Bayesian Network Applications for
Traffic Prediction and Traffic Loads in Bridges
in the Netherlands

Master Thesis

Liuxin Si
Bayesian Network Applications for 
Traffic Prediction and Traffic Loads in Bridges 
in the Netherlands

THESIS

submitted in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

in

APPLIED MATHEMATICS

by

Liuxin Si

born in Zhejiang, China
Bayesian Network Applications for Traffic Prediction and Traffic Loads in Bridges in the Netherlands

Author: Liuxin Si  
Student NO.: 4118731  
Email: liuxinsi@gmail.com  
Study: Risk and Environmental Modelling  
University: Delft University of Technology  
Date: August, 2012

Abstract

Traffic loads is an important factor in assessing the reliability of infrastructures. The gravity causes the pressure to the bridge deck when traffic pass by, which further leads to a hidden danger to bridge safety. Besides the prediction of traffic situation is valuable not only for bridge safety valuation, but also for travelers, since it can help drivers to arrange the route and road managers to execute traffic controls. Therefore, an accurate model for measuring and prediction is necessary.

In this thesis, Continuous Non-parametric Bayesian Networks (CNBNs) are employed to investigate the dependence relations and the predictability of target variables. Two case studies are conducted on the basis of simulated data for “Traffic Loads in Bridges” project. The BN structure is learned from data. Three copula testing methods are applied to evaluate the Gaussian copula assumption. The BN model with mixed Normal margins is able to predict better than the model with Normal margins when conditionalizing on out of sample values. We propose three models for traffic prediction purpose. The Naive model turns out to predict the best with prediction time horizon below 20 minutes. When the prediction time horizon falls in [20, 40] minutes, the BN model presents an approximate accuracy of 90%-95%. The Historical model is a good choice for longer prediction time horizon prediction.

The results of traffic loads project can be used for bridge safety assessment. And the traffic speed prediction can provide information to travel time prediction for dynamic routing.

Thesis Committee:

Chair: Dr. Dorota Kurowicka, TU Delft  
Committee Member/Daily Supervisor: Dr. Oswaldo Morales Napoles, TNO, TU Delft  
Committee Member: Dr. Ghada el Serafy, TU Delft  
Committee Member: Dr. Raphael Steenbergen, TNO  
Committee Member: Dr. Taoufik Bakri, TNO
I would like to thank Dorota Kurowicka, for her guidance and support during these two years in Delft, and also her invaluable insights for this thesis project. I wish to express my gratitude to my daily supervisor Oswaldo Morales Napoles. Without his continued support, constructive criticism and consecutive encourage during my graduation project, the work will not be accomplished so smoothly. I truly enjoyed our weekly inspiring discussions.

I also want to acknowledge TNO, and especially Raphael Steenbergen and Taoufik Bakri not only for sharing their expertise, but more important, also for their professional insights and recommendations from engineers’ point of view.

In addition, I would like to thank Ghada el Serafy for being my thesis committee member. Also a special thank goes out to A.M. Hanea from Risk and Environmental Modeling of TU Delft, for sharing her knowledge and her help in answering my questions during these two years.

Last, but not least, I specially would like to thank my husband Kun Wu, my parents my sister, and my friends, for their endless support during my graduation project.

Liuxin Si
Delft, the Netherlands
August 5, 2012
Table of Contents

Table of Contents  
i

List of Figures  
v

List of Tables  
x

Chapter I: INTRODUCTION  
1
  1.1 Background  ................................................................. 1
  1.2 Methodologies ............................................................. 3
  1.3 Aim of the Projects ....................................................... 4
  1.4 Supports and Contributions ............................................ 5
  1.5 Organization of Report ................................................. 6

Chapter II: NON-PARAMETRIC BAYESIAN NETWORKS  
8
  2.1 Bayesian Networks ..................................................... 8
     2.1.1 Probabilistic background ......................................... 9
     2.1.2 Directed acyclic graph and d-separation ........................ 14
     2.1.3 Bayesian network basics ......................................... 16
  2.2 Copula and Gaussian copula ............................................ 22
     2.2.1 Copula ................................................................. 22
4.1 Problem Statement ............................................. 80
4.2 Traffic Modelling ............................................. 81
  4.2.1 Traffic theory ........................................... 81
  4.2.2 Early works ............................................ 85
  4.2.3 Problem statement ..................................... 90
4.3 The Data ....................................................... 92
  4.3.1 Data acquisition ....................................... 92
  4.3.2 Data processing ....................................... 94
  4.3.3 Preliminary Analysis ................................... 97
4.4 Model Preparation ......................................... 102
  4.4.1 Nodes selection ...................................... 102
  4.4.2 Model construction and validation .................... 106
4.5 Implementation with Filtered Data ....................... 110
  4.5.1 Performance measures ............................... 112
  4.5.2 Prediction at one location ........................... 115
  4.5.3 Prediction over the network ......................... 124
4.6 Implementation with Raw Data .......................... 133

Chapter V: CONCLUSIONS AND FUTURE WORK ............... 138
5.1 Conclusions ............................................... 138
  5.1.1 Theoretical part .................................... 138
  5.1.2 Project 1: Traffic Loads in Bridges ................. 138
  5.1.3 Project 2: Traffic Information Modelling ............ 139
5.2 Future Work ................................................. 140
List of Figures

2.1 A directed un-acyclic graph on 3 variables .......................... 14
2.2 Serial connection ......................................................... 14
2.3 Diverging connection .................................................... 15
2.4 Converging connection .................................................. 15
2.5 A DAG with $X_2$ and $X_{10}$ instantiated ............................ 16
2.6 A BN on 6 discrete variables. The probabilities to specify are
   $P(X_1), P(X_2), P(X_3|X_1, X_2), P(X_4|X_3), P(X_6), P(X_5|X_3, X_4, X_6)$ .................. 17
2.7 A simple discrete BN on 3 variables. ................................. 19
2.8 Conditional probability table specified in Netica. .................... 20
2.9 Result from Netica, which is the same as the result from calculation. ... 20
2.10 Scatter plots of Gaussian copula with different rank correlation ....... 26
3.1 A visual bridge model .................................................. 38
3.2 A maximal bending moment example on a Dutch bridge ............ 39
3.3 NPBN model for quantifying data generated from WIM ............... 41
3.4 A realization of $S_t$ with $r(S_t, S_{t-1}) = 0.2$ .......................... 42
3.5 A realization of the simulation ........................................ 43
3.6 Histogram of $W_f, W_s, C_f$ and $qq$ .................................. 46
3.7 Histogram of average weight of each axle ................................. 46
3.8 BN with no arcs and empirical correlation matrix of case $N = 1$ ........ 50
3.9 BN on 13 nodes with no arcs of case $N = 3$ ............................... 50
3.10 Empirical correlation matrix of case $N = 3$ ............................... 51
3.11 Saturated BN model of case $N = 1$ ........................................ 52
3.12 Saturated BN model of case $N = 3$ ........................................ 52
3.13 Validation results of saturated BN model of case $N = 1$ ............ 53
3.14 Validation results of saturated BN model of case $N = 3$ ............ 54
3.15 Semi-saturated BN model of case $N = 3$, after removing arcs whose
  absolute (conditional) rank correlation below 0.1 .......................... 55
3.16 Validation results of semi-saturated BN model of case $N = 3$ ...... 56
3.17 Normal distribution fitting to variables $W_f$ and $W_s$ .................... 59
3.18 NlogL line with respect to the number of components, for variables $W_f,$
  $W_s$, $C_f$ and $qq$ .......................................................... 61
3.19 Goodness-of-fitting comparison between normal fitting and
  mixture-normal fitting to $W_f$, by compare survival function ............ 61
3.20 Inference results implemented in UNINET, given $qq = 28(kNm)$, case $N = 1$ 63
3.21 Inference results implemented in UNINET, given $W_f = 1835(kN)$, $W_s =$
  $1440(kN)$ and $C_f = 25$, for case $N = 1$ ................................. 67
3.22 Inference results implemented in UNINET, given $qq = 28(kNm)$ for case
  $N = 3$ ........................................................................... 69
3.23 Inference results comparison between mixture-normal model and normal model, given $qq = 28(kNm)$, $N = 3$, with UNINET result as the benchmark ................................................................. 70

