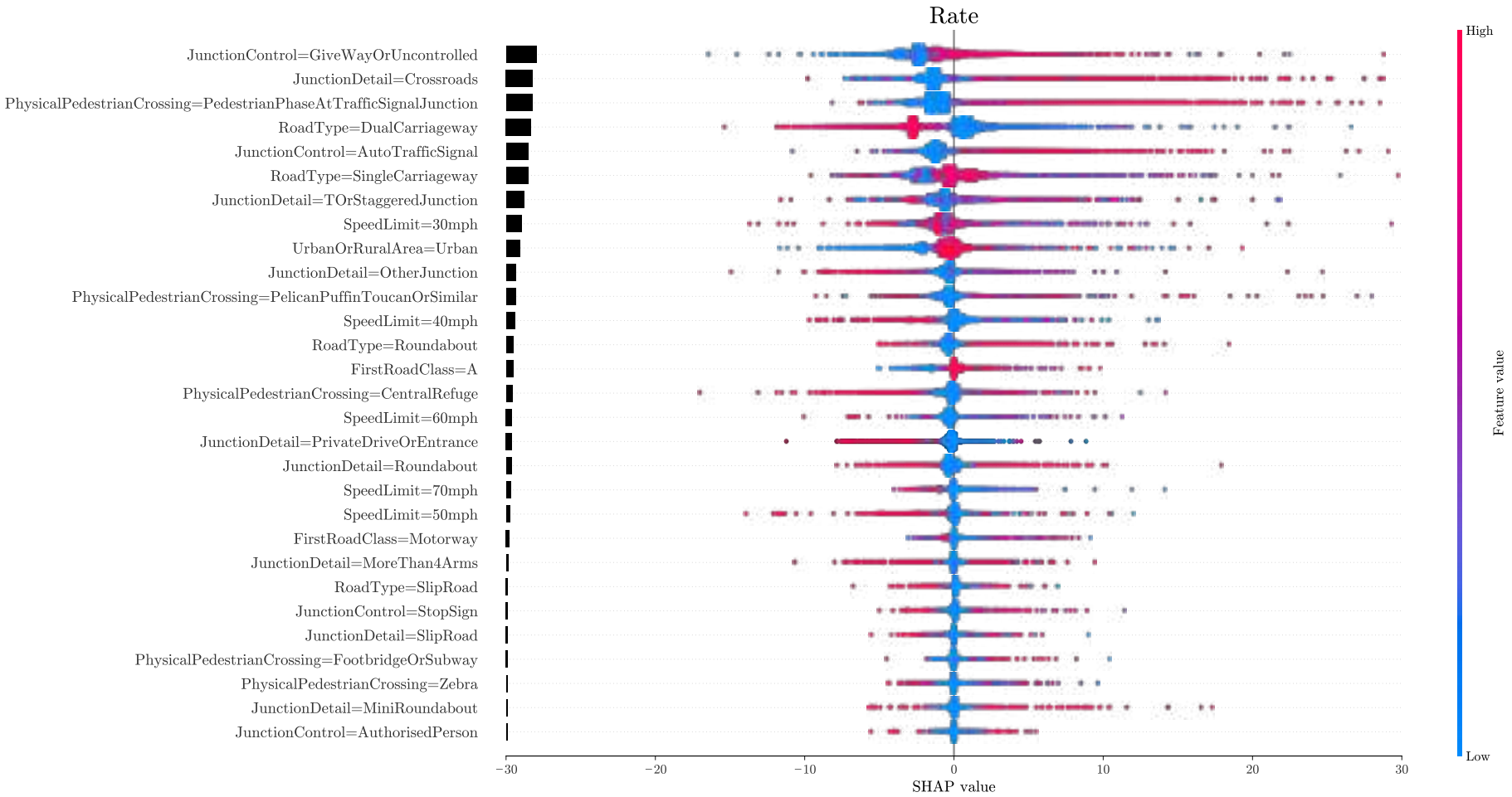


FIGURE G.4: *SHAP value results for the kNN learning model trained to predict accident rate, where each road segment observation is plotted according to each semi-spatial attribute.*

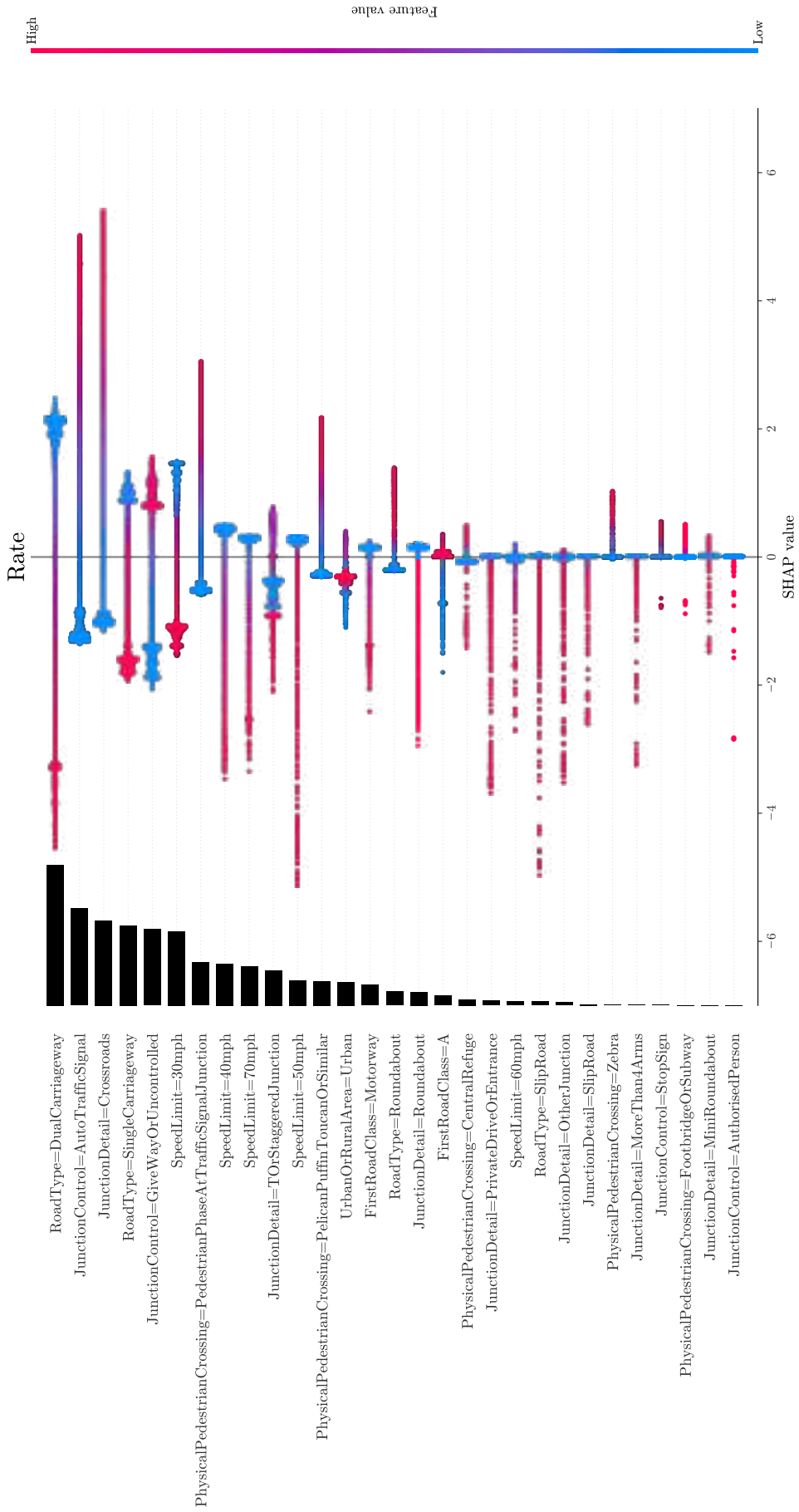


FIGURE G.5: *SHAP value results for the SVM learning model trained to predict accident rate, where each road segment observation is plotted according to each semi-spatial attribute.*

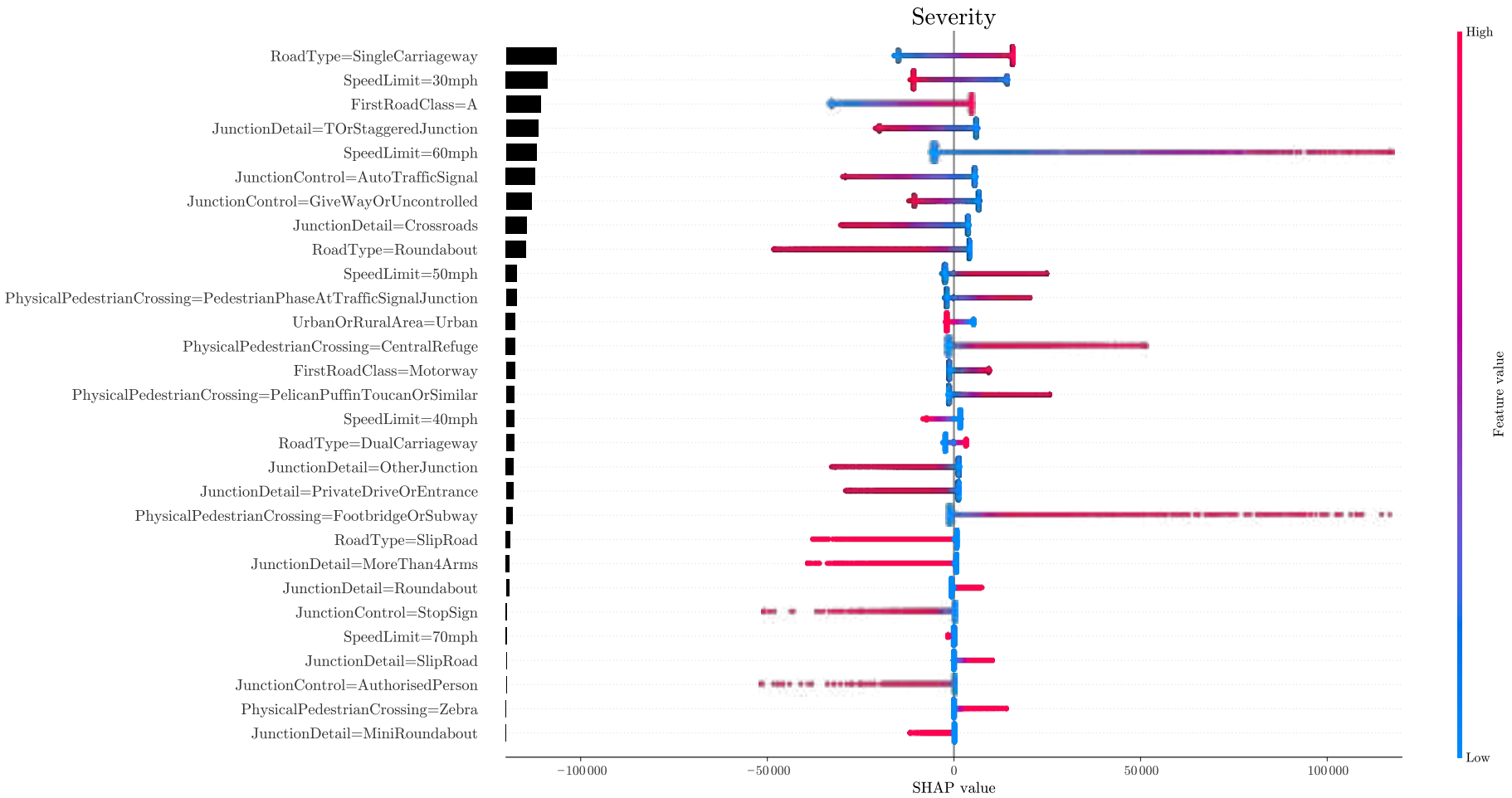


FIGURE G.6: *SHAP value results for the LR learning model trained to predict accident severity, where each road segment observation is plotted according to each semi-spatial attribute.*

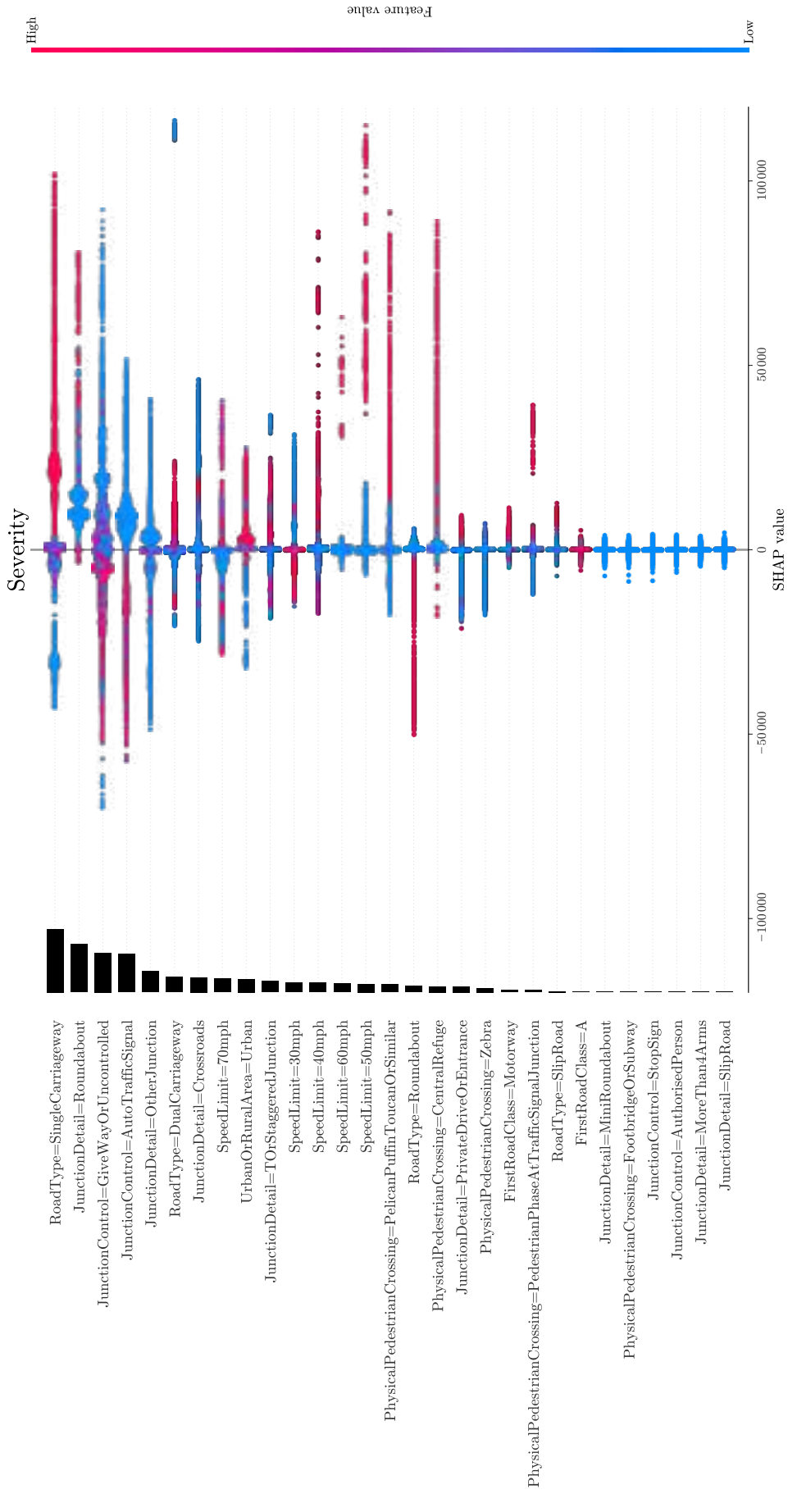


FIGURE G.7: *SHAP value results for the DT learning model trained to predict accident severity, where each road segment observation is plotted according to each semi-spatial attribute.*

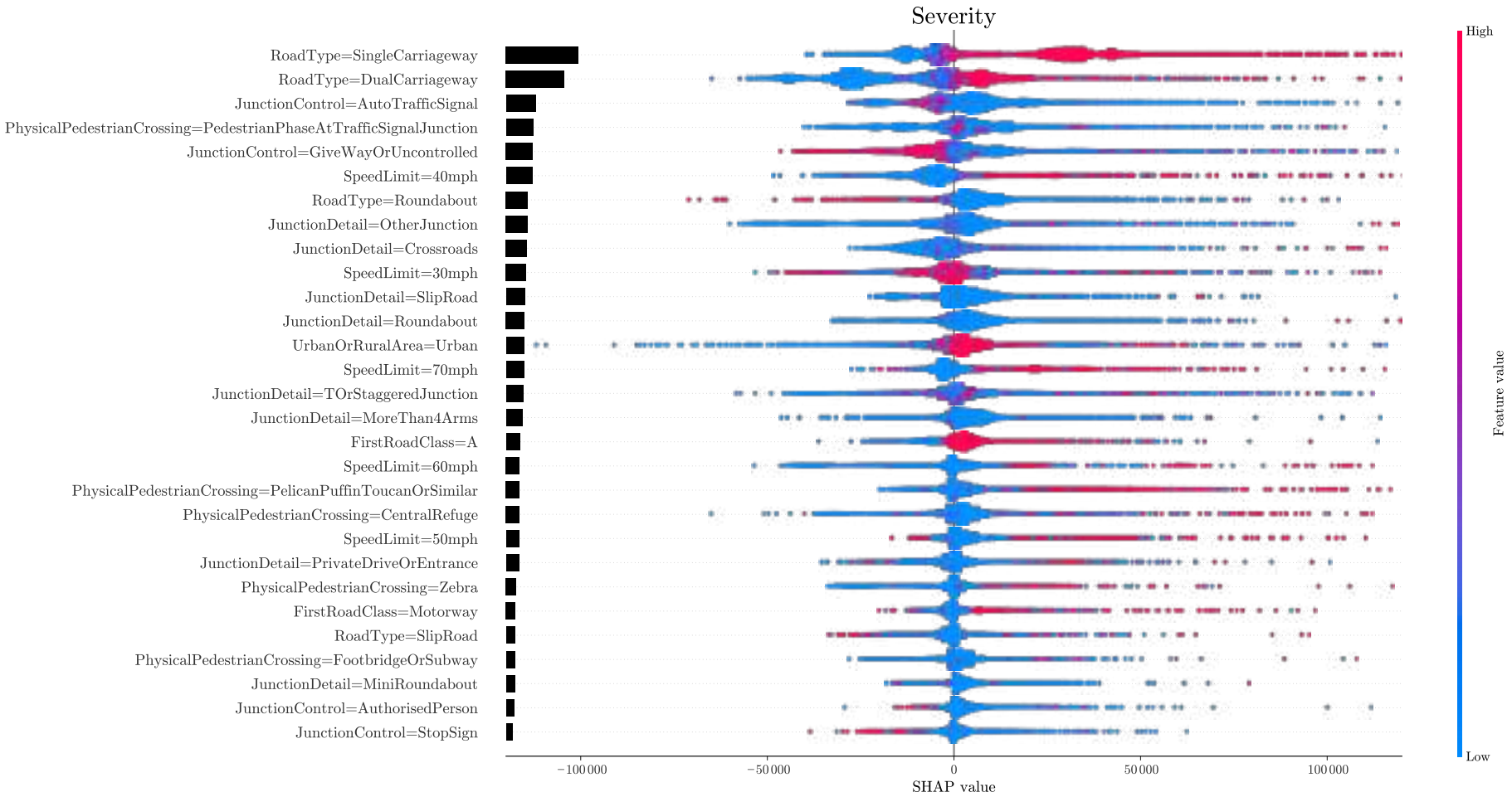


FIGURE G.8: *SHAP value results for the RF learning model trained to predict accident severity, where each road segment observation is plotted according to each semi-spatial attribute.*

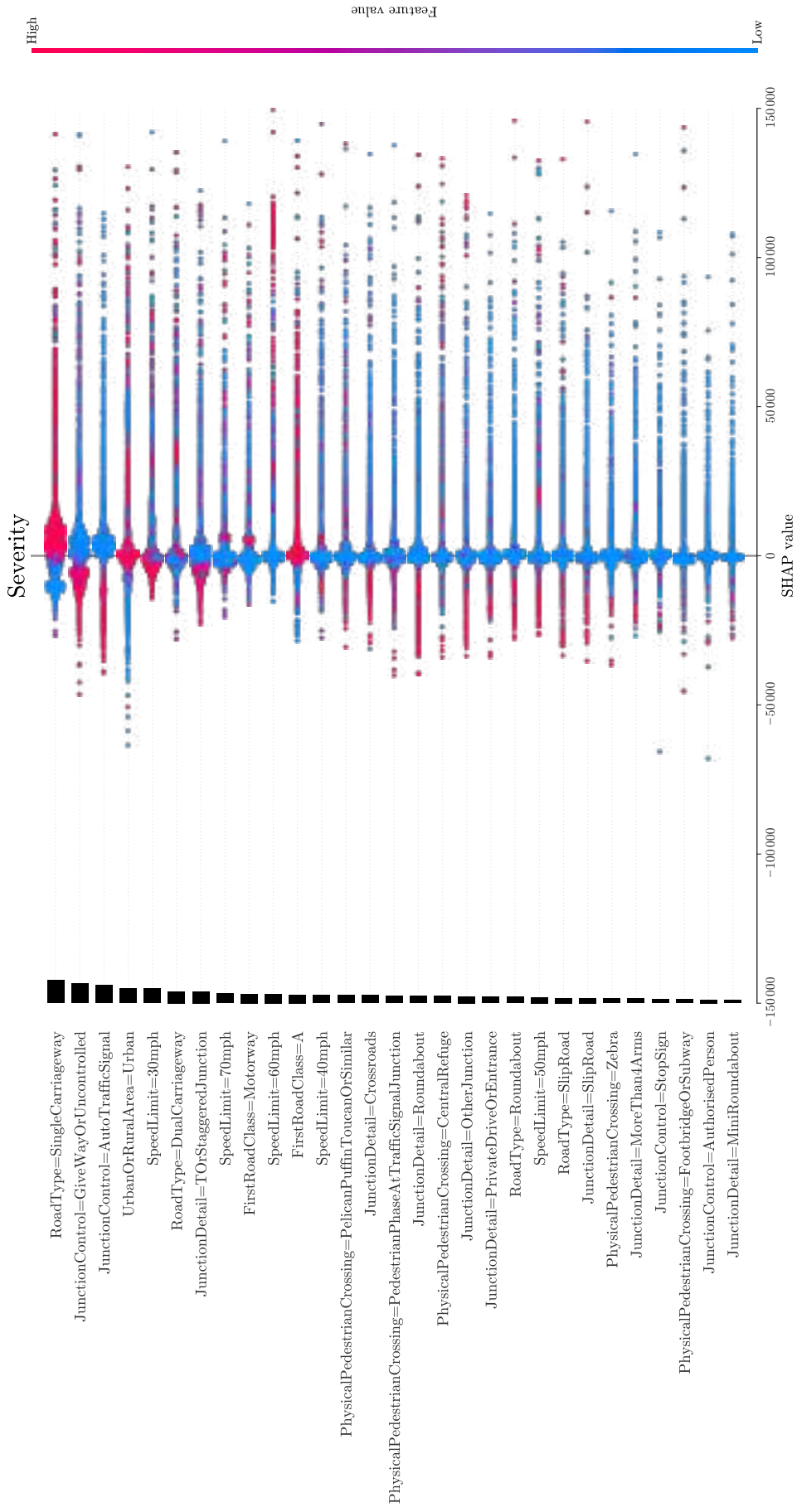


FIGURE G.9: *SHAP value results for the kNN learning model trained to predict accident severity, where each road segment observation is plotted according to each semi-spatial attribute.*

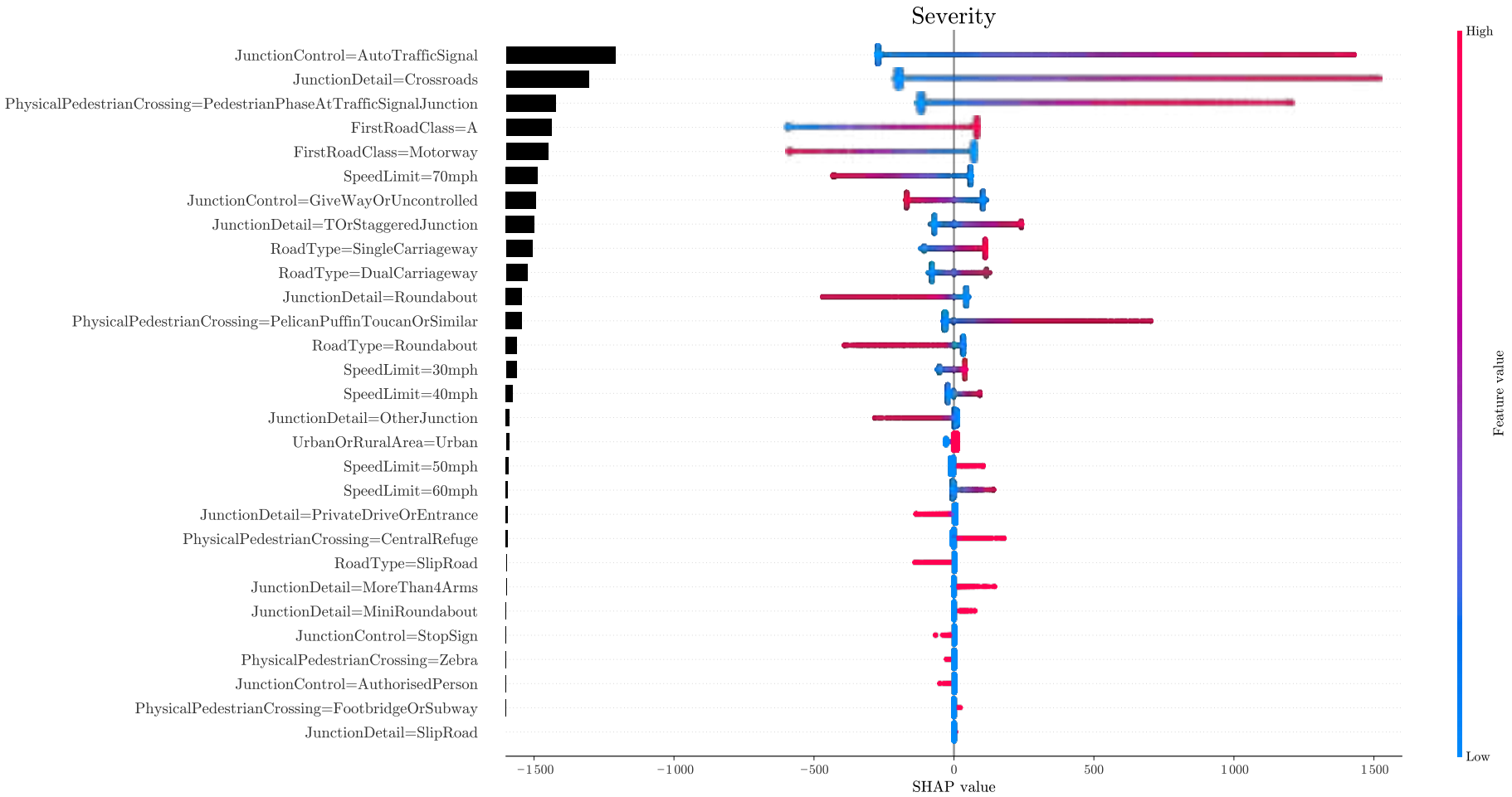


FIGURE G.10: *SHAP value results for the SVM learning model trained to predict accident severity, where each road segment observation is plotted according to each semi-spatial attribute.*

APPENDIX H

Dissimilarity threshold and similarity measure empirical study

An emperical study was conducted in order to determine an appropriate junction dissimilarity threshold β and similarity function Λ. In this study, the Lin, Goodall and Occurrence Frequency similarity measures where considered at different junction dissimilarity threshold values, the results of which may be found in Figures H.1, H.2 and H.3, respectively.

(a) $\beta = 0.0$

(b) $\beta = 0.5$

(c) $\beta = 0.10$

Figure H.1: *Junction clusterings at different junction dissimilarity thresholds for the Lin similarity measure.*

(d) $\beta = 0.15$

(e) $\beta = 0.20$

(f) $\beta = 0.25$

FIGURE H.1: *Junction clusterings at different junction dissimilarity thresholds for the Lin similarity measure.*

(g) $\beta = 0.30$

(h) $\beta = 0.35$

(i) $\beta = 0.40$

Figure H.1: *Junction clusterings at different junction dissimilarity thresholds for the Lin similarity measure.*

(j) $\beta = 0.45$

(k) $\beta = 0.50$

(l) $\beta = 0.55$

FIGURE H.1: *Junction clusterings at different junction dissimilarity thresholds for the Lin similarity measure.*

(m) $\beta = 0.60$

(n) $\beta = 0.65$

(o) $\beta = 0.70$

Figure H.1: *Junction clusterings at different junction dissimilarity thresholds for the Lin similarity measure.*

(p) $\beta = 0.75$

(q) $\beta = 0.80$

(r) $\beta = 0.85$

FIGURE H.1: *Junction clusterings at different junction dissimilarity thresholds for the Lin similarity measure.*

(s) $\beta = 0.90$

(t) $\beta = 0.95$

(u) $\beta = 1.00$

Figure H.1: *Junction clusterings at different junction dissimilarity thresholds for the Lin similarity measure.*

(a) $\beta = 0.0$

(b) $\beta = 0.5$

(c) $\beta = 0.10$

FIGURE H.2: *Junction clusterings at different junction dissimilarity thresholds for the Goodall similarity measure.*

(d) $\beta = 0.15$

(e) $\beta = 0.20$

(f) $\beta = 0.25$

Figure H.2: *Junction clusterings at different junction dissimilarity thresholds for the Goodall similarity measure.*

(g) $\beta = 0.30$

(h) $\beta = 0.35$

(i) $\beta = 0.40$

FIGURE H.2: *Junction clusterings at different junction dissimilarity thresholds for the Goodall similarity measure.*

(j) $\beta = 0.45$

(k) $\beta = 0.50$

(l) $\beta = 0.55$

Figure H.2: *Junction clusterings at different junction dissimilarity thresholds for the Goodall similarity measure.*

(m) $\beta = 0.60$

(n) $\beta = 0.65$

(o) $\beta = 0.70$

FIGURE H.2: *Junction clusterings at different junction dissimilarity thresholds for the Goodall similarity measure.*

(p) $\beta = 0.75$

(q) $\beta = 0.80$

(r) $\beta = 0.85$

Figure H.2: *Junction clusterings at different junction dissimilarity thresholds for the Goodall similarity measure.*

(s) $\beta = 0.90$

(t) $\beta = 0.95$

(u) $\beta = 1.00$

FIGURE H.2: *Junction clusterings at different junction dissimilarity thresholds for the Goodall similarity measure.*

(a) $\beta = 0.0$

(b) $\beta = 0.5$

(c) $\beta = 0.10$

Figure H.3: *Junction clusterings at different junction dissimilarity thresholds for the Occurrence Frequency similarity measure.*

(d) $\beta = 0.15$

(e) $\beta = 0.20$

(f) $\beta = 0.25$

FIGURE H.3: *Junction clusterings at different junction dissimilarity thresholds for the Occurrence Frequency similarity measure.*

(g) $\beta = 0.30$

(h) $\beta = 0.35$

(i) $\beta = 0.40$

Figure H.3: *Junction clusterings at different junction dissimilarity thresholds for the Occurrence Frequency similarity measure.*

(j) $\beta = 0.45$

(k) $\beta = 0.50$

(l) $\beta = 0.55$

FIGURE H.3: *Junction clusterings at different junction dissimilarity thresholds for the Occurrence Frequency similarity measure.*

(m) $\beta = 0.60$

(n) $\beta = 0.65$

(o) $\beta = 0.70$

Figure H.3: *Junction clusterings at different junction dissimilarity thresholds for the Occurrence Frequency similarity measure.*

(p) $\beta = 0.75$

(q) $\beta = 0.80$

(r) $\beta = 0.85$

FIGURE H.3: *Junction clusterings at different junction dissimilarity thresholds for the Occurrence Frequency similarity measure.*