3.24 Inference results comparison between mixture-normal model and normal model, given $Wf2 = 1369(kN)$ and $Ws2 = 0(kN)$, with UNINET result as the benchmark ................................................................. 73

3.25 The color map of qq from different data sets shown in table 3.15 ................. 76

3.26 The percentage of qq falls in intervals. Data sets are generated with $L =$ 50$(m)$, but different auto correlation coefficients ......................................................... 77

3.27 The percentage of qq falls in intervals. Data sets are generated with $L =$ 100$(m)$, but different auto correlation coefficients ......................................................... 77

4.1 Trajectories example in time-space diagram [26] ........................................... 82

4.2 Two dimension relationship among density, intensity and speed, within Greenshield model ................................................................. 85

4.3 A demonstration of Bayesian Model from [22] - A schematic overview ........ 88

4.4 One example dynamic BNs in [23] ................................................................. 89

4.5 Schematic overview of experiment road section A10 ring, and demonstration of prediction location B ................................................................. 91

4.6 ILD example on Dutch highways [22] ............................................................. 93

4.7 Dynamic Route Information Panel ............................................................... 94

4.8 Application of the Treiber and Helbing filter: speed color map on day 5th of March 2011 at A10 ring Amsterdam ........................................................ 96

4.9 A10 ring Amsterdam overview ................................................................. 99
4.10 Mean speed plots for the total set, weekdays, weekend days, days of the week, public holidays, Non public holidays, public holiday in week days, Non public holidays in week days, in 2011 .................................................. 100

4.11 Mean plots for speed for the total set, public holidays, Non public holidays, week days, public holiday in week days, Non public holidays in week days, with high percentage of public holidays in both total set and week day set, in 2011 .................................................. 101

4.12 Schematic view of the data flows between the sensors and the data base and the applications. ................................................................. 102

4.13 Application of Traffic Flow theory at A10 ring Amsterdam .......................... 103

4.14 Speed and density product moment correlation of locations A and C, with different distances between A and C ................................................. 105

4.15 Density-Speed, Intensity-Speed plots, from 12:00 to 24:00 on 4th of March in 2011, at A10 ring 28.08 hectometers ........................................... 107

4.16 Two BN types demonstration ................................................................. 107

4.17 A demonstration: Pearson correlation vary in time .................................... 108

4.18 Model I: One time step model at time 16:00, 4th of March (Friday) in 2011, at hectometer 28.08 on A10 ring .................................................. 109

4.19 Validation results of model I ................................................................. 110

4.20 Model II: Two time steps model at time 16:00, 4th of March (Friday) in 2011, at hectometer 28.08 on A10 ring .................................................. 110

4.21 Validation results of model II ................................................................. 111
4.22 MAPE performance measures for chosen models, experimented on testing days and locations; BN model refers to BN model I .......................... 118
4.23 sMAPE performance measures for chosen models, experimented on testing days and locations; BN model refers to BN model I .......................... 119
4.24 FN performance measures for chosen models, experimented on testing days and locations; BN model refers to BN model I .......................... 120
4.25 FP performance measures for chosen models, experimented on testing days and locations; BN model refers to BN model I .......................... 121
4.26 An example of traffic information over A10 ring Amsterdam on the 3rd of March, in 2011 ................................................................. 125
4.27 Predictions over the whole network of A10 ring, on testing day 18th of July 2011, Monday. Prediction time horizon: 20 minutes ............... 129
4.28 Predictions over the whole network of A10 ring, on testing day 4th of March 2011, Friday. Prediction time horizon: 30 minutes ............... 130
4.29 Predictions over the whole network of A10 ring, on testing day 10th of November 2011, Tuesday. Prediction time horizon: 30 minutes ....... 131
4.30 Predictions over the whole network of A10 ring, on testing day 30th of September 2011, Friday. Prediction time horizon: 40 minutes ........... 132
4.31 Hectometers and speed map on 4th of March, of raw data .................. 134
4.32 Predictions over the whole network of A10 ring, on testing day 4th of March 2011, Friday. Prediction time horizon: 30 minutes ............... 136
4.33 Predictions over the whole network of A10 ring, on testing day 18th of July 2011, Monday. Prediction time horizon: 30 minutes ............... 137
List of Tables

2.1 Prior distribution of variables $X_1$ and $X_2$, example 1 ................. 19
2.2 $P(X_1,X_2,X_3)$ in example 1 ........................................ 19
2.3 CPT in example 1 ......................................................... 20
3.1 One simulation sample of axles on the bridge ................................. 44
3.2 Notations and definitions of variables involved in case $N=1$ .............. 45
3.3 Notations and definitions of variables involved in case $N=3$ .............. 47
3.4 Bivariate Gaussian copula testing results for model 2 in case $N=1$ (10000 bootstrap steps) ......................................................... 57
3.5 Bivariate Gaussian copula testing results for model 3 in case $N=3$ (10000 bootstrap steps) ......................................................... 58
3.6 Optimal number of components for variables in case $N=3$, chosen from interval [2,25] ............................................................. 62
3.7 The comparison of inference results given $qq = 28(kNm)$ for case $N=1$, among different margin fitting models ............................ 65
3.8 Inference results given $qq = 30(kNm)$ for case $N=1$ ...................... 65
3.9 Inference results given $W_f = 1835.8(kN)$, $W_s = 1440(kN)$ and $C_f = 25$, for case $N=1$ ......................................................... 66
<table>
<thead>
<tr>
<th>Table No.</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.10</td>
<td>Notations and definitions of variables involved in case $N = 3$</td>
<td>68</td>
</tr>
<tr>
<td>3.11</td>
<td>Inference results given $qq = 28(kNm)$, for case $N = 3$</td>
<td>71</td>
</tr>
<tr>
<td>3.12</td>
<td>Inference results given $qq = 30(kNm)$, for case $N = 3$</td>
<td>72</td>
</tr>
<tr>
<td>3.13</td>
<td>Inference results given $Wf_2 = 1369(kNm)$, for case $N = 3$</td>
<td>74</td>
</tr>
<tr>
<td>3.14</td>
<td>Inference results given $Wf_2 = 11369(kNm)$, $Ws_2 = 0$, for case $N = 3$</td>
<td>75</td>
</tr>
<tr>
<td>3.15</td>
<td>Sample size, maximal 'qq' and mean value of 'qq', for data sets with different length and correlation parameters $r(S_t, S_{t-1})$</td>
<td>76</td>
</tr>
<tr>
<td>3.16</td>
<td>The mean and $[0.05 \ 0.95]$ percentiles' values of 'Wf', conditionalized on observed 'qq'</td>
<td>78</td>
</tr>
<tr>
<td>3.17</td>
<td>The mean and $[0.05 \ 0.95]$ percentiles' values of 'qq', conditionalized on observed 'Wf'</td>
<td>78</td>
</tr>
<tr>
<td>4.1</td>
<td>MAPE performance measures for chosen models, experimented during afternoon rush hours on testing days and locations, with prediction time horizon 30 minutes</td>
<td>117</td>
</tr>
<tr>
<td>4.2</td>
<td>sMAPE performance measures for chosen models, experimented during afternoon rush hours on testing days and locations, with prediction time horizon 30 minutes</td>
<td>122</td>
</tr>
<tr>
<td>4.3</td>
<td>FP performance measures for chosen models, experimented during afternoon rush hours on testing days and locations, with prediction time horizon 30 minutes</td>
<td>122</td>
</tr>
<tr>
<td>4.4</td>
<td>FN performance measures for chosen models, experimented during afternoon rush hours on testing days and locations, with prediction time horizon 30 minutes</td>
<td>123</td>
</tr>
</tbody>
</table>
4.5 Error measures’ performance for chosen models, experimented on the
18th of July, 2011, with prediction time horizon 20 minutes .......... 128
4.6 Error measures’ performance for chosen models, experimented on the
4th of March, 2011, with prediction time horizon 30 minutes .......... 128
4.7 Error measures’ performance for chosen models, experimented on the
10th of November, 2011, with prediction time horizon 30 minutes ....... 133
4.8 Error measures’ performance for chosen models, experimented on the
30th of September, 2011, with prediction time horizon 40 minutes ....... 133
1.1 Background