(s) $\beta = 0.90$

(t) $\beta = 0.95$

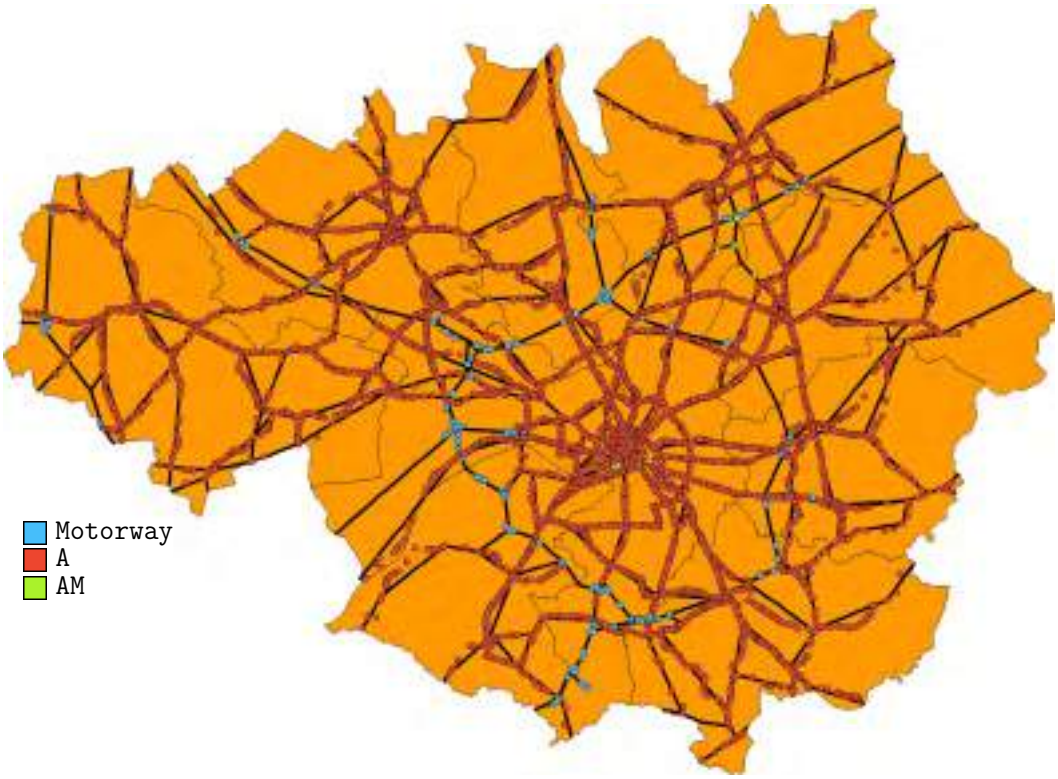
(u) $\beta = 1.00$

Figure H.3: *Junction clusterings at different junction dissimilarity thresholds for the Occurrence Frequency similarity measure.*

APPENDIX I

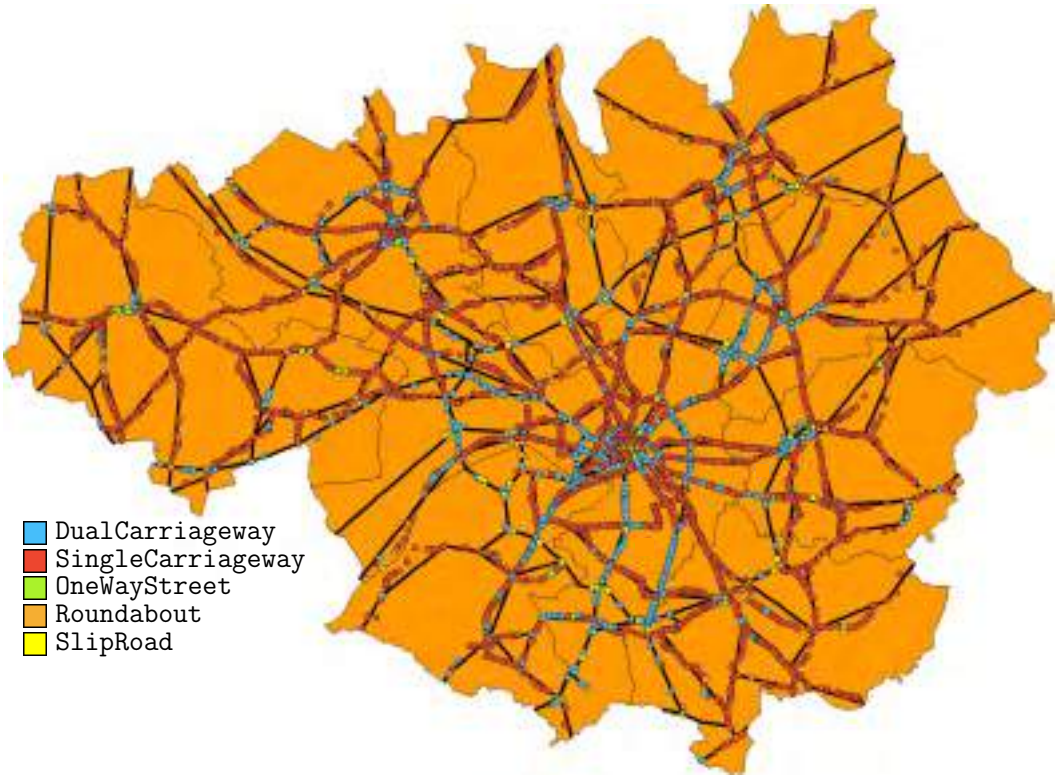
Road junction spatial representation

The spatial results corresponding to the `FirstRoadClass`, `RoadType`, `SpeedLimit` and `UrbanOrRuralArea` semi-spatial attributes determined during the demonstration of Sub-module 8.5 in §8.1.2 are shown in Figure F.1.

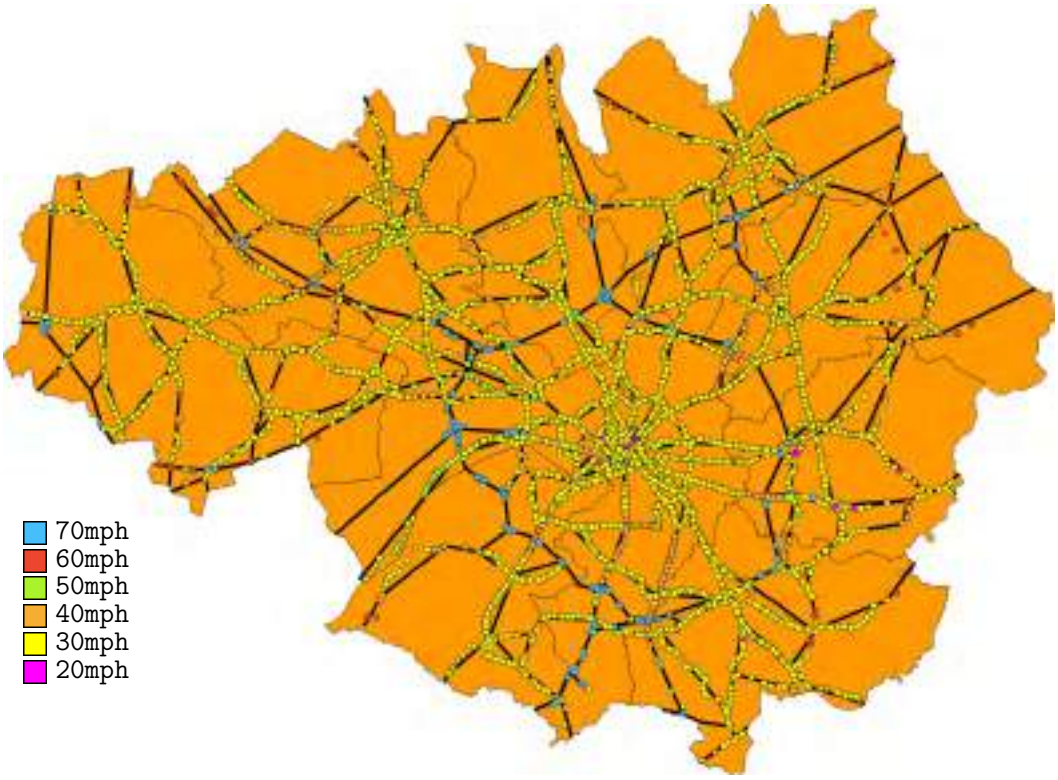


(a) `FirstRoadClass`

FIGURE I.1: *A spatial representation of different road junction attribute values in Greater Manchester.*

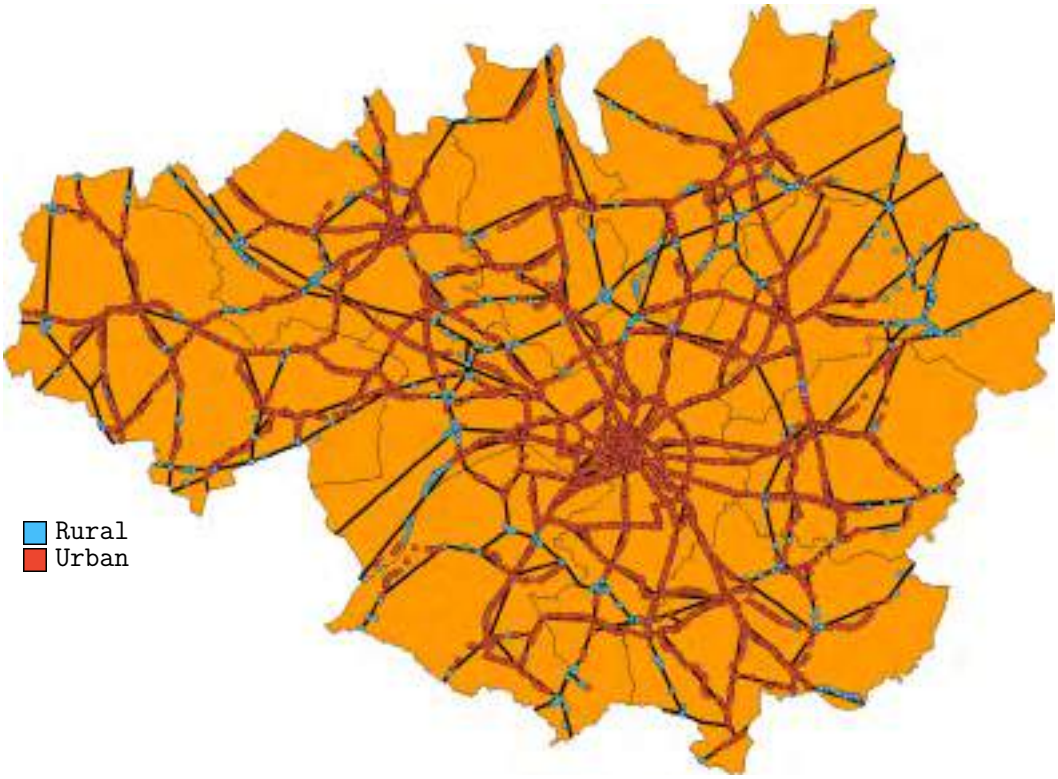


(b) RoadType



(c) SpeedLimit

FIGURE I.1: *A spatial representation of different road junction attribute values in Greater Manchester.*



(d) `UrbanOrRuralArea`

FIGURE I.1: *A spatial representation of different road junction attribute values in Greater Manchester.*

APPENDIX J

Road junction SHAP values

The SHAP values corresponding to the LR, DT, RF, kNN and SVM learning models trained to predict accident rate and severity at a road junction during the demonstration of Sub-module 9.5 in §8.1.5 are contained in this appendix. The learning models trained to predict accident rate are shown in Figures J.1–J.5, while those that were trained to predict accident severity are shown in Figures J.6–J.10, respectively.

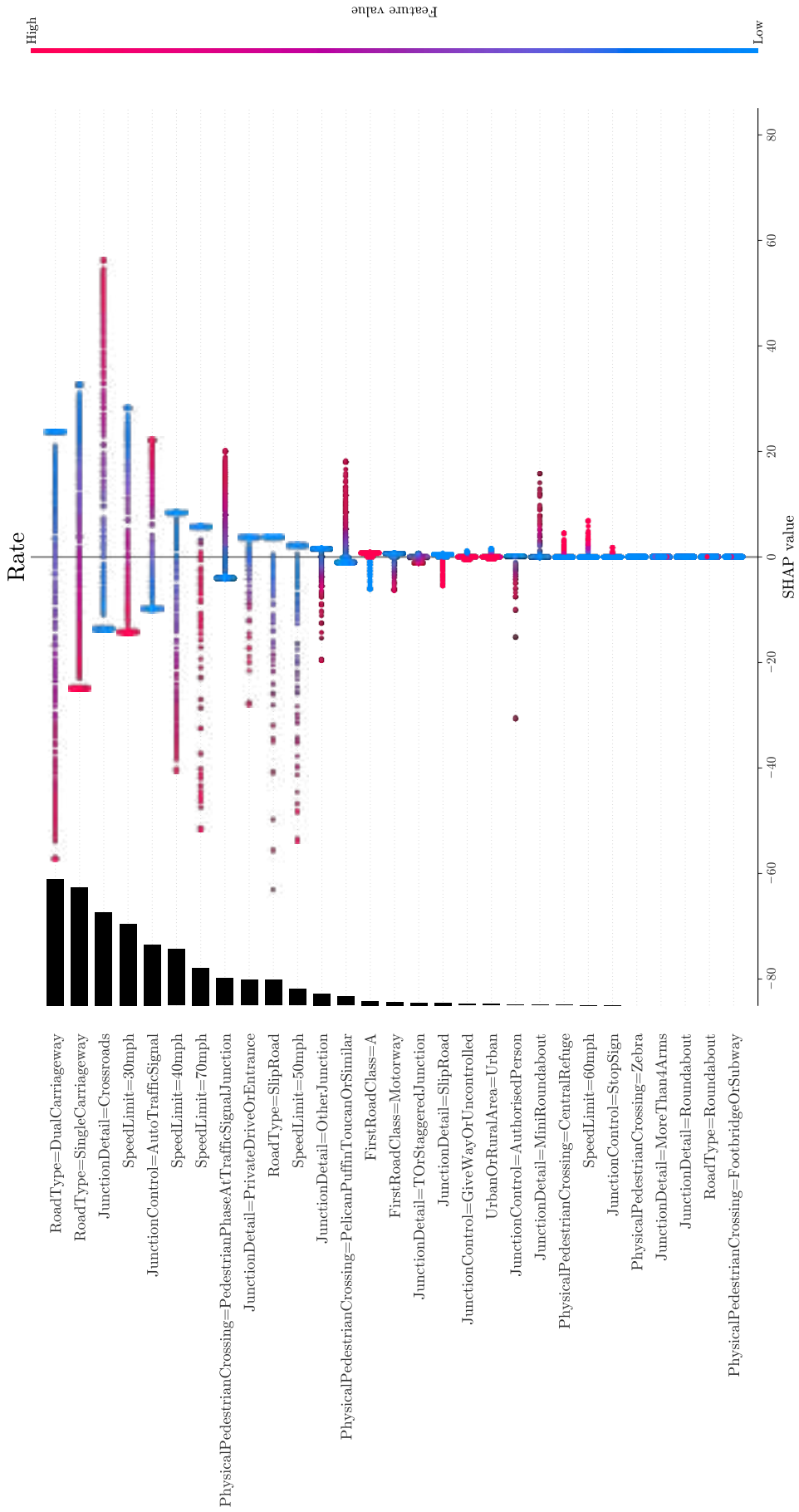


FIGURE J.1: *SHAP value results for the LR learning model trained to predict accident rate, where each road junction observation is plotted according to each semi-spatial attribute.*