When assessing the reliability of infrastructure, there are two main components to consider. First, it is the characteristics of the construction itself, for instance the type and quality of material used in the construction process, its design and possible errors performed during the construction process. The second component is on the side of the loads. To which type of loads are buildings, roads, dykes, dams’ bridges, and other types of infrastructures subject to? There are environmental loads, such as wind, rain or snow. And there are also those that are due to human activity. For example the use of vehicles over roads and bridges, the use of dykes and dams for water distribution, etc.

In this thesis we will investigate two problems which are interrelated to each other: traffic prediction and traffic loads in bridges. The traffic prediction problem may be looked at, and will be looked at in this thesis, independently of the traffic loads. This problem is of interest because of the importance of making it possible for people to use roads optimally in terms of time and safety. For the investigation of traffic loads in bridges it is necessary first to understand the configuration of traffic and at the same time its load composition.

Traffic loads in bridges

Over the last years a considerable increase in traffic loads and intensities has been observed in the Netherlands. Moreover it is expected that in the future this tendency will continue. The increase in traffic loads and traffic intensity places pressure in the correct assessment of the reliability of infrastructure. The purpose is to keep an appropriate balance between available economic resources and the safety of
users of infrastructure. In particular, in this thesis, we will be concerned with investigating reliability of bridges.

From the bridge reliability point of view, bending moments lead to deformation in the beams beneath bridges. The deformation brings crumpling on the upper and lower surfaces of the bridge, and might further result in a bridge collapse. The failure of a bridge could be a disaster, and the repair of such damage are hugely expensive. In general, the traffic load is the most significant variable to be considered when tolerated limit states are under investigation. Two of the variables used in bridge reliability estimates are the bending moments and the traffic configurations leading to certain bending moments. These information are helpful not only in the design of new bridges but also in the reliability assessment of existing structures.

**Traffic information prediction**

It is hardly necessary to emphasize how important the transportation is in our lives. Economically, the transportation plays an important role in impacting on development and the welfare of populations. Efficient transport systems provide economic and social opportunities and benefits that result in positive effects such as better accessibility to markets, employment and additional investments, and it also reduces economic costs. While the inefficient transportation increases costs as it brings a delay in travelling, which may result in not only opportunities missing, but also negative environmental effects like producing more CO2. Low efficiency transportation usually comes from bad road condition, low-efficient road management, traffic congestion and so on.

Traffic is known to be a dynamic interplay between traffic demand and traffic supply. An undesirable traffic situation is congestion. A traffic congestion is a situation where there is either not enough traffic supply in terms of road capacity or an excessive amount of traffic demand[23]. Among all European and North American countries, the Netherlands ranks second worst country for traffic congestion. The Netherlands has one of the most advanced freeway networks in the world, however, it still suffers from severe traffic jam at some times of the day, as the country has one of the highest population densities and many freeways are lacking sufficient capacity. While travelling there is
nothing more frustrating than being caught in a traffic congestion, not to mention the economic loss due to delay and the negative environmental influence.

The Dutch government is trying something to ease the country’s overload road networks, through reducing the number of cars by taxing and making use of the hard shoulders as an additional lane during rush hour. Some transport experts argue that road improvement projects, such as building better links connecting the main highways that crisscross the country, would be more effective at reducing congestions. All these methods have their own feasibilities but also negative effects. Taxing on cars increases people's expense and it may do little to reduce traffic, since driving will still be cheaper and more convenient than using public transport for the most part. Shoulder running has been proven to be a short-term solution because of the traffic volumes growing. Building new connections needs huge expenditure and faces space constraints. An alternative way is to predict the congestion in a large enough time horizon, thus giving travellers the possibility to reroute.

Besides travellers, the transportation companies and road managers benefit from the reliable traffic information prediction system as well [22]. Based on the accurate traffic situation prediction, the transportation companies can plan their delivers and pickups reasonably to satisfy their customers, with a respective reduction in the economic costs and time. Also, an accurate traffic prediction system would provide road managers with a whole traffic network view, thus they could be able to react to extreme cases such as potential traffic congestion in advance. Possible measures include diversing the traffic flow, better controlling road signalization to warn drivers for upcoming traffic jams.

1.2 Methodologies

In both problems, dependence between different random variables needs to be investigated. Traffic flow and traffic load are stochastic in nature and therefore a probabilistic analysis is needed. As a user friendly and popular tool for specifying high dimensional probabilistic model, Bayesian Network (BN) is employed in this research. BNs encode the probability density or mass function of a set of variables by specifying a number of conditional independence statements in the form of an acyclic directed
graph and a set of conditional probability functions. The problem at hand (as we will see later) contains continuous variables only. These types of domains are often referred to as continuous BN. In solving continuous BNs discretization is often used. However, it brings problems in the quantification process when dealing with high dimensional BNs. Additionally, exact inference becomes challenging and often unfeasible. To avoid some of the limitations, Continuous Non-parametric Bayesian Networks (CNBN) are applied.

CNBNs are initially investigated from the copula-vine approach. Up to now, quite some evidence [8][9] shows nice properties of the Gaussian copula, which makes it attractive for applications. In a CNBN, nodes are associated with arbitrary continuous invertible distribution functions and arcs with (conditional) rank correlations, which are realized by a copula with the zero independence property. Quantifying in this way separates one dimensional marginal distributions’ assessment and the dependence structure. The change of margins will not influence the dependence structure, and the vise versa. In this way, influences are represented as conditional rank correlations, and the joint Gaussian copula enables rapid sampling and conditionalization.

The Gaussian copula does not always represent the data fairly. In case of blindly using Gaussian copula assumption, Gaussian copula tests are needed, since it would bring very wrongly results if the Gaussian copula is not suitable for presenting the data dependence relations. Three common used types, Bivariate Gaussian copula test, Blanket test and D-test are applied.

1.3 Aim of the Projects

Traffic loads in bridges

There are two goals of interest in this project. The first one is the margins’ difference in predictions, and the second one is the conditionalization results of extreme values.

Since the copula approach separates the margins and dependence structure, it is possible to assign different parametric distributions to variables. Difference margins may result in different predictions. Thereby, we are aiming to find out the optimal
fitting margins for this specific project.