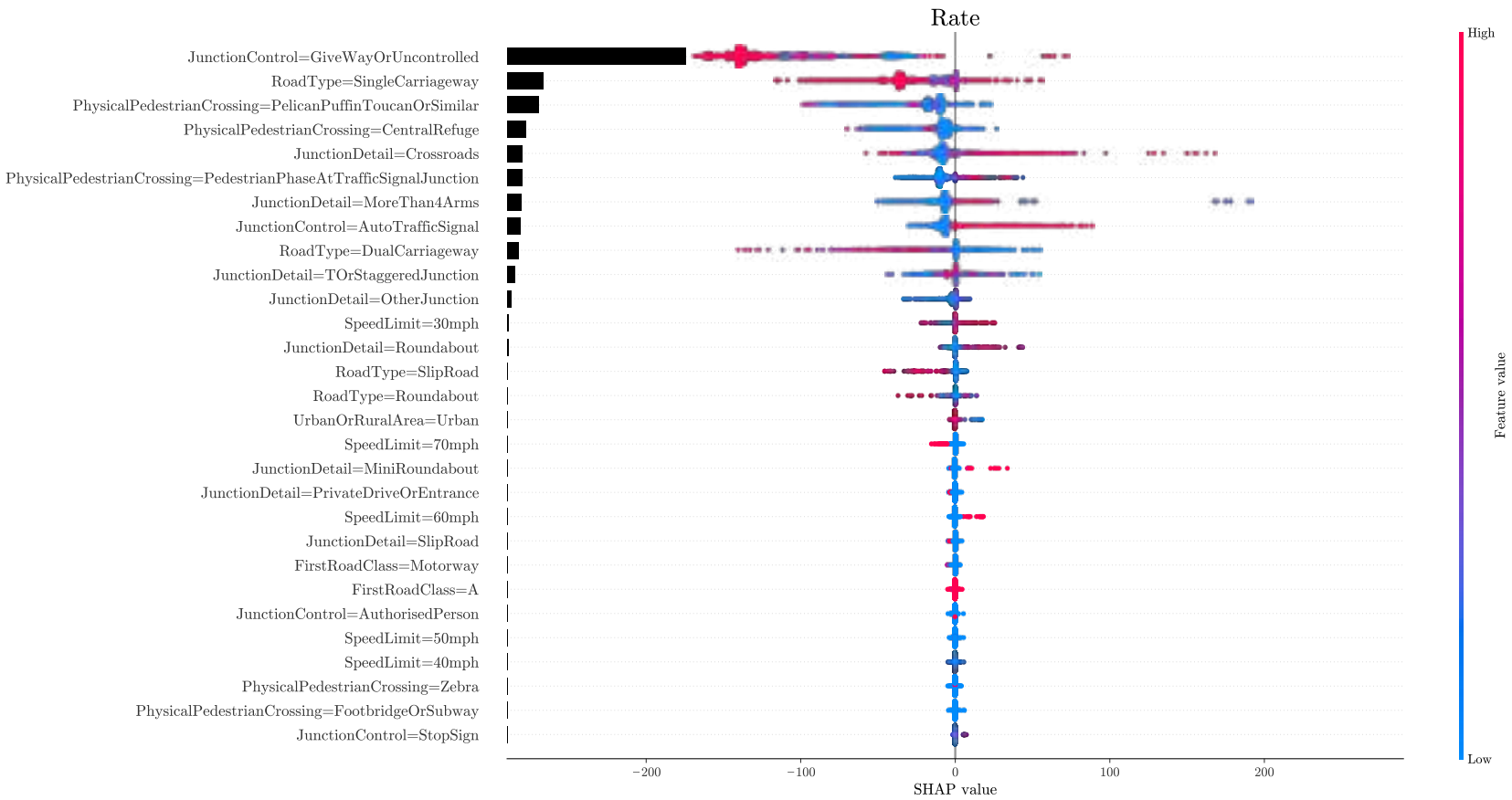


FIGURE J.2: *SHAP value results for the DT learning model trained to predict accident rate, where each road junction observation is plotted according to each semi-spatial attribute.*

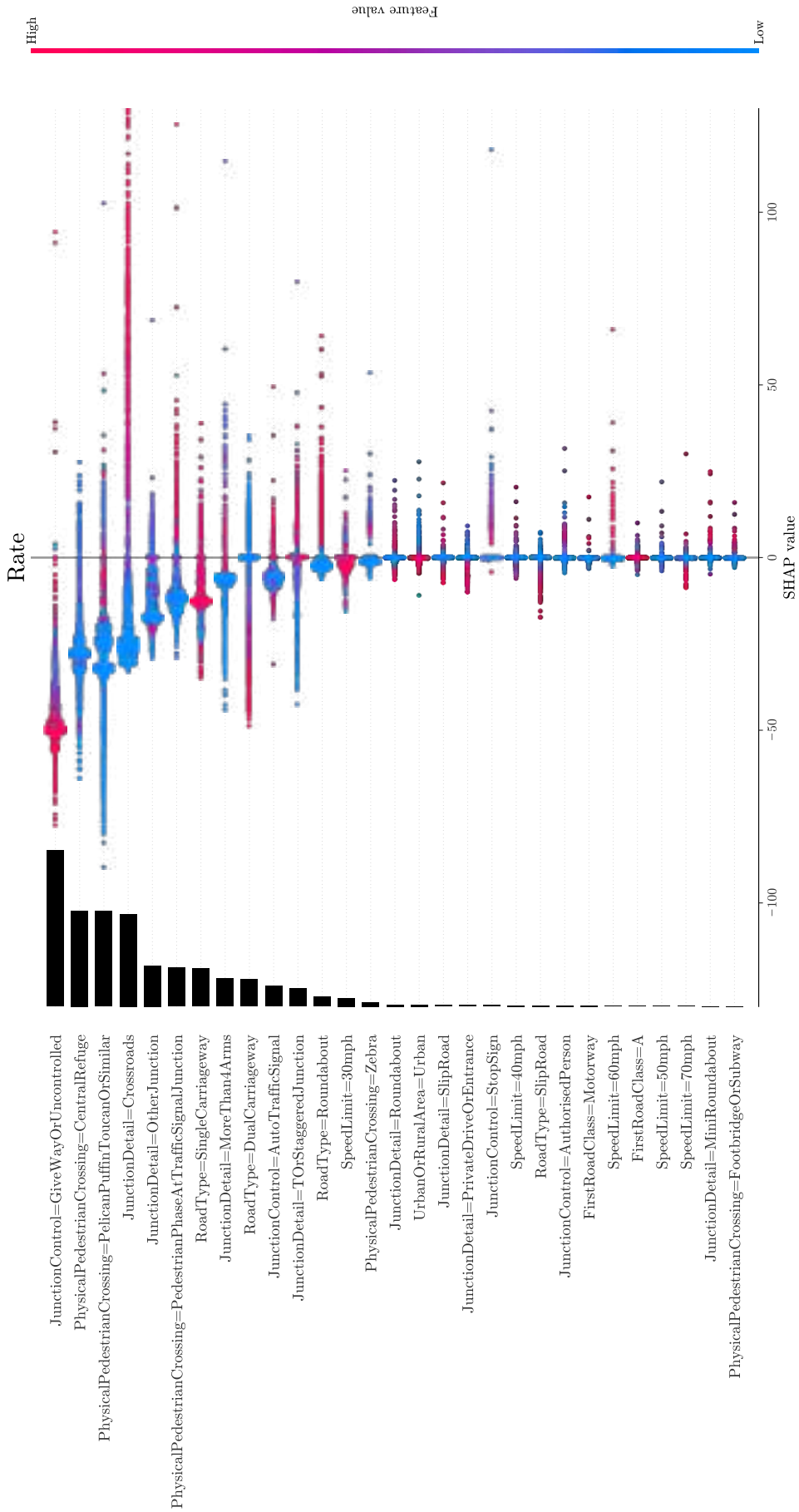


FIGURE J.3: *SHAP value results for the RF learning model trained to predict accident rate, where each road junction observation is plotted according to each semi-spatial attribute.*

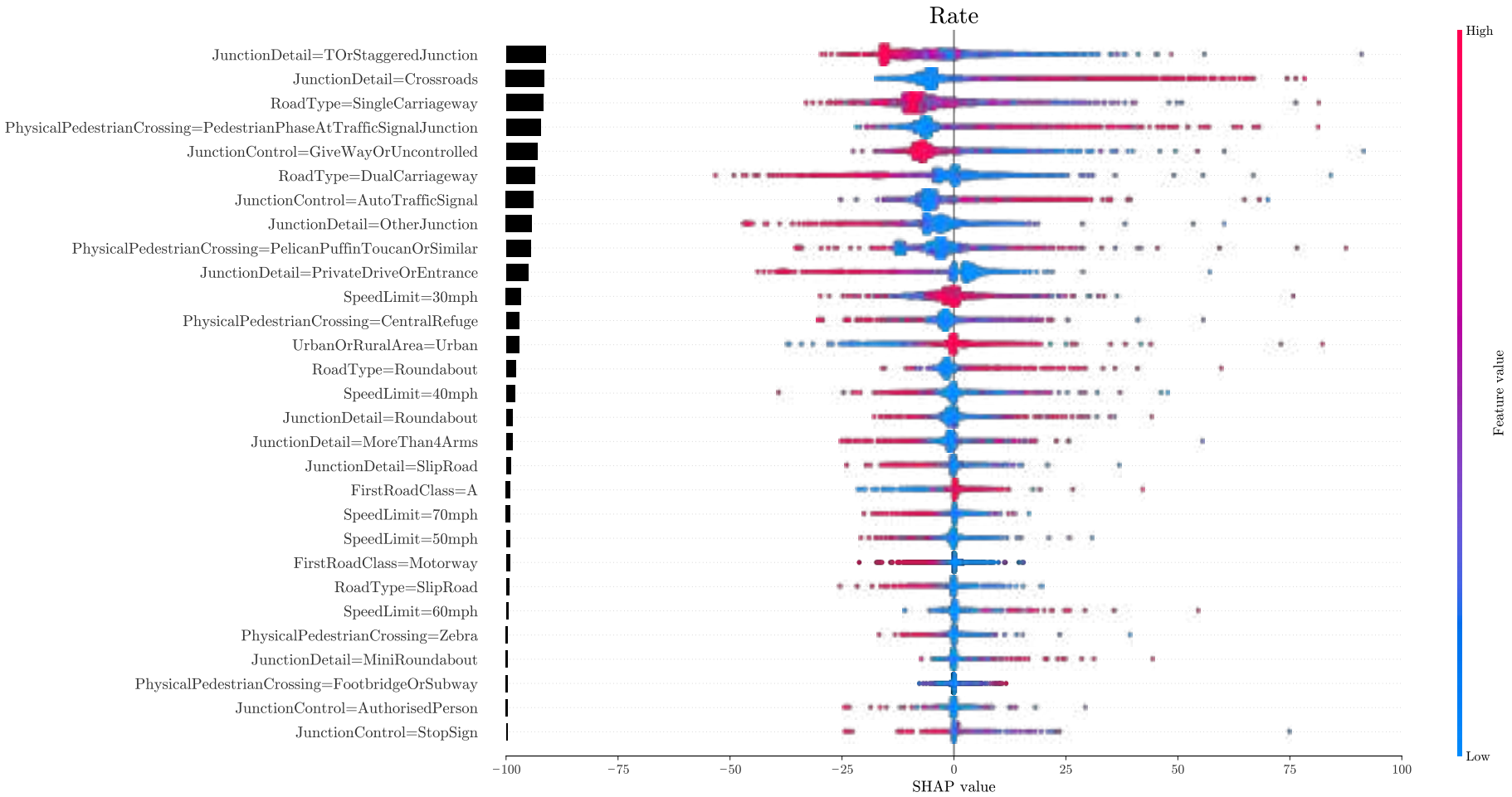


FIGURE J.4: *SHAP value results for the kNN learning model trained to predict accident rate, where each road junction observation is plotted according to each semi-spatial attribute.*

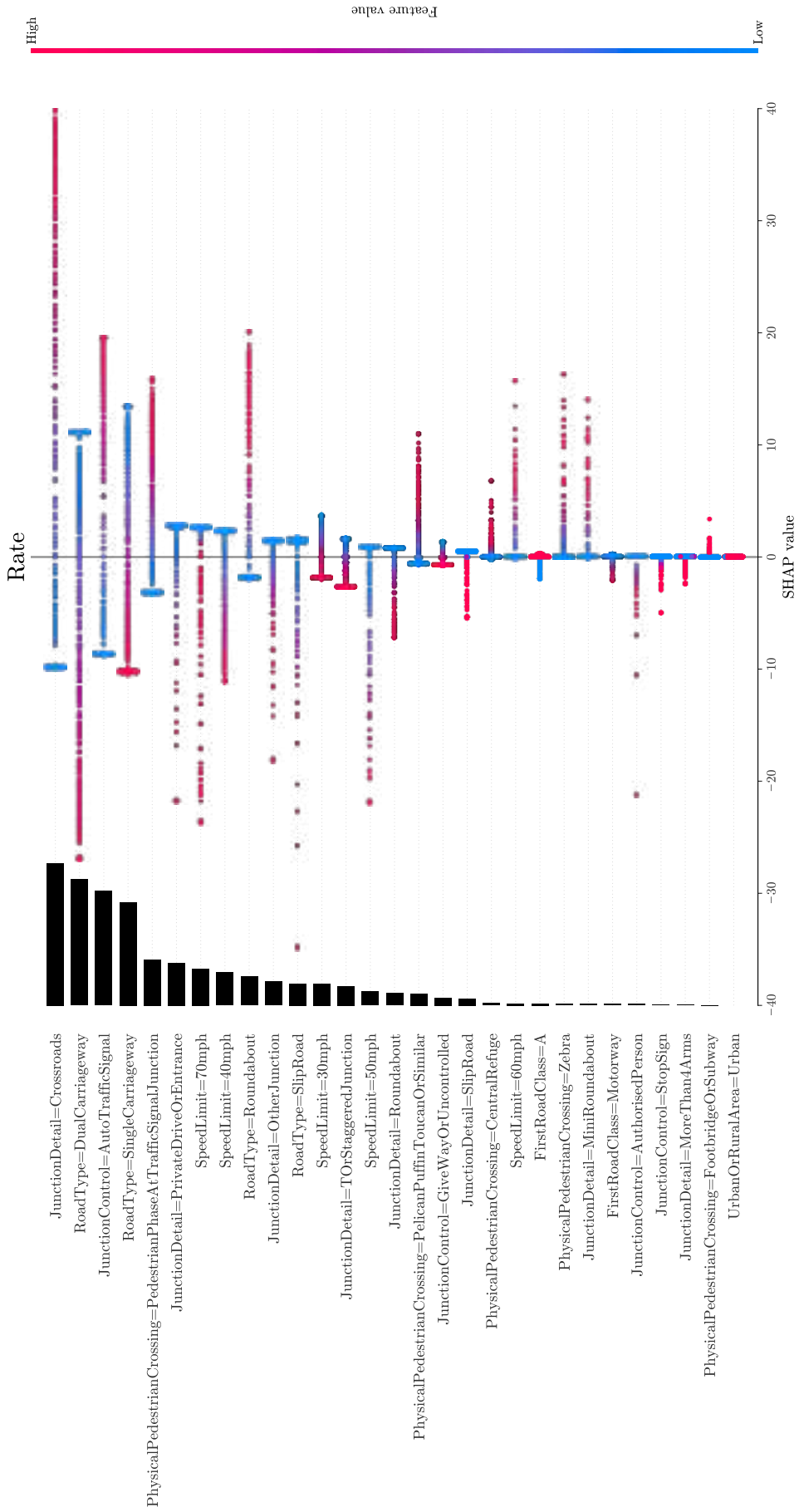


FIGURE J.5: *SHAP value results for the SVM learning model trained to predict accident rate, where each road junction observation is plotted according to each semi-spatial attribute.*