Another goal is to investigate the relationship between traffic configuration and the maximal bending moment. Here maximal bending moment refers to the maximal bending moment per day, and traffic configuration means the configuration of vehicles and loads. The over-dense traffic distributed on the bridge may result in a high bending moment, which is a potential danger to bridge. On the other hand, we are also interested to see the traffic configuration given certain amount of bending moment.

**Traffic information prediction**

Traffic information can be presented by traffic speed, traffic intensity or traffic density. From traffic theory, any one of these three variables can be obtained by knowing the other two. Traffic congestion is one of the most interesting and important situations in forecasting traffic information. If we are able to predict any one of the above mentioned three variables, then the traffic congestion is also predicable.

Traffic speed is a common used variable, and very easy to understand. In this project, we are aiming to predict the traffic speed with a given prediction time horizon. Predictions at single location and over a piece of road network are implemented. In order to observe accuracy of BN model predictions, Naive prediction is employed as the benchmark.

### 1.4 Supports and Contributions

This thesis is supported by Structural Reliability and Smart Mobility in TNO with abundant data, and it contributes to the project Graphical Methods for Systems Risk and Reliability (GAMES2R) within Enabling Technologies Program - models (ETP-models) program of TNO. The main objective of GAMES2R up to 2015 is to establish a generic toolbox of probabilistic graphical models for application in modelling systems risk and reliability. The Netherlands Organization for Applied Scientific Research (TNO) is an independent research organization whose expertise and research make an important contribution to the competitiveness of companies and organizations, to the economy and to the quality of society as a whole.

New technologies are now enabling traffic loads and traffic information to deal
with the requirements of the road users and designers. The TNO approach fits in with the concept of Sustainable and Safe, which was developed some ten years ago. The process and results in this thesis project helps investigating the bridge safety to prevent as much as possible the mistakes that are make by human participants.

Predicting traffic information is a field in which TNO investigates both the user needs as well as new technologies such as prediction models in the department of Intelligent Transport Systems (ITS). TNO is a market leader in traffic information and prediction systems in the Netherlands and has extensive cooperation with governmental organizations such as the Dutch Ministry of Transport, Public Works and Water Management [22]. The CNBP approach discussed in this project provides a new way to go for traffic information prediction.

The Risk and Environmental Modelling (REM) group of Delft University of Technology conducts research in physical modelling, risk and decision making under uncertainty. The researches within REM include expert judgement methods, stakeholder preference modelling, uncertainty analysis, dependence modelling and graphical models. The group is internationally known for its success in applying these techniques to multi and large scope areas, such as predicting natural disasters, tracking the movement of hazardous materials, assessing the adequacy of a flood defense system, transportation prediction and so on.

The combination of the theoretical knowledge in the risk analysis field of REM group at Delft University of Technology and the knowledge, practical experience and data of TNO is a strong basis for this thesis. And further the research in this thesis makes positive contributions to both institutions.

### 1.5 Organization of Report

This report is organized as:

Chapter I is the introduction where the background, aim of project, contributions and supports are described and explained.

Chapter II contains the necessary theoretical background that underlies the simulation and analysis presented in the latter stage of the report. The general knowledge of CNBNs and Gaussian copula testing are presented in this chapter.
In chapter III, the CNBN application study in traffic loads in bridges is accomplished. This chapter consists of 6 parts. Part 1 states the problem which needs to be solved in the later stages. Data simulation and processing are presented in part 2. Models for later using are constructed in part 3. In part 4 two Gaussian copula tests: Bivariate Gaussian copula test and Blanket test are done to check the Gaussian copula hypothesis. Part 4 presents the process of generating optimal margins to fit the variables. Inference results are shown in part 5. And an investigation into data sets generated by different parameter is done in the last part.

Chapter IV gives a brief description of traffic information prediction case study. This chapter includes 5 parts. In the first part the traffic problem of investigation is introduced. The second part introduces basics about traffic theory. Data preparation is accomplished in part 3. Part 4 prepare the model for further prediction. Model results and performance are shown in part 5.

Finally, this thesis is closed in chapter V with conclusions, recommendations and future work.
2.1 Bayesian Networks

Bayesian Networks (also known as Bayesian Belief Networks, Causal Probabilistic Networks, Causal Nets, Graphical Probability Networks and BBNs) are an emerging directed graphical modelling approach of artificial intelligence research that aims to provide a decision-support framework for problems involving uncertainty, complexity and probabilistic reasoning [1]. We will use the name Bayesian Network and the abbreviation BN. The name of Belief Networks honors Rev. Thomas Bayes (1702-1761), whose rule for updating probabilities in the light of new evidence is the foundation of the approach.

BNs are directed acyclic graphics, containing nodes and arcs. Nodes in BNs represent univariate random variables, which can be discrete and continuous\(^1\). The arcs, connecting two variables, represent direct dependency relationships (often, but not necessary, causal relationships) between the corresponding variables. The absence of an arc between two nodes means that their dependency is mediated via some other variables. With these nodes and arcs, BNs provide a compact representation of high dimensional uncertainty distributions over the set of variables [6] and encode the probability density or mass function on these variables by specifying a set of conditional independence statements in a form of an acyclic directed graph and a set of probability functions.

One of the most important features of BNs is its capability for statistical inference. Given a well structured BN, the evidence can be efficiently propagated

---

\(^1\)BNs can also contain functional nodes, i.e nodes which are functions of other nodes. The ensuing discussion refers to probabilistic nodes
throughout the network following probability calculus and Bayes theorem. Both ‘forward’ (in terms of the directionality of arcs) and ‘backward’ inferences are accepted in BNs. If the inference is done ‘backward’ (usually), the BN is used for diagnosis, whereas if it is done ‘forward’, the BN serves for prediction. The feature of inference led to the rapid application of Bayesian Networks as the method of choice for uncertain reasoning in AI and expert systems.

Often the graphical aspect of a Bayesian network is referred to as its qualitative aspect, while the probabilistic, numerical parts as its quantitative aspect. General basics about Bayesian Networks are introduced in the following sections.

2.1.1 Probabilistic background

The uncertainty calculus used in BNs is conducted on the basis of probability theory. The basic concept in the Bayesian treatment with uncertainty is that of conditional probability. A conditional probability is the probability that event \( A \) occurs when the sample space is limited to event \( B \). This is read ‘the probability of \( A \), given \( B \)’, commonly notated as \( P(A|B) \). For two jointly distributed continuous random variables \( X \) and \( Y \), the conditional density of \( X \) given \( Y \) is the probability distribution of \( X \) when \( Y \) is known to be a particular value, simply written as \( f_X(X|Y = y) \).

The independency of two events or two variables leads to:

\[
P(A|B) = P(A) \quad \Rightarrow \quad P(B|A) = P(B)
\]

or

\[
f_Y(Y|X) = f_Y(Y) \quad \Rightarrow \quad f_Y(Y|X) = f_Y(Y)
\]

The fundamental rule of probability calculus reads:

\[
f(X,Y) = f_X(X|Y)f_Y(Y) = f_Y(Y|X)f_X(X)
\]

We restrict our attention to the continuous case. For two arbitrary continuous variables \( X \) and \( Y \), the fundamental rule of probability calculus reads:

\[
f(X,Y) = f_X(X|Y)f_Y(Y) = f_Y(Y|X)f_X(X)
\]

A extension of fundamental rule to high dimension, i.e. a set of variables \((X_1, X_2, \ldots, X_n)\), generates the chain rule as:

\[
f(X_1, \ldots, X_n) = f(X_n|X_1, \ldots, X_{n-1})f(X_{n-1}|X_1, \ldots, X_{n-2}) \cdots f(X_1)
\]
Bayes’ Rule/Theorem follows if \( f_X(X) \neq 0, f_Y(Y) \neq 0 \):

\[
f_X(X|Y) = \frac{f_Y(Y|X) \cdot f_X(X)}{f_Y(Y)} \tag{2.5}
\]

\[
f_Y(Y|X) = \frac{f_X(X|Y) \cdot f_Y(Y)}{f_X(X)} \tag{2.6}
\]

The significance of Bayes’ Rule was first appreciated by the British cleric Thomas Bayes in his posthumously published masterwork, "An Essay Toward Solving a problem in the Doctrine of Chance" (Bayes 1764). It can be seen as a way of understanding how the probability that a theory is true is affected by a new piece of evidence.

"Given the number of times in which an unknown event has happened failed [... Find] the chance that the probability of its happening in a single trial lies somewhere between any two degrees of probability that can be named."

Further, we dwell a little on how Bayes’ Rule can be used and understood since it is always underlined as a core to inference in Bayesian Probability calculus. Assume that we have two (possibly, sets of) continuous variables \( X \) and \( Y \), a joint distribution \( f(X, Y) \) given in the factorized form \( f_X(X|Y)f_Y(Y) \), and that we observe \( X = x \). We would then typically want to compute \( f_Y(Y|X = x) \). Before we enter into interpretation of Bayes’ Rule, we introduce related notations first.

A **prior distribution** is an initial distribution originally obtained before any additional information is obtained.

A **posterior distribution** is a distribution that has been revised by using additional information that is later obtained.

Thus in our case, the prior distribution, \( f_Y(Y) \), expresses our initial about \( Y \), and the posterior distribution \( f_Y(Y|X = x) \), expresses our revised belief about \( Y \) in light of the observation \( X = x \). Bayes’ Rule tells us how to obtain the posterior distribution by multiplying the prior \( f_Y(Y) \) by the ratio \( \frac{f_X(X=x|Y)}{f_X(X=x)} \), where \( f_X(X=x|Y) \equiv L(Y|X = x) \) is called the **likelihood** for \( Y \) given \( X = x \). The general formula is:

\[
\text{posterior} \propto \text{prior} \times \text{likelihood}
\]

Bayes’ formula is always used to update a joint distribution on the basis of evidence, to see how the new information affects the posterior probability (as opposed
to prior probabilities). All these mechanics apply to discrete case as well.

BNs contain random variables, and the links represent direct dependence among variables. There are several statistical measures that summarize dependence (or independence) between random variables. They are commonly referred to as dependence measures or correlation measures. The Pearson product-moment correlation coefficient is a measure of the correlation between two variables X and Y, giving a value between \([-1, 1]\). It is widely used as a measure of the strength of linear dependence between two variables.

*Pearson's product-moment correlation coefficient* between two variables X and Y is defined as the covariance of the two variables divided by the product of their standard deviations:

\[
\rho_{X,Y} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}
\]  

(2.7)

where \(\mu_X\) and \(\mu_Y\) state the expectations of variables X and Y, and \(\sigma_X, \sigma_Y\) are standard deviations correspondingly.

The product moment correlation coefficient has the following mathematical properties:

- \(-1 \leq \rho_{X,Y} \leq 1\)
- Symmetric: \(\rho_{X,Y} = \rho_{Y,X}\)
- If \(\rho_{X,Y} = 1\), then \(X = a + bY\), \(a\) and \(b\) are constants, \(b > 0\).
- If \(\rho_{X,Y} = -1\), then \(X = a + bY\), \(a\) and \(b\) are constants, \(b < 0\).
- If \(\rho_{X,Y} = -1\), then X and Y are independent.

The product moment correlation is popular because it is often straight-forward to calculate. For uncertainty analysis it has somehow a few disadvantages.

- The product moment correlation is not defined if the expectations and variance of X and Y are not finite. (e.g. Cauchy distribution)
- The product moment correlation is not invariant under nonlinear strictly increasing transformations.
• The possible values of the product moment correlation depend on marginal distributions.[3]

To avoid these defects, the Spearman rank correlation was introduced by Spearman (1904).

**Definition 2.1 (Rank correlation)** The rank correlation of random variables $X$, $Y$ with cumulative distribution functions $F_X$ and $F_Y$ is

$$ r(X, Y) = \rho(F_X(X), F_Y(Y)) \quad (2.8) $$

We denote the rank correlation as $r(X, Y)$.

Besides the general form, the rank correlation can also be defined in two versions, one is population version in case of continuous variables, while the other one is named sample version.

Consider a population distributed according to two variables $X$ and $Y$ with two members $(X_1, Y_1)$ and $(X_2, Y_2)$. The *population version of Spearman rank correlation* $r$ is defined as:

$$ r = 3 \cdot \left[ P((X_1 > X_2, Y_1 > Y_2) \text{ or } (X_1 < X_2, Y_1 < Y_2)) ight. $$
$$ \left. -P((X_1 < X_2, Y_1 > Y_2) \text{ or } (X_1 > X_2, Y_1 < Y_2)) \right] \quad (2.9) $$

where $(X_1, Y_1)$ and $(X_2, Y_2)$ are independent.

The above definition of Spearman's rank correlation is appropriate only for populations for which the probabilities of $X_1 = X_2$ or $Y_1 = Y_2$ are zero.

Besides, Spearman also proposed a way to determine the sample rank correlation based on replacing measurements with their ranks. Suppose we get $N$ samples of random vector $X$ and $Y$: $(x_1, y_1), \ldots, (x_N, y_N)$, and these samples are converted to ranks $(x_1', y_1'), \ldots, (x_N', y_N')$. The sample rank correlation is computed from these:

$$ r = \frac{\sum_i (x_i' - \bar{x})(y_i' - \bar{y})}{\sqrt{\sum_i (x_i' - \bar{x})^2 \sum_i (y_i' - \bar{y})^2}} \quad (2.10) $$

where $\bar{x}$ is the mean of $x_1', \ldots, x_N'$, and $\bar{y}$ is the mean of $y_1', \ldots, y_N'$. 
From definitions there exists a close relationship between the product moment correlation and the rank correlation. The rank correlation is a product moment correlation of random variables transformed to uniform random variables. Hence we get immediately that rank correlation is symmetric and takes values from the interval \([-1, 1]\).

Rank correlation is invariant under non-linear strictly increasing transformations. In contrast to the product moment correlation, the rank correlation always exists and is independent of marginal distributions.

A partial correlation can be defined in a way similar to product moment correlation in terms of partial regression coefficients. Consider variables \(X_i\) with zero means and standard deviations \(\sigma_i, i = 1, \cdots, n\). Let the numbers \(b_{12;3,\cdots,n}, \cdots, b_{1n;2,\cdots,n-1}\) minimize

\[
E((X_1 - b_{12;3,\cdots,n}X_2 - \cdots - b_{1n;2,\cdots,n-1}X_n)^2)
\]

Then the partial correlation is defined as:

\[
\rho_{12;3,\cdots,n} = \text{sgn}(b_{12;3,\cdots,n})\sqrt{b_{12;3,\cdots,n}b_{21;2,\cdots,n}}
\]

Equivalently we could define the partial correlation as:

\[
\rho_{12;3,\cdots,n} = -\frac{C_{12}}{\sqrt{C_{11}C_{22}}}
\]

(2.11)

where \(C_{ij}\) denotes the \((i,j)\)th cofactor of the correlation matrix; that is, the determinant of the sub-matrix after removing row \(i\) and column \(j\).