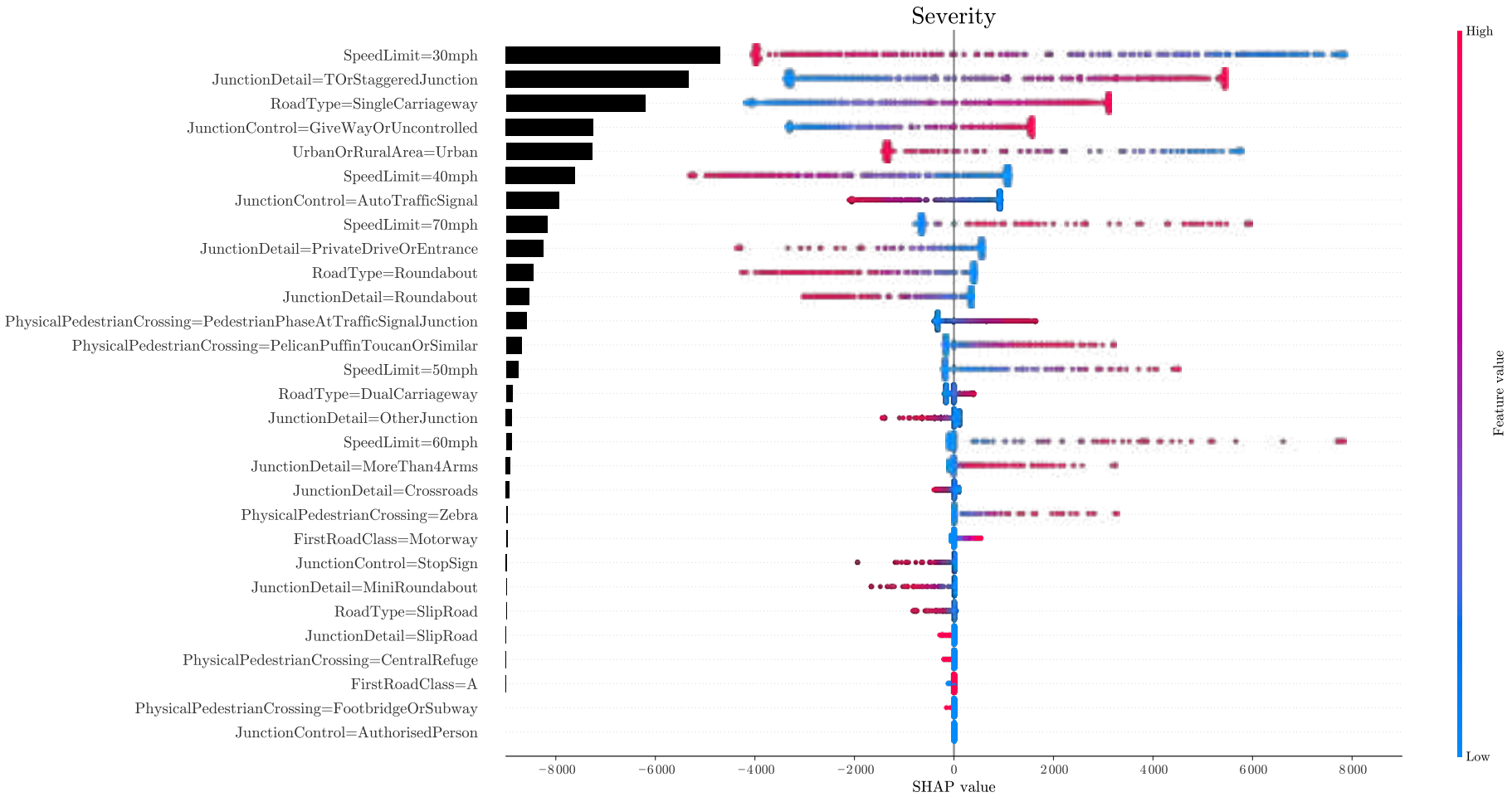


FIGURE J.6: *SHAP value results for the LR learning model trained to predict accident severity, where each road junction observation is plotted according to each semi-spatial attribute.*

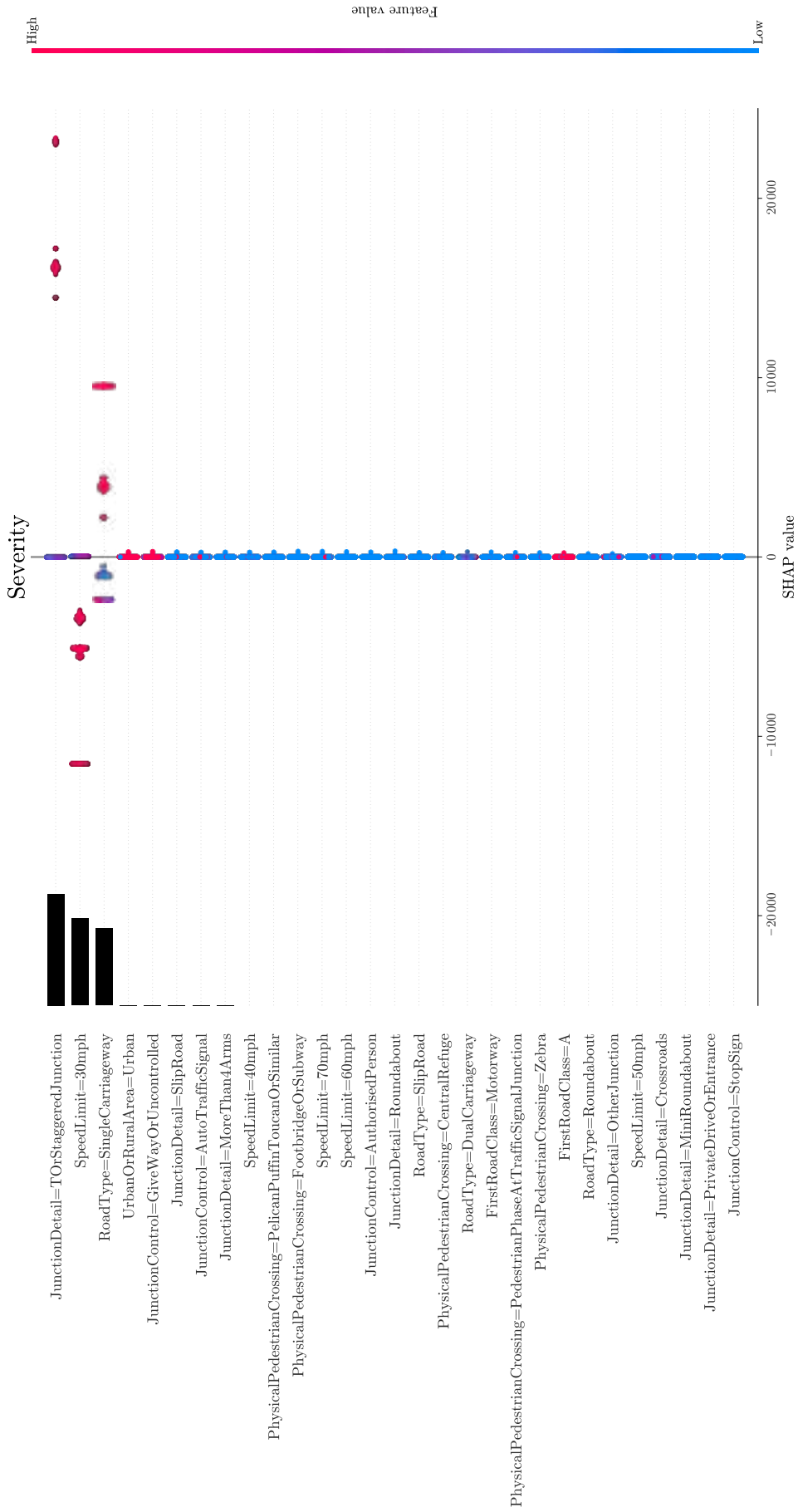


FIGURE J.7: *SHAP value results for the DT learning model trained to predict accident severity, where each road junction observation is plotted according to each semi-spatial attribute.*

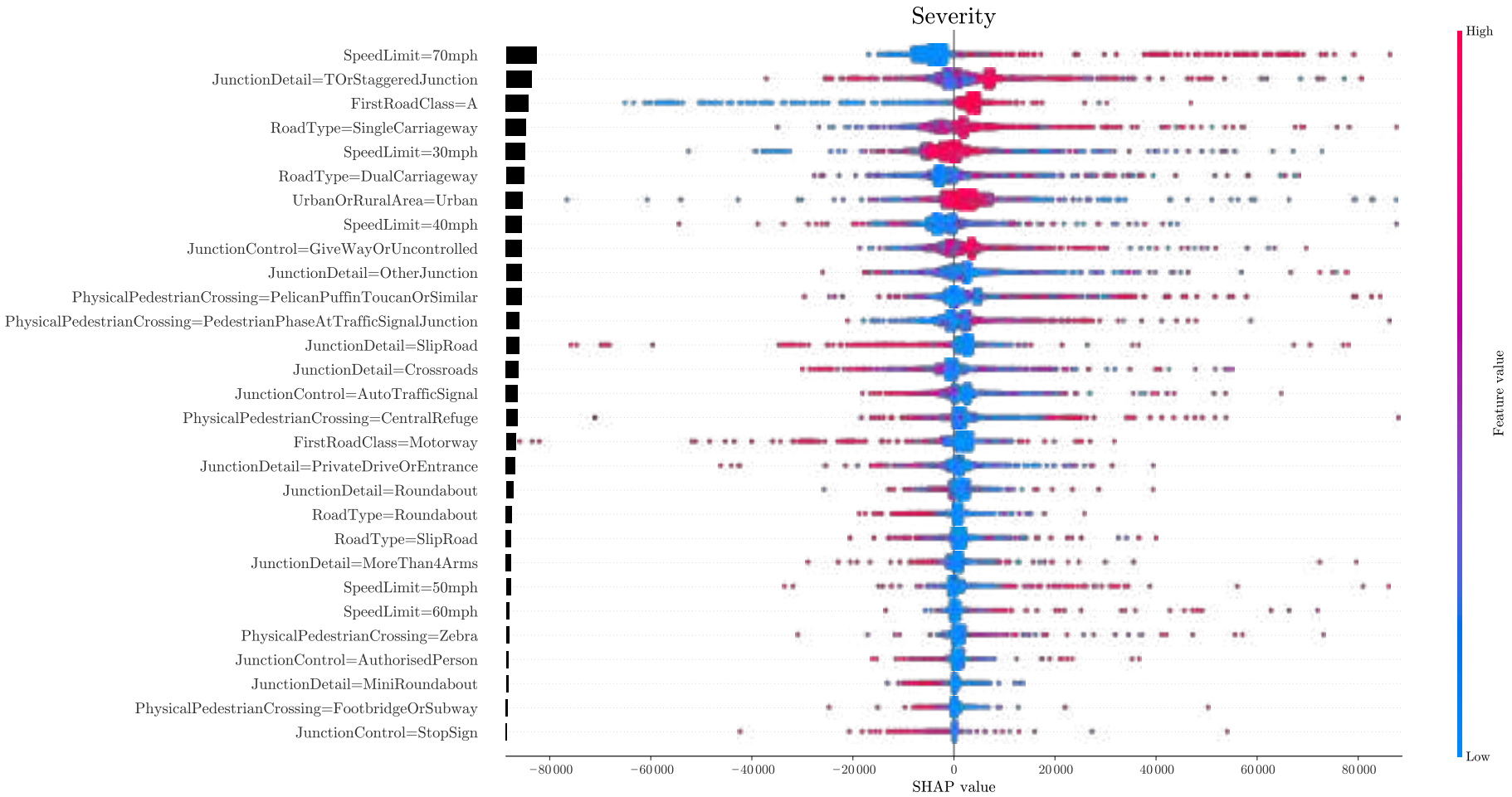


FIGURE J.8: *SHAP value results for the RF learning model trained to predict accident severity, where each road junction observation is plotted according to each semi-spatial attribute.*

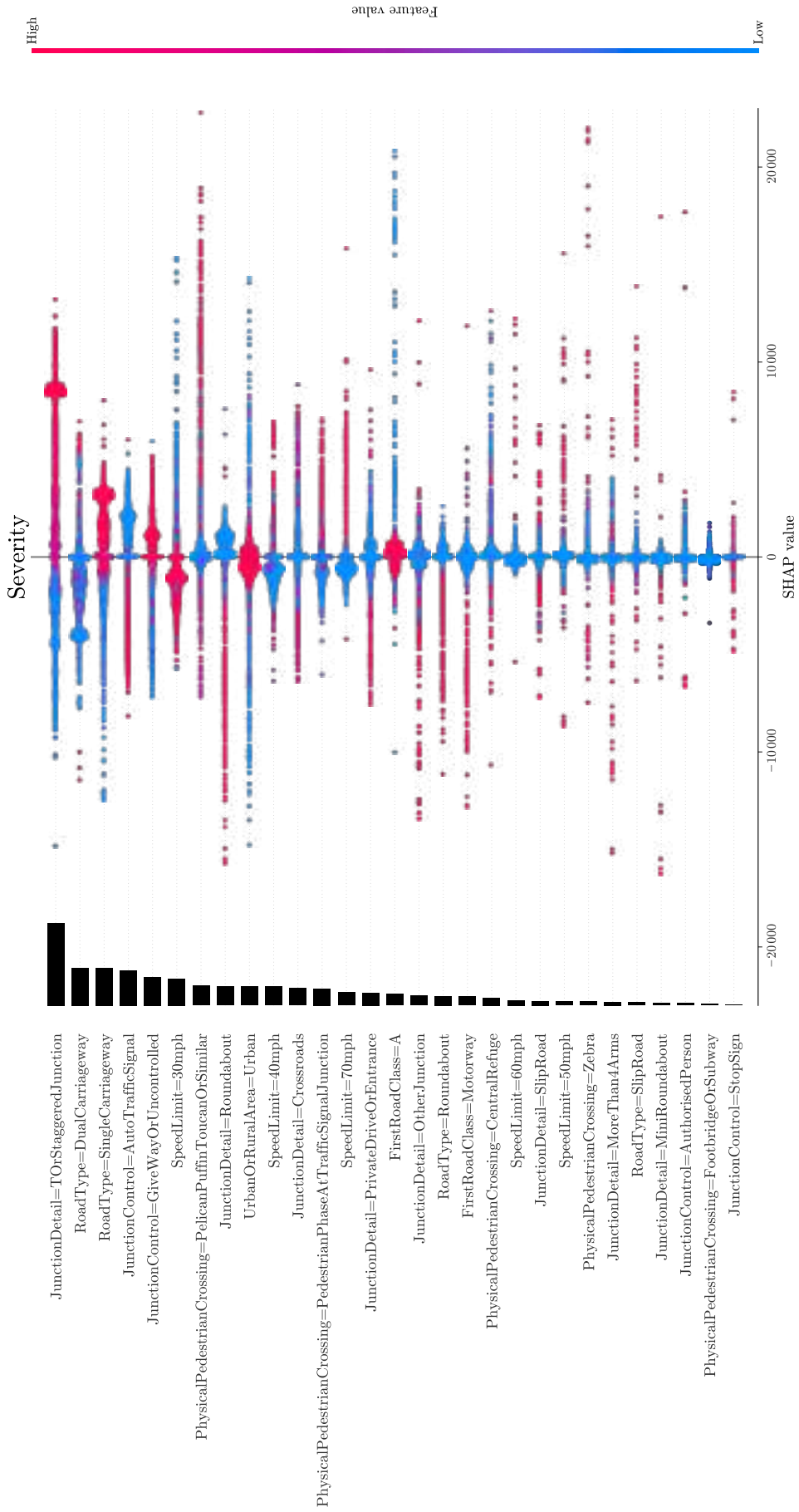


FIGURE J.9: *SHAP value results for the kNN learning model trained to predict accident severity, where each road junction observation is plotted according to each semi-spatial attribute.*