Partial correlations can also be computed from correlations with the following recursive formula (Yule and Kendall (1965)):

\[
\rho_{12;3,\cdots,n} = \frac{\rho_{12;3,\cdots,n-1} - \rho_{1n;3,\cdots,n-1} \cdot \rho_{2n;3,\cdots,n-1}}{\sqrt{1 - \rho_{1n;3,\cdots,n-1}^2} \sqrt{1 - \rho_{2n;3,\cdots,n-1}^2}}
\]

(2.13)

As mentioned at the beginning of this chapter, BNs consists of quantitative part and qualitative aspect. The qualitative aspect of BNs is named as directed acyclic graphs (DAGs), stem from causal networks. We get into details about DAGs in the following section.
2.1.2 Directed acyclic graphs and d-separation

In mathematics, a DAG is a directed acyclic graph (also name as causal network) with no directed cycles. It is formed by a collection of variables and directed arcs. Each arc connects one variable to another, such that there is no way to start at some variable \( X_i \) and follow a sequence of arcs that eventually loops back to \( X_i \) again.[4] Figure 2.1 shows an example of directed graph, but not acyclic. We use the wording of family to represent the relationship between two variables, if there is an arc from \( X_1 \) to \( X_2 \), we say that \( X_1 \) is a parent of \( X_2 \) and, \( X_2 \) is a child of \( X_1 \).

![Figure 2.1](image)

**Figure 2.1:** A directed un-acyclic graph on 3 variables

DAGs can be used for reasoning or influence. We present a set of rules of reasoning corresponding to different type of connections. DAGs consist of serial connections, diverging connections, converging connections and their combinations. We define a variable instantiated when the state of this variable is known. Consider the situation in figure 2.2, variable \( X_1 \) has an influence on \( X_2 \), which in turn has an influence on \( X_3 \), and vice versa. However, if the state of \( X_2 \) is known, then the channel is blocked, and \( X_1 \) and \( X_3 \) become d-separated. Details about d-separation will be discussed later in this section. The evidence may be transmitted through a serial connection unless the state of a variable in its path is known. Figure 2.3 represents a diverging connection. The children are connected through parent \( X_1 \). If \( X_1 \) is instantiated, then the influence between its children is blocked. That is, children \( X_2, X_3, \ldots, X_n \) are d-separated given \( X_1 \).

![Figure 2.2](image)

**Figure 2.2:** Serial connection

The converging connection situation, as described in figure 2.4, brings a opposite
Parents $X_2, X_3, \ldots, X_n$ are d-separated if their child $X_1$ is unknown. That is, if $X_1$ is unknown, evidence about one of its parents cannot influence the certainties of the others through $X_1$. However, if $X_1$ is instantiated, knowledge of one cause may tell us something about the other causes. Simply saying, in converging connections, parents are independent given the child is uncertain, but become dependent when the evidence is learned for child (Finn V. Jensen et al. 2007).

The rules of evidence transmission makes it possible to decide the independence or dependence property of pair variables in DAGs given the evidence entered into the network. The rules are formulated in the following definition.

**Definition 2.2 d-separation** (Finn V. Jensen et al. 2007). Two distinct variables $X$ and $Y$ in a DAG are d-separated (‘d’ for ‘directed graph’) if for all paths between $X$ and $Y$, there is an intermediate variable $Z$ (distinct from $X$ and $Y$) such that either

- the connection is serial or diverging and $Z$ is instantiated
- the connection is converging, and neither $Z$ nor any of $Z$’s descendants have received evidence.

If $X$ and $Y$ are not d-separated, we call them d-connected. For d-separated case, $X$ has no influence of $Y$. When it comes to d-connected, however, changes in the belief
in $X$ will not necessarily change the belief in $Y$.

Figure 2.5 presents an example of large DAG. The dotted lines indicate insertion of evidence, that is, $X_2$ and $X_{10}$ are instantiated. The evidence entered at $X_1$ can be transmitted to $X_5$, but it cannot pass through $X_2$ to $X_6$ because this channel is blocked by instantiated $X_2$. However, the evidence may be transmitted to $X_6$ since the child of $X_8$ has received evidence. Further, it may pass through $X_3$, $X_7$ and $X_9$ to $X_{11}$ at the end. So, the path $X_1 - X_5 - X_8 - X_6 - X_3 - X_7 - X_9 - X_{11}$ is a d-connecting path, and pair variables $X_1$ and $X_{11}$ are d-connected.

**Figure 2.5:** A DAG with $X_2$ and $X_{10}$ instantiated

d-separation can be extended to sets of variables. Take the network in figure 2.5 as an example again. Note that $X_2$ and $X_{10}$ were instantiated. If $X_6$ is known as well, then the variable set \{$X_1, X_5, X_8$\} is d-separated from the variable set \{$X_3, X_4, X_7, X_9, X_{11}$\}, that is, there is no path to transmit the evidence between the node from the former set to the node from the latter set. In Bayesian networks, we state this scenario as these two sets are conditionally independent given \{ $X_2, X_6, X_{10}$\} are known.

### 2.1.3 Bayesian network basics

DAGs present a dependence structure of variables. Together with the quantitative aspect, namely their strength, DAGs lead to Bayesian networks.

**Definition 2.3** Bayesian networks (Finn V. Jensen et al. 2007). A Bayesian network
consists of the following:

- A set of variables represented by nodes and a set of directed arcs between nodes
- The variables together with the directed arcs form an acyclic directed graph
- To each variable $X_i$ with parents $X_j, \ldots, X_n$, a conditional probability table or an conditional probability function is attached

Note that if $X_i$ has no parent, then the conditional table or probability function reduces to unconditional ones.

The definition of BNs does not refer to causality, and there is no requirement that the links represent causal impact. However, one needs to check the d-separation properties and ensure the conditional independence to be hold in real world.

![Figure 2.6: A BN on 6 discrete variables. The probabilities to specify are $P(X_1), P(X_2), P(X_3|X_1, X_2), P(X_4|X_3), P(X_6), P(X_5|X_3, X_4, X_6)$](image)

In the BNs’ construction process, first of all, a list of the relevant variables is made by starting with the objectives of the analysis, and then describing the primary factors that might influence these objectives. Further, the secondary relevant factors are described and, then go on until all variables have been found. These chance nodes, and the arcs drawn between nodes representing qualitative influences, must be quantified by model builder. That is, the prior probability of each variable and the strength of the arcs need to be specified. Sometimes the strength of the arcs refer to dependence measures such as (conditional) rank correlation. And sometimes it means the conditional probability table.

In general, a BN is a Directed Acyclic Graph, whose nodes represent univariate random variables and arcs represent direct qualitative dependence relationships.
between variables represented by adjacent nodes. The absence of arcs guarantees a set of (conditional) independence facts. All these features embedded in BNs can determine a quantified joint probability distribution of variables set \( \phi = \{X_1, \ldots, X_n\} \). The quantitative information can be retrieved from data, or from expert judgement by quantifying the random variables and dependence structure when data is not available. Together with the qualitative and quantitative aspects of BNs provide a compact representation of high dimensional uncertainty distribution over variable set \( (X_1, X_2, \ldots, X_n) \) [6]. We have indicated that a BN is built by first determining a graphical model, and then specifying the quantitative part. By doing this, we have the basic background to deduce the joint probability involving all variables through the chain rule.

**Proposition 2.1 (The chain rule for BNs)** Let BN be a BN over \( \phi = \{X_1, \ldots, X_n\} \). Then BN specifies a unique joint probability distribution \( f(\phi) \) given by the product of all conditional probability specified in BN:

\[
f(\phi) = \prod_{i=1}^{n} f(X_i|\text{pa}(X_i))
\]

where \( \text{pa}(X_i) \) is the parents set of variable \( X_i \), and \( f(X_i|\text{pa}(X_i)) \) is the conditional probability of \( X_i \) given its parents \( \text{pa}(X_i) \) known.

BNs can be categorized into discrete BNs, continuous BNs and hybrid BNs based on nodes property. As can be seen directly from the name, the nodes in a discrete BN are discrete variables, continuous BNs include continuous variables only and hybrid BNs include both.

In discrete BNs, the quantitative part are specified by conditional probability tables (CPT). We present a discrete BN to show how the joint probability can be generated following the chain rule.

**Example:** Application of Chain Rule in a simple discrete BN.

Assume all three variables \( X_i, i = 1, 2, 3 \) take only two values ‘true’ and ‘false’ and the prior distribution of \( X_i, i = 1, 2, 3 \) are:

And the CPT, here is just the conditional distribution of \( X_3 \) given \( X_1 \) and \( X_2 \),
Figure 2.7: A simple discrete BN on 3 variables.

<table>
<thead>
<tr>
<th>Variables</th>
<th>True</th>
<th>False</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_1$</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>$X_2$</td>
<td>0.3</td>
<td>0.7</td>
</tr>
<tr>
<td>$X_3$</td>
<td>?</td>
<td>?</td>
</tr>
</tbody>
</table>

Table 2.1: Prior distribution of variables $X_1$ and $X_2$, example 1

given below is required as well.

<table>
<thead>
<tr>
<th></th>
<th>True</th>
<th>False</th>
<th>True</th>
<th>False</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_1$</td>
<td>1</td>
<td>0.6</td>
<td>0.2</td>
<td>0</td>
</tr>
<tr>
<td>$X_2$</td>
<td>0</td>
<td>0.4</td>
<td>0.8</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 2.2: $P(X_1, X_2, X_3)$ in example 1

The simple Bayesian model in figure 2.6, together with table 2.1 and table 2.3, is a complete representation of the joint distribution. With these information, we can get the marginal distribution of child node $X_3$:

$$P(X_3 = 1) = \sum P(X_3 = 1 | X_1, X_2)P(X_1, X_2)$$

$$= 1 \times 0.5 \times 0.3 + 0.6 \times 0.5 \times 0.7 + 0.2 \times 0.5 \times 0.3$$

$$= 0.39$$

The same result can also be achieved by using Netica\(^2\).

Now the prior distribution of all variables and conditional probability table are ready for the calculation of the joint distribution by using the chain rule. Note that

\(^2\)Netica is an easy-to-use, complete program for working with BNs and influence diagrams. See www.norsys.com
the parents of $X_3$ are $X_1$ and $X_2$, and both $X_1$ and $X_2$ have parent. So we just use the marginal distribution of $X_1$ and $X_2$ in application of the chain rule. In the following we illustrate one scenario in calculation of joint distribution and leave the others to readers.

\[
P(X_1 = true, X_2 = true, X_3 = true) = P(X_3 = true | X_1 = true, X_2 = true)P(X_1 = true)P(X_2 = true) \]

\[
= 1 \times 0.5 \times 0.3 \]

\[
= 0.15
\]

By applying the same procedure step by step, we get the following joint probability distribution table.

<table>
<thead>
<tr>
<th>$X_1$</th>
<th>True</th>
<th>False</th>
<th>$X_2$</th>
<th>True</th>
<th>False</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>True</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$X_3$</td>
<td></td>
<td></td>
<td>True</td>
<td>True</td>
<td>0.15</td>
</tr>
<tr>
<td></td>
<td>False</td>
<td></td>
<td>True</td>
<td>False</td>
<td>0.21</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>False</td>
<td>True</td>
<td>0.14</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>False</td>
<td>False</td>
<td>0.03</td>
</tr>
<tr>
<td>Probability</td>
<td>0.15</td>
<td>0</td>
<td>0.21</td>
<td>0.14</td>
<td>0.03</td>
</tr>
</tbody>
</table>

Table 2.3: CPT in example 1

In the case of simple BN as showed in figure 2.7 with 3 binary variables, at least
four values have to be specified in a consistent manner for CPT. However, if the variables take 10 possible states for each, then the conditional probability table contains $10^3$ entries, i.e. 1000 conditional probabilities must be acquired and maintained. Furthermore, if the number of variables increases, the entries of CPT goes up exponentially. This would cause a big challenging in data construction or expert assessment. Besides, discrete BNs suffer other serious disadvantages (A.Hanea 2008):

- Applications involving high complexity in data-sparse environments are severely limited by the excessive assessment burden which leads to rapid, informal and indefensible quantification.

- The marginal distributions can often be retrieved from data, but not full interactions between children and parent nodes.

- Discrete BNs take marginal distribution only for source nodes, who have no parent. Marginal for other nodes are computed from the conditional probability tables. When these margins are available from data, this imposes difficult constraints on the conditional probabilities.

- Whereas BNs are very flexible with respect to recalculation and updating, they are not flexible with respect to changes in modeling: if we add on parent node, then we must re-do all previous quantification for the children of this node.

In many situations continuous nodes would be more appropriate. A BN constructed by continuous nodes is named as continuous BN. If variables in a BN follow a joint normal distribution assumption, we talk of Gaussian BNs (or normal BNs) [6]. From quantitative aspect, Gaussian BNs require means, conditional variances and partial regression coefficients to be specified. Given these information, the conditional probability function associated with variables in the BNs are calculable.

Let a BN’s variable set $X = (X_1, \cdots, X_n)$ follow multivariate normal distribution. Then the conditional probability function of variables are of the form:

$$f(X_i|pa(X_i)) \sim N(\mu_i + \sum_{j \in pa(X_i)} b_{ij}(X_j - \mu_j); \nu_i) \quad (2.15)$$
where \( \mu = (\mu_1, \cdots, \mu_n) \) is the mean vector, \( \nu = (\nu_1, \cdots, \nu_n) \) is a vector of conditional variances and \( b_{ij} \) are partial regression coefficients.

However, Gaussian BNs have the constraints of its restriction to the joint normal distribution and, in the absence of data, the estimation of quantitative coefficients by experts. If the normality assumption does not hold, then:

- The individual variables must be transformed to normals.

- The conditional variance in Normal Units must be constant.

- The partial regression coefficients apply to the normal units of the transformed variables, not to the original units.

- If a parent node is added or removed, after quantification, then the previously assessed partial regression coefficients must be re-assessed.

Kurowicka and Cooke (2004) introduced an approach to continuous BNs by using vines (Cooke 1997; Bedford and Cooke 2002) and copulas. In the following section 2.2, details about copula and vines are presented.

### 2.2 Copula and Gaussian copula

Copula is a tool needed to reduce computational cost caused by discretization, when dealing with high dimensional continuous BNs. The study of copulas and their applications is a quite new phenomenon. During past around 30 years interest grows significantly seeing from the publications and international conferences devoted to this topic.

From one point of view, copulas are functions that join or ‘couple’ multivariate distribution functions to their one-dimensional marginal distribution functions. Literature suggests two main reasons that make the study of these functions interesting: "First, as a way of studying scale-free measures of dependence; and secondly, as a starting point for constructing families of bivariate distributions, sometimes with a view to simulation" (Fisher 1997).

#### 2.2.1 Copula
The notion of ‘copula’ was introduced to separate the effect of dependence from the effect of marginal distributions in a joint distribution. A copula is simply a distribution on the unit square with uniform marginal distributions.