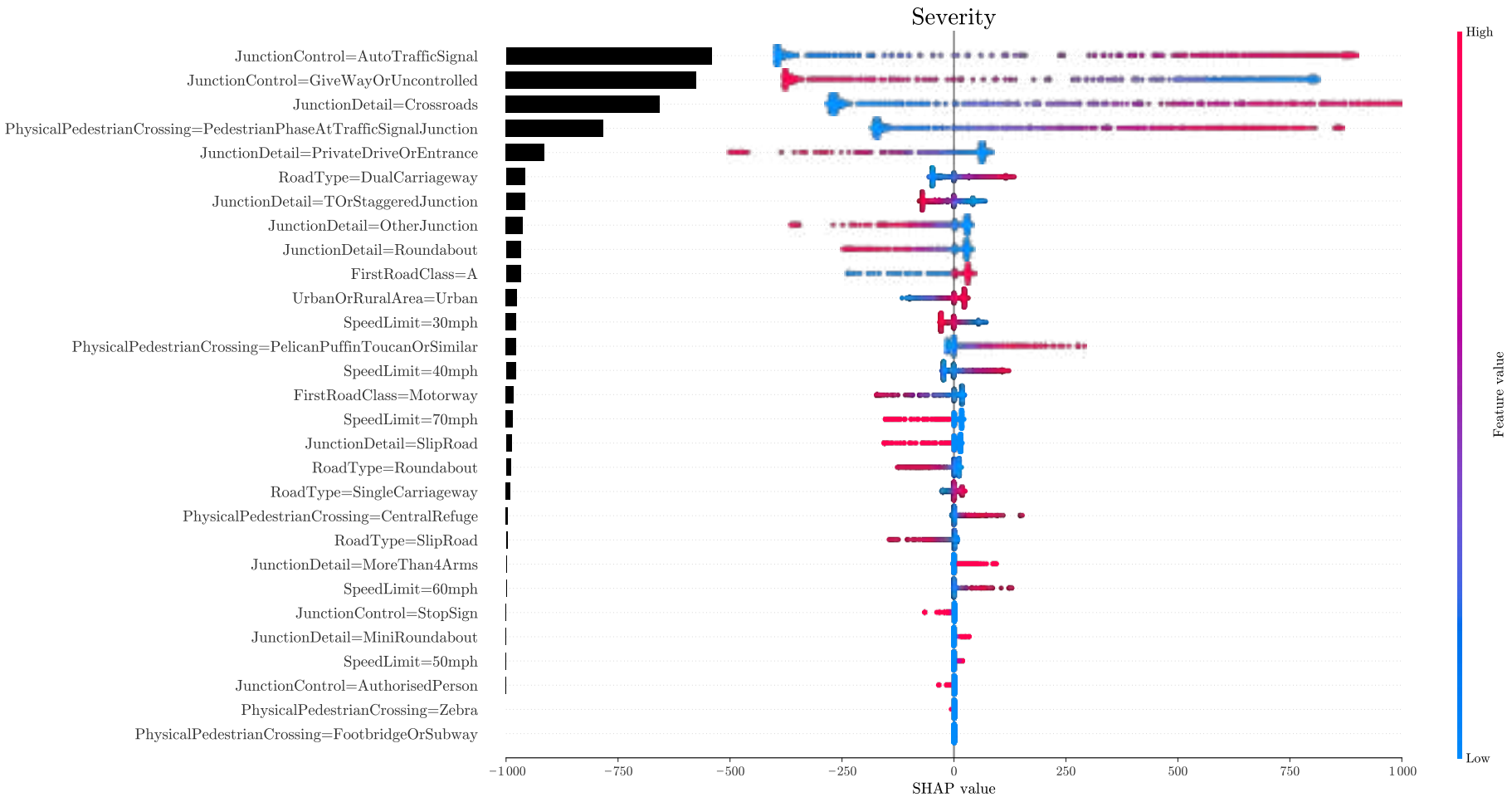


FIGURE J.10: *SHAP value results for the SVM learning model trained to predict accident severity, where each road junction observation is plotted according to each semi-spatial attribute.*

APPENDIX K

Road accident SHAP values at crossroads

The SHAP values corresponding to the OR, NB, CT, kNN, SVM and MLP learning models trained to predict accident severity of a two-vehicle RA at a crossroads during the demonstration of Sub-module 9.5 in §8.3.1 are contained in this appendix. The learning models trained to predict accident severity are shown in Figures K.1–K.6.

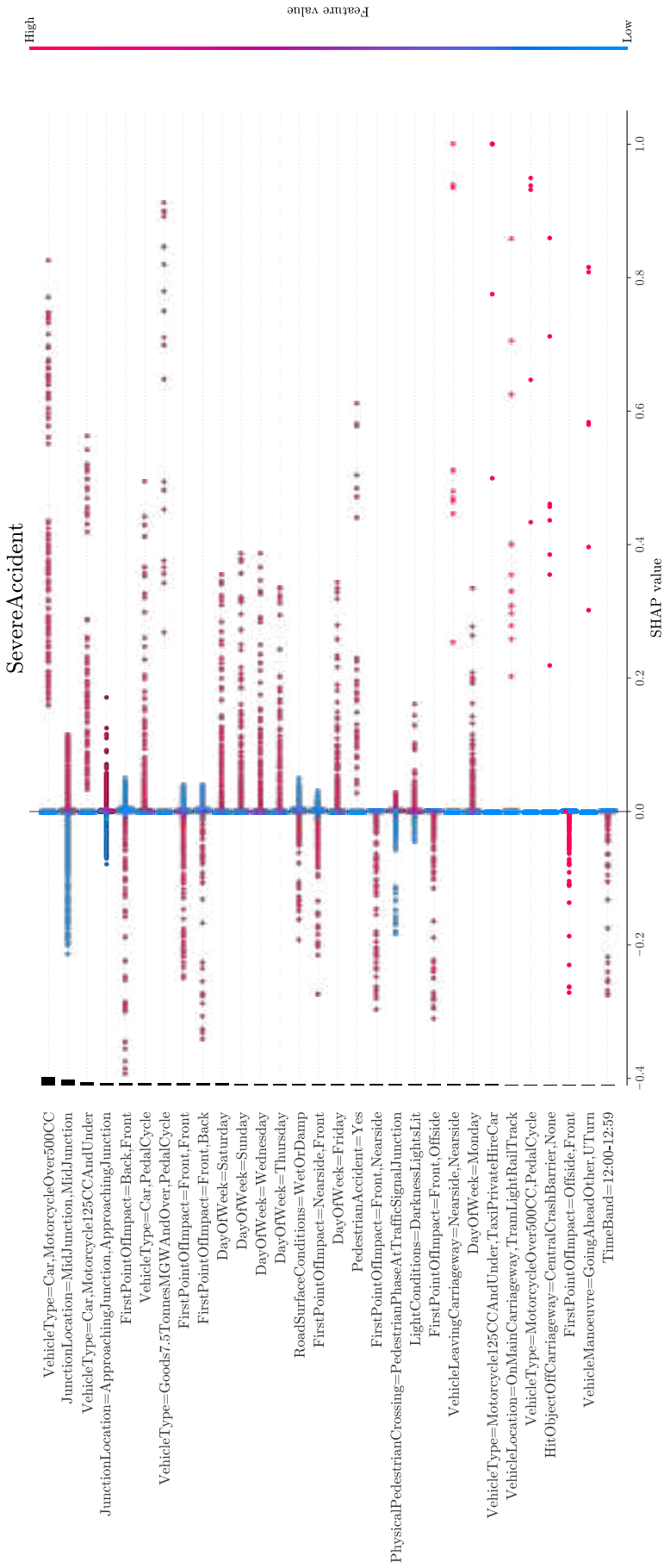


FIGURE K.1: *SHAP value results for the OR learning model trained to predict accident severity at crossroads, where each accident observation is plotted according to each non-spatial attribute.*

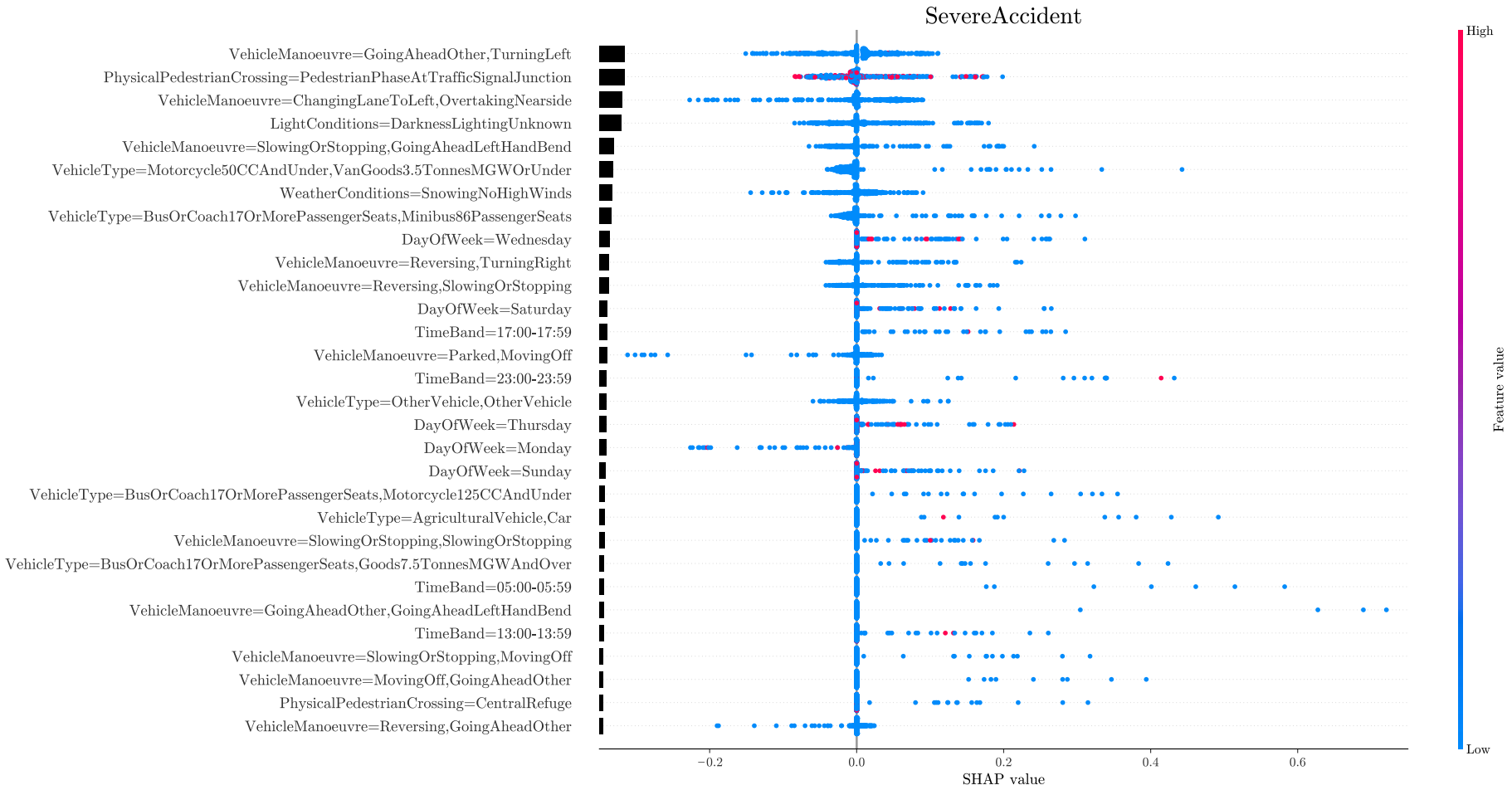


FIGURE K.2: *SHAP value results for the NB learning model trained to predict accident severity at crossroads, where each accident observation is plotted according to each non-spatial attribute.*

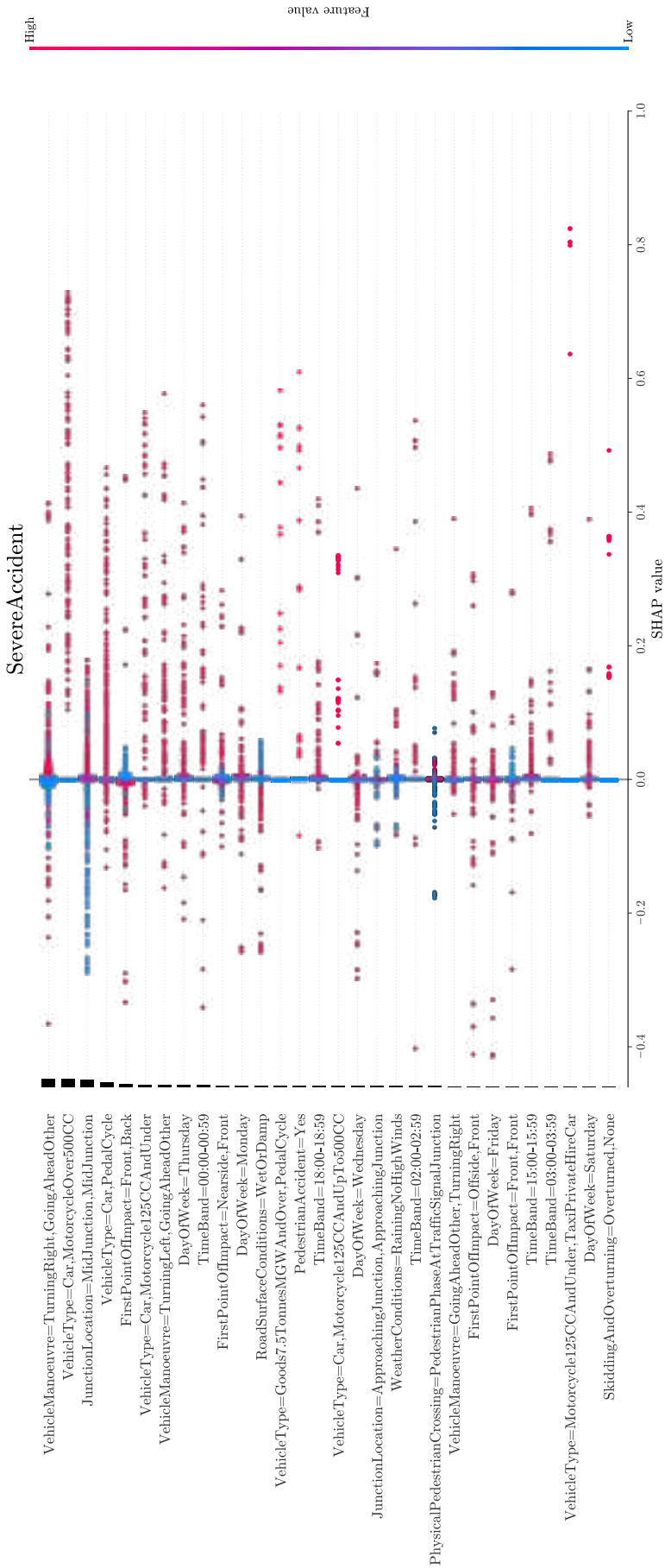


FIGURE K.3: *SHAP value results for the CT learning model trained to predict accident severity at crossroads, where each accident observation is plotted according to each non-spatial attribute.*

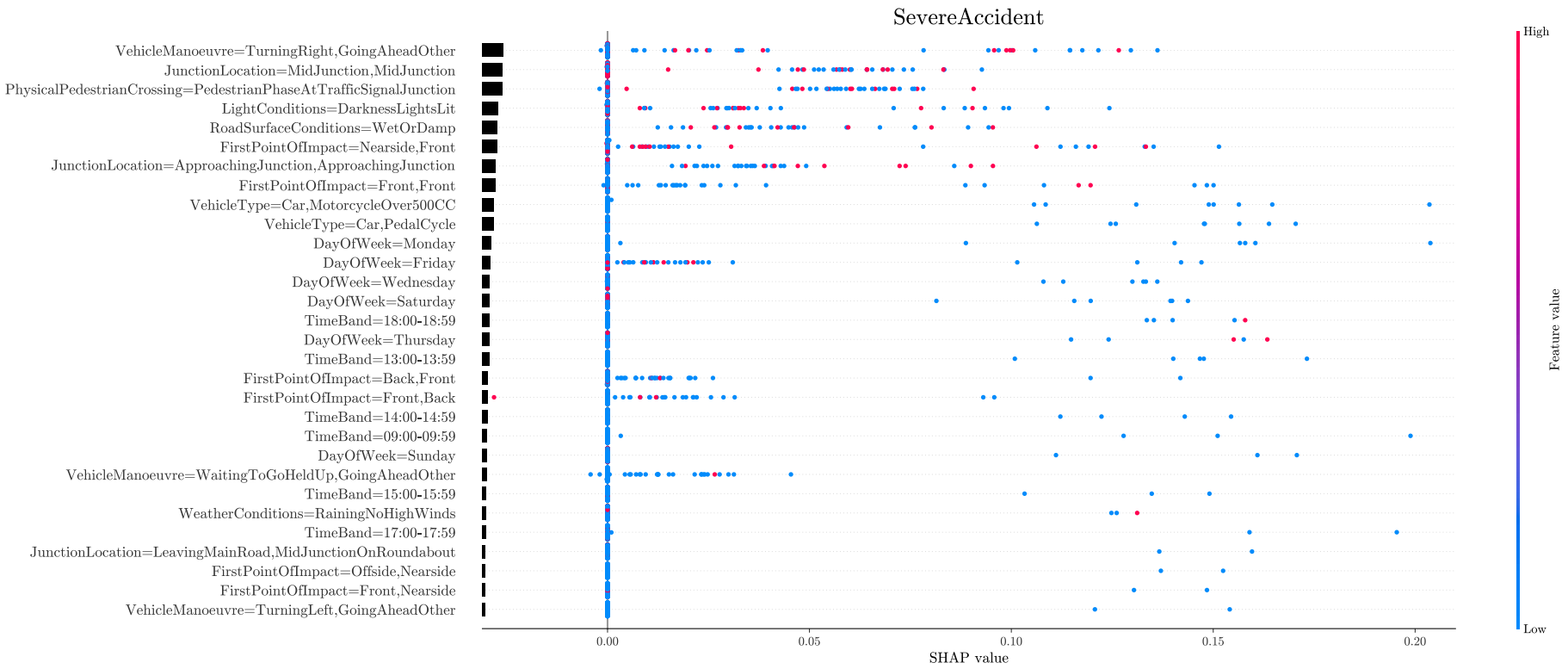


FIGURE K.4: *SHAP value results for the kNN learning model trained to predict accident severity at crossroads, where each accident observation is plotted according to each non-spatial attribute.*

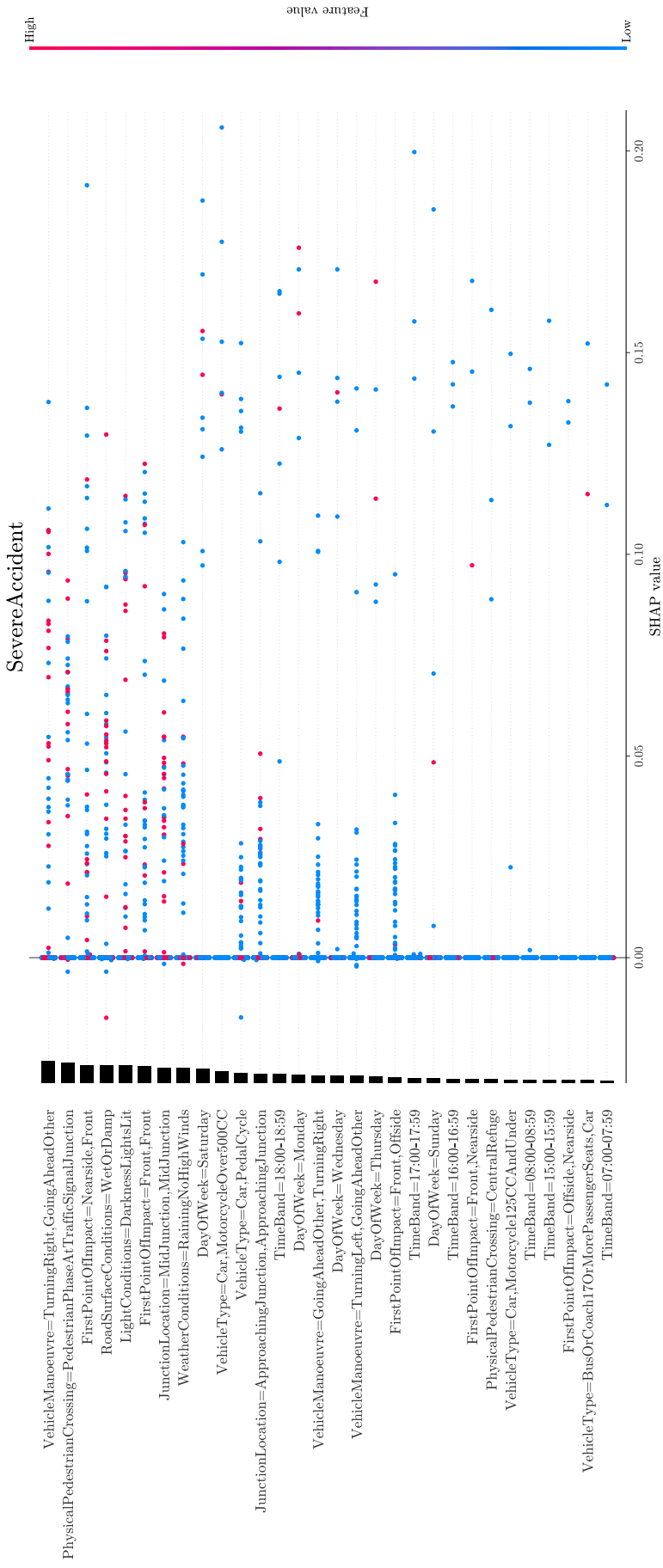


FIGURE K.5: *SHAP value results for the SVM learning model trained to predict accident severity at crossroads, where each accident observation is plotted according to each non-spatial attribute.*

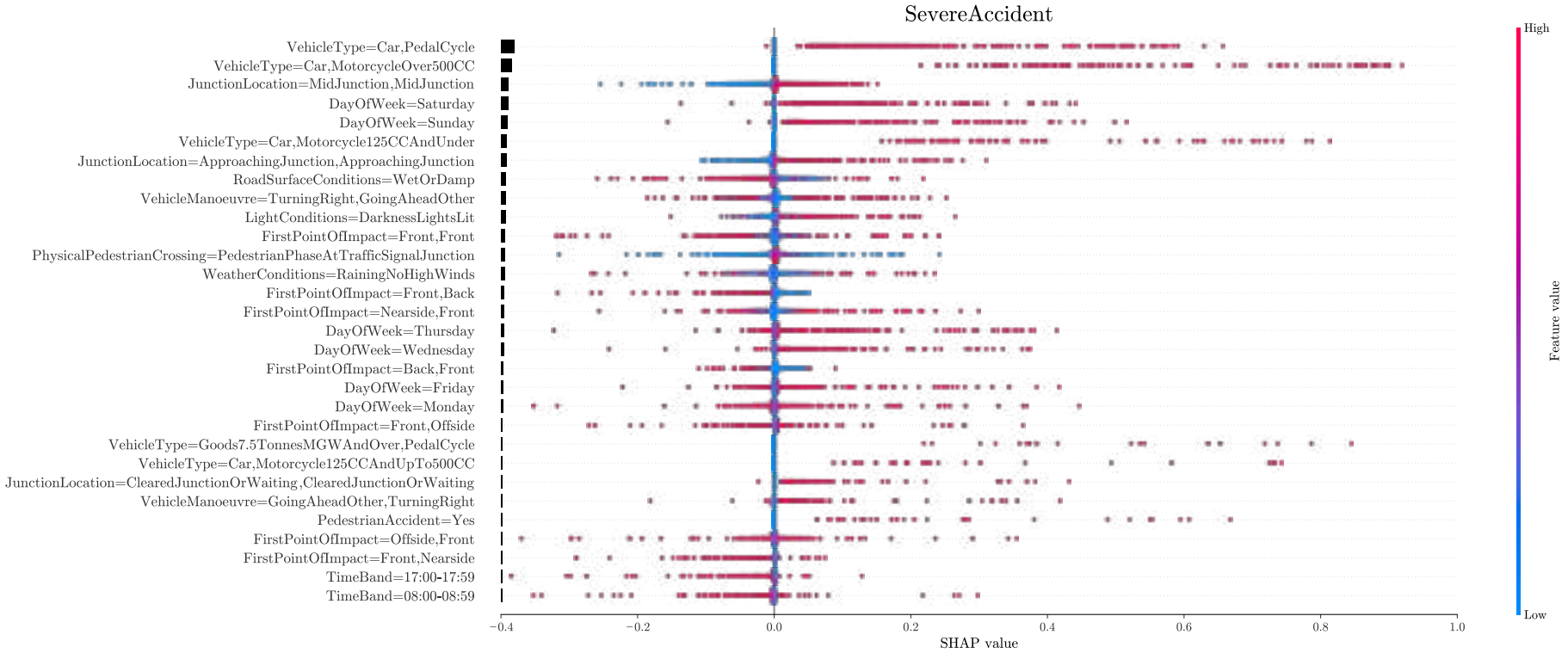


FIGURE K.6: *SHAP value results for the MLP learning model trained to predict accident severity at crossroads, where each accident observation is plotted according to each non-spatial attribute.*

APPENDIX L

Road accident SHAP values on motorways

The SHAP values corresponding to the OR, NB, CT, kNN, SVM and MLP learning models trained to predict accident severity of a one-vehicle RA on a motorway during the demonstration of Sub-module 9.5 in §8.3.2 are contained in this appendix. The learning models trained to predict accident severity are shown in Figures L.1–L.6.

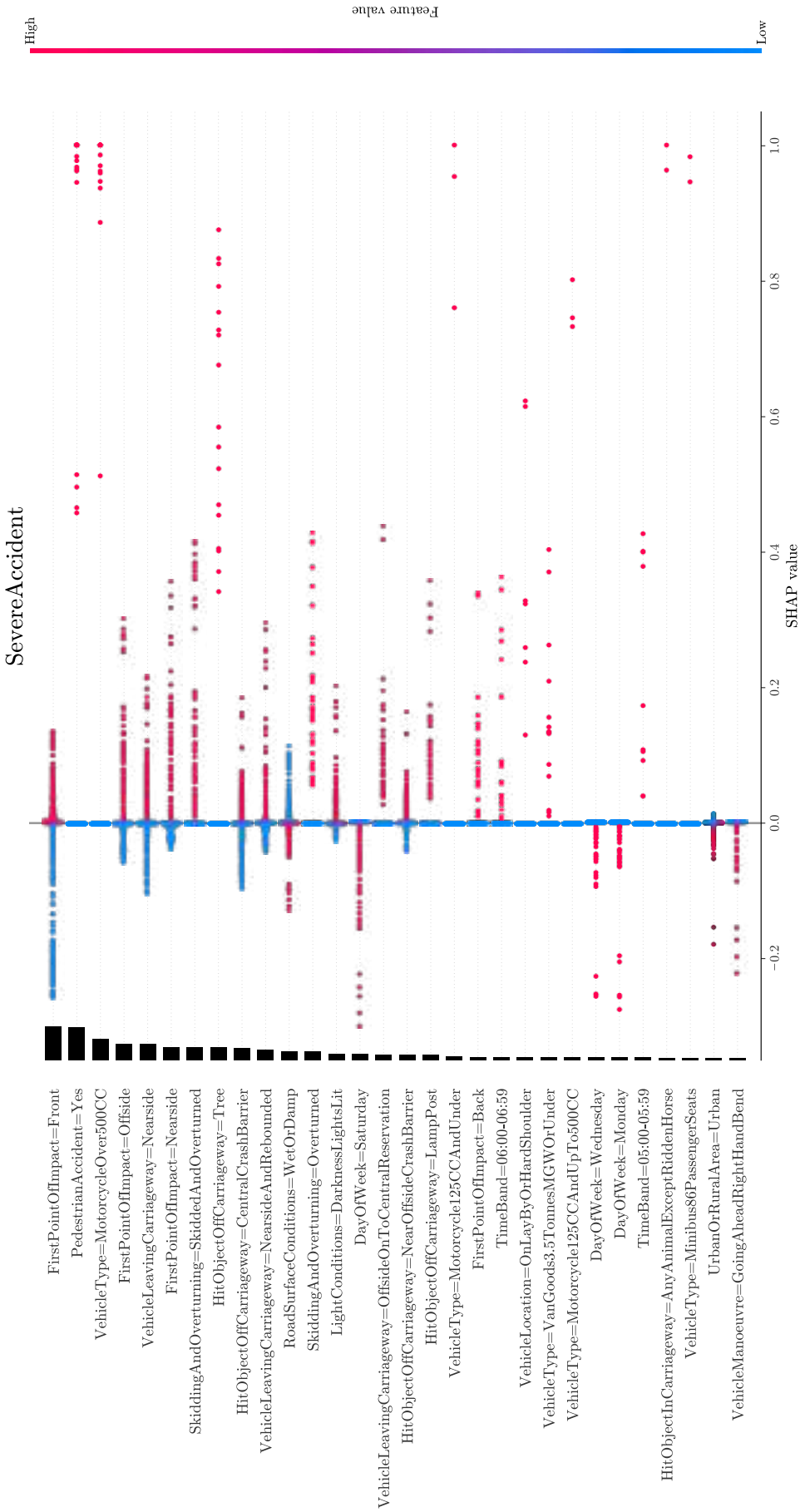


FIGURE L.1: *SHAP value results for the OR learning model trained to predict accident severity on motorways, where each accident observation is plotted according to each non-spatial attribute.*

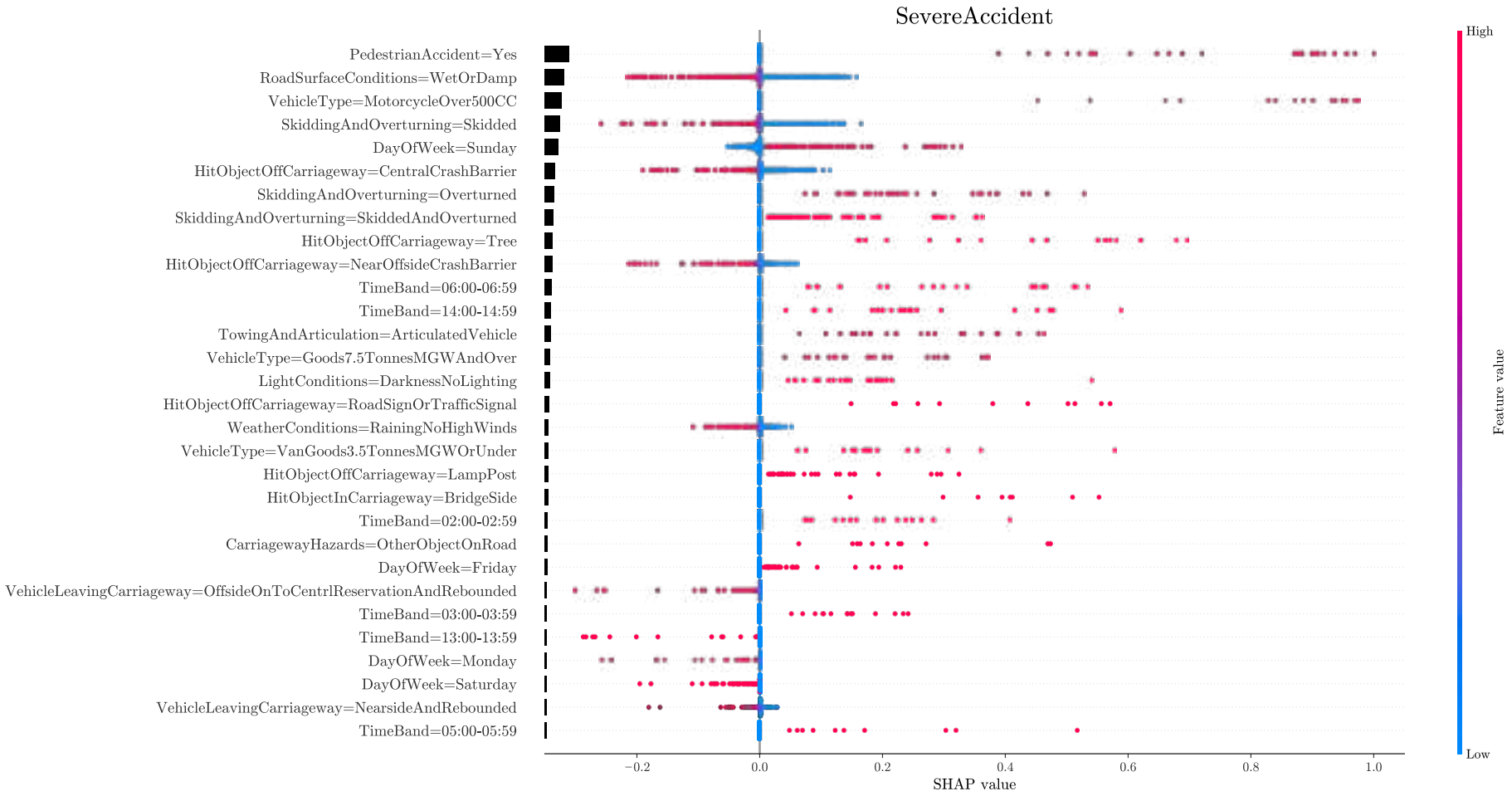


FIGURE L.2: *SHAP value results for the NB learning model trained to predict accident severity on motorways, where each accident observation is plotted according to each non-spatial attribute.*

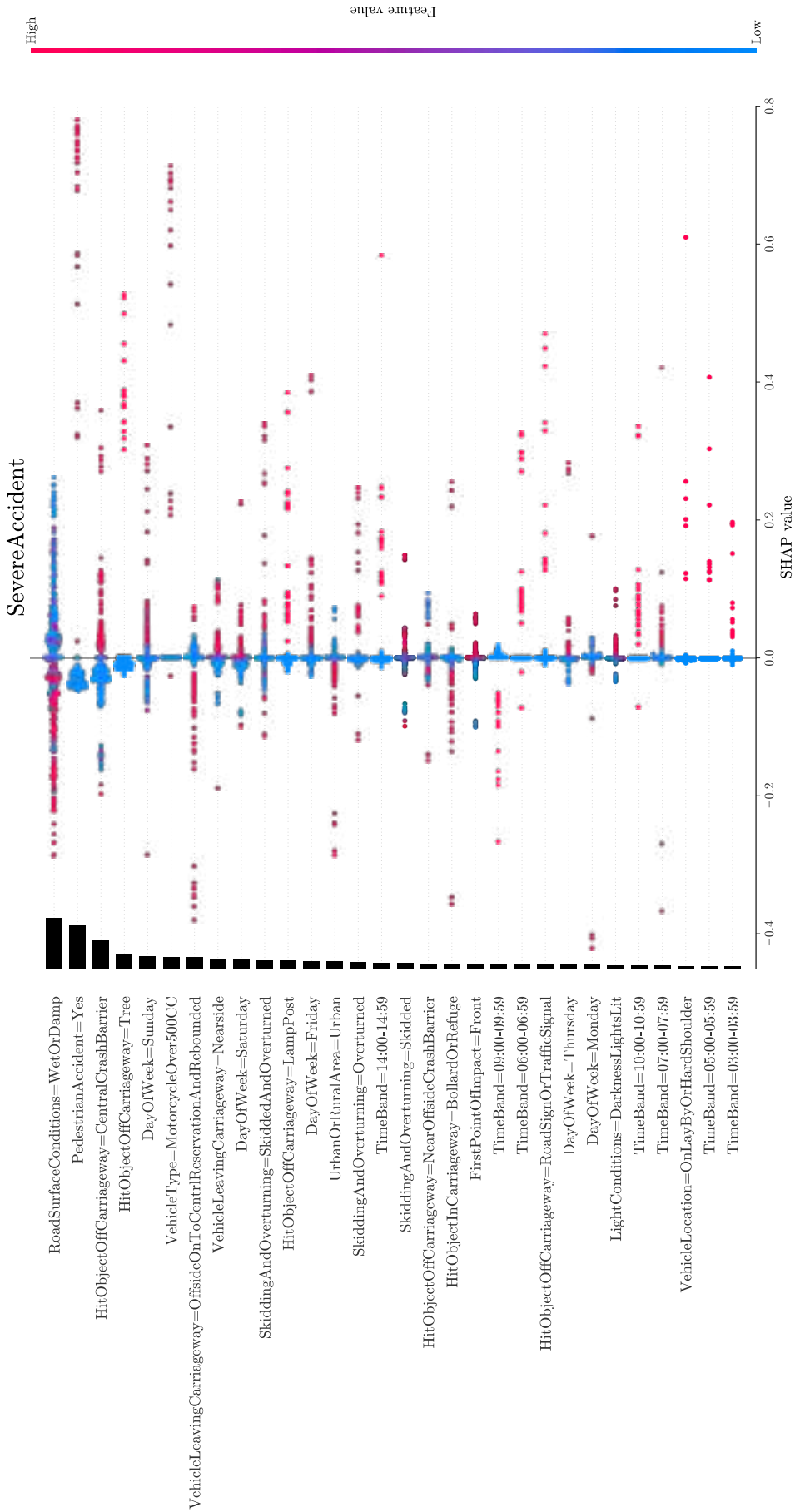


FIGURE L.3: *SHAP value results for the CT learning model trained to predict accident severity on motorways, where each accident observation is plotted according to each non-spatial attribute.*

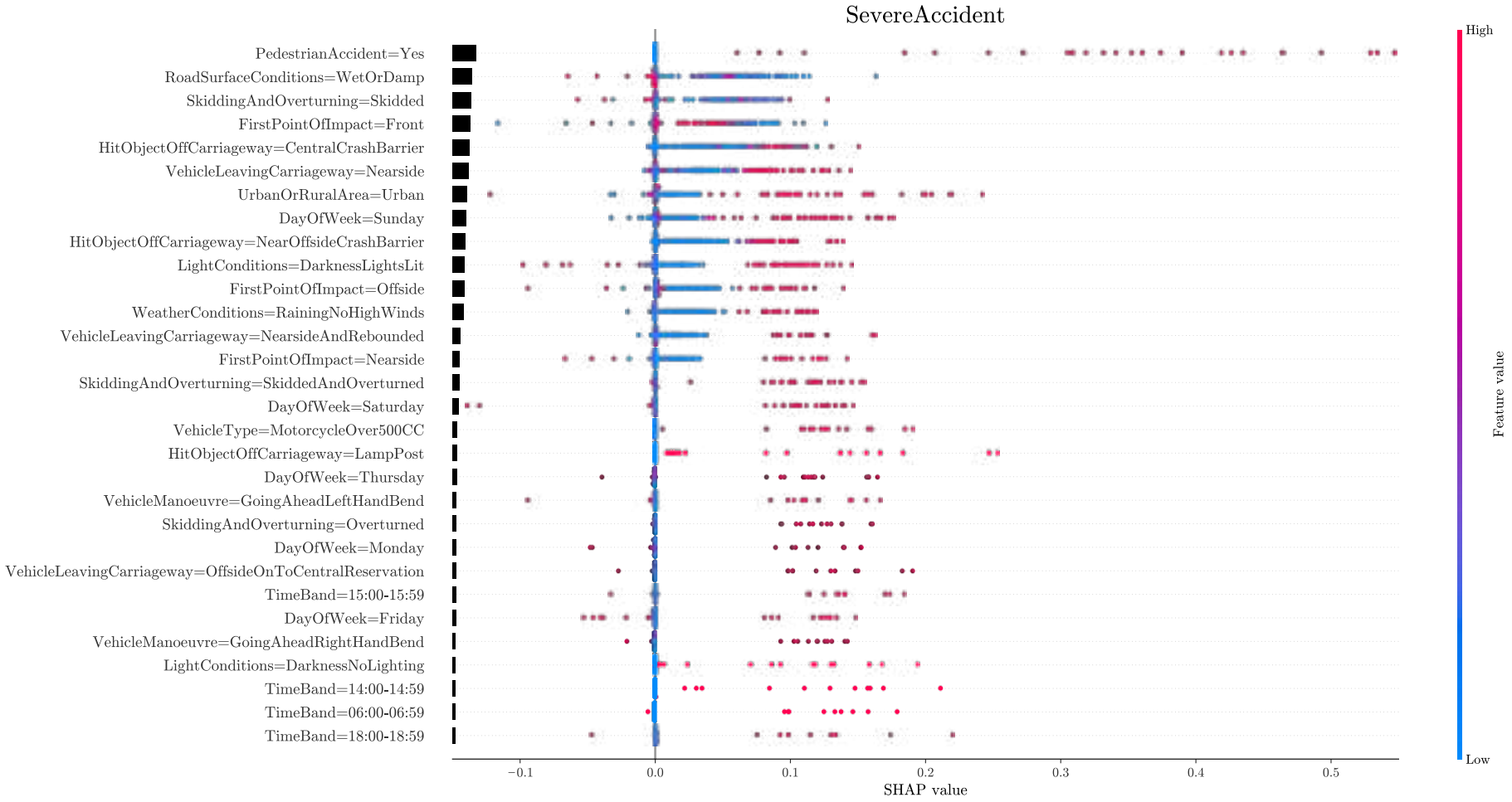


FIGURE L.4: *SHAP value results for the kNN learning model trained to predict accident severity on motorways, where each accident observation is plotted according to each non-spatial attribute.*

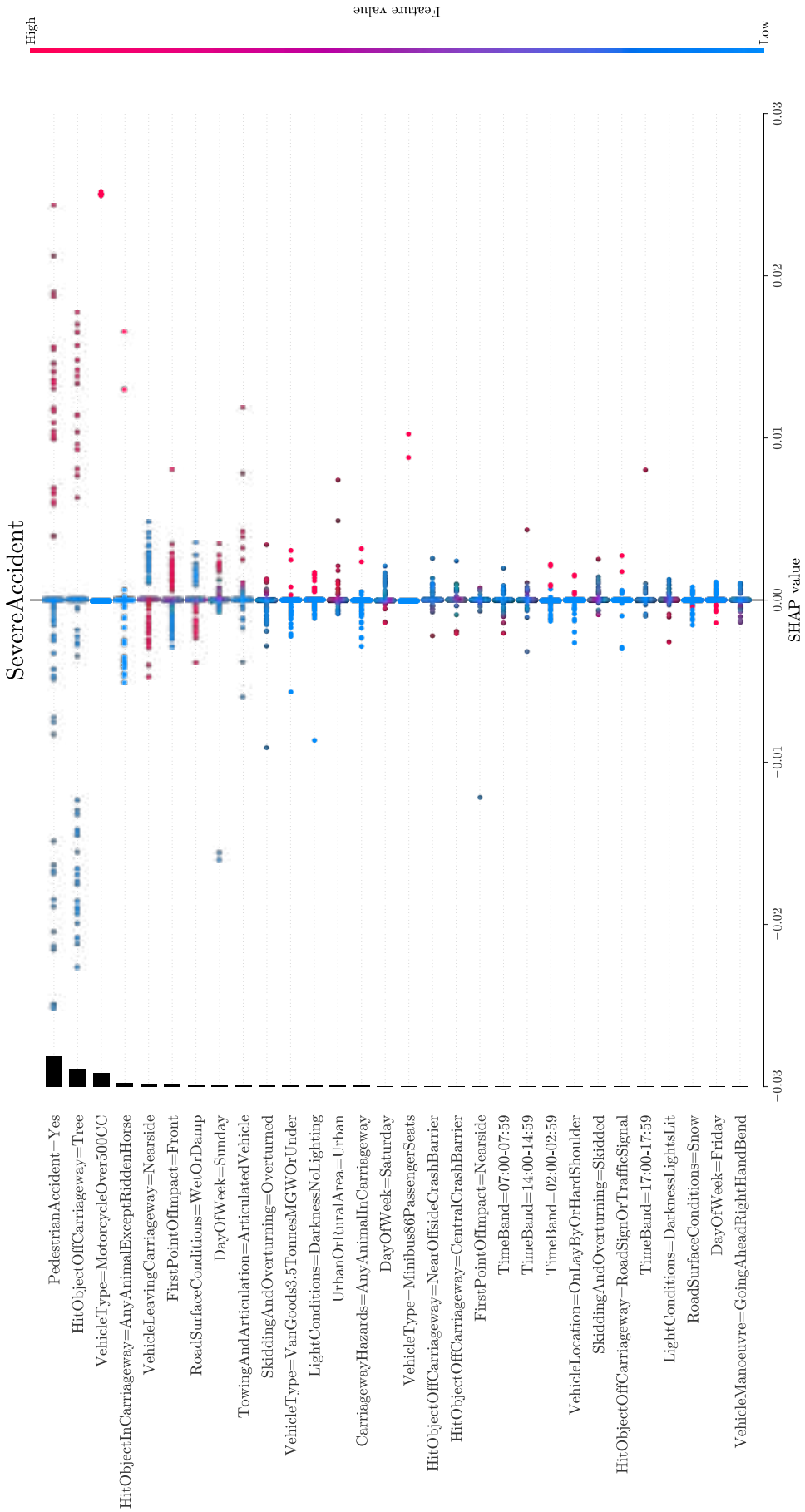


FIGURE L.5: *SHAP value results for the SVM learning model trained to predict accident severity on motorways, where each accident observation is plotted according to each non-spatial attribute.*

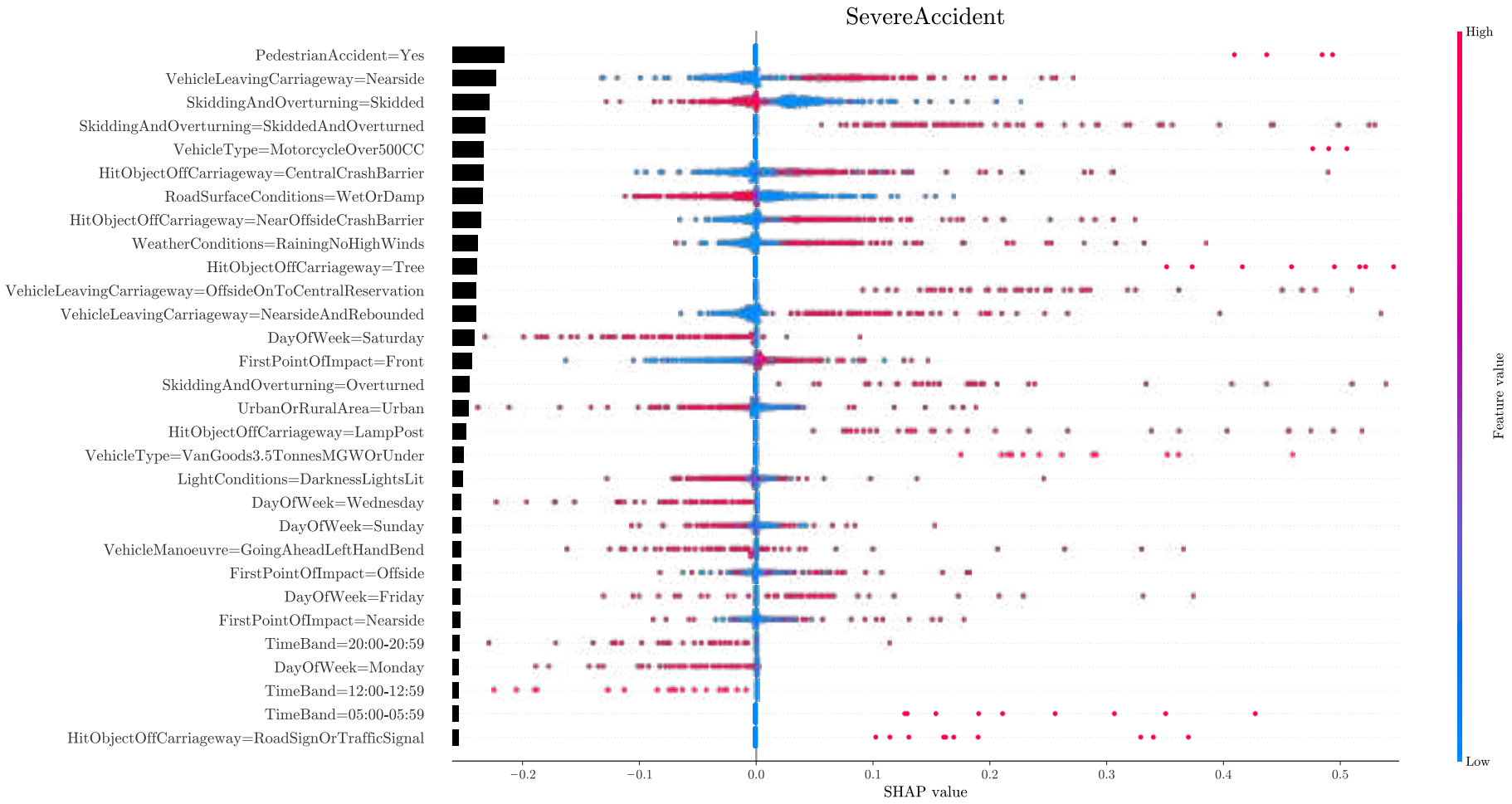


FIGURE L.6: *SHAP value results for the MLP learning model trained to predict accident severity on motorways, where each accident observation is plotted according to each non-spatial attribute.*