**Definition 2.4** (R.B. Nelson 2006) A copula is a function $C$ from $[0,1]^2$ to $[0,1]$ with the following properties:

1. For every $u, v$ in $[0,1]$,
   \[ C(u,0) = 0 = C(0,v) \]
   (2.16)
   and
   \[ C(u,1) = u \text{ and } C(1,v) = v; \]
   (2.17)

2. For every $u_1, u_2, v_1, v_2$ in $[0,1]$ such that $u_1 \leq u_2$ and $v_1 \leq v_2$,
   \[ C(u_2,v_2) - C(u_2,v_1) - C(u_1,v_2) + C(u_1,v_1) \geq 0. \]
   (2.18)

3. $C$ is non-decreasing in each variable.

The definition of copula can be extended to high dimensional, but we restrict on 2-dimensional copula in this report. High dimensional copula can be constructed by 2-dimensional copulas.

Loosely speaking, a copula incorporates the information on dependence structure of related random variables. Suppose we have two continuous random variables $X$ and $Y$, and the corresponding distribution functions $U = F_X(X)$ and $V = F_Y(Y)$, and the inverse $F_X^{-1}(X)$ and $F_Y^{-1}(Y)$. It follows from the probability integral transform that $U$ and $V$ are uniformly distributed on $[0,1]$. With this in mind, we obtain:

\[ P(X \leq F_X^{-1}(u), Y \leq F_Y^{-1}(v)) = P(U \leq u, V \leq v) \]
\[ = C(u,v) \]
(2.19)
and

\[
P(X \leq x, Y \leq y) = P(F_X(X) \leq F_X(x), F_Y(Y) \leq F_Y(y)) = C(F_X(x), F_Y(y))
\] (2.20)

There is a one-to-one correspondence between copula \( C \) and the distribution of \((X, Y)\), as stated in the fundamental theorem of Sklar.

**Theorem 2.1** (Sklar 1959) Random variables \( X \) and \( Y \) are joined by copula \( C \) if their joint distribution can be written as:

\[
c(F_X(x), F_Y(y)) = \frac{f_{XY}(x, y)}{f_X(x)f_Y(y)}
\] (2.21)

Sklar theorem enables us to construct the joint distribution given the specified margins and the underlying copula. On the other hand, it also reveals how to extract the copula of a given bivariate distribution. The same applies to density transformation.

If \( f_{X,Y} \) and \( f_X, f_Y \) denote the nonzero density and marginal densities of \((X, Y)\) with joint distribution function \( F_{X,Y} \), then:

\[
c(F_X(x), F_Y(y)) = \frac{f_{XY}(x, y)}{f_X(x)f_Y(y)}
\] (2.22)

is the copula density of \( C \). Hence the density can be written in terms of the copula density and the marginal densities as:

\[
f_{X,Y}(x, y) = c(F_X(x), F_Y(y))f_X(x)f_Y(y)
\] (2.23)

The rank correlation for \( X \) and \( Y \), as a measure of dependence, can be expressed as:

\[
r = 12 \int_{[0,1]^2} uvC(u, v) - 3 = 12 \int_{[0,1]^2} C(u, v)dudv - 3
\] (2.24)

Copula has quite some families such as Diagonal band copula, Archimedean copula, Frechet, Gaussian and so on. So far, Gaussian copula is the most popular one in real application.

### 2.2.2 Gaussian copula
The Gaussian copula is constructed from a multivariate normal distribution over \( \mathbb{R}^d \) by using the probability integral transform.

For a given correlation matrix \( \Sigma \in \mathbb{R}^{d \times d} \), the Gaussian copula with parameter matrix \( \Sigma \) can be written as

\[
C_{\Sigma}(\mathbf{u}) = \Phi_{\Sigma}(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_d)); \mathbf{u} \in [0,1]^d
\]

where \( \Phi^{-1} \) in the inverse cumulative distribution function of a standard normal and \( \Phi_{\Sigma} \) is the joint cumulative distribution function of a multivariate normal distribution with mean vector zero and covariance matrix equal to the correlation matrix \( \Sigma \).

Reduced to 2-dimensions, the bivariate Gaussian copula can be expressed as:

\[
C_{\rho}(u, v) = \Phi_{\rho}(\Phi^{-1}(u), \Phi^{-1}(v)); u, v \in [0,1]
\]

where \( \rho \) defines the product moment correlation between variables \( X \) and \( Y \). We replace \( \Sigma \) with \( \rho \) for the reason that the product moment correlation determines the covariance matrix for the bivariate joint normal distribution.

\[
\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}
\]

For the Gaussian copula the relationship between the product moment and rank correlation coefficients is given by:

\[
\rho = 2 \cdot \sin(\frac{\pi}{6} \cdot r)
\]

Besides, for the joint normal variables the partial correlation is just equal to the conditional product moment correlation (A.M.Hanea et.al 2006).

\[
\rho_{XY|Z} = \rho_{XY|Z}
\]

Gaussian copula is widely used in BNs’ application because of these two nice properties (mainly but not only), which help us to get rid of computation constraints in quantifying BNs. In this research, only Gaussian copula is employed.

### 2.3 Inference algorithm for Non-parametric BNs
When dealing with continuous BNs, both discrete and joint normal methods present disadvantages, as discussed in section 2.1.3. Kurowicka and Cooke (2004) introduced a Copula-Vine approach to continuous BNs, in which the copula represent (conditional) independence with zero (conditional) rank correlation.

For a non-parametric BN (NPBNs), no joint distribution is assumed. According to the approach proposed by Kurowicka and Cooke (2004), the joint distribution is specified through assembling bivariate pieces. These pieces are modeled through bivariate copula, which are characterized by parametric families. The embedded parameters represent the association between random variables. Such measures of association can be, for example, Spearman’s rank correlation \( r \) or Kendall’s tau. In order to quantify NPBNs, one needs to associate each variable with a one dimensional marginal distribution, each arc in BNs with a (conditional) rank correlation realized by the chosen copula. In continuous NPBNs nodes are associated with continuous invertible distribution functions.

We know that in a NPBN, every arc is assigned with a (conditional) rank correlation. For each variable \( X_i \) with \( m \) parents \( X_k = pa_1(X_i), \ldots, X_l = pa_m(X_i) \) associate the arc \( pa_j(X_i) \rightarrow X_i, \ j = 1, \ldots, m \) with the rank correlation:

\[
r(X_i, pa_j(X_i)), \quad j = 1
\]

\[
r(X_i, pa_j(X_i)|pa_1(X_i), \ldots, pa_{j-1}(X_i)), \quad j = 2, \ldots, m
\]

The order of the parents can be different; hence the assignment above is not unique. The assignment is vacuous if \( pa(X_i) = \emptyset \).
The following theorem is crucial for the copula-vine approach to CNBNs. It shows that these assignments uniquely determine the joint distribution and are algebraically independent [5].

**Theorem 2.2**

Given:

1. a directed acyclic graph with $n$ nodes specifying conditional independence relationships in a BN;
2. $n$ variables, assigned to the nodes, with continuous invertible distribution functions;
3. the specification 2.30 2.31, $i = 1, \cdots, n$ of conditional rank correlations on the arcs of the BN;
4. a copula realizing all correlations $[-1, 1]$ for which correlation 0 entails independence; the joint distribution of the $n$ variables is uniquely determined. This joint distribution satisfies the characteristic factorization 2.14 and the conditional rank correlations in 2.30 and 2.31 are algebraically independent. (see proof in [5])

Both the relatively simple visualization of the complicated dependence relationships between variables and information updating possibility are, probably, the most appealing features of a BN model. Inference (information updating, conditionalizing) means updating the marginal distributions of a subset of variables when the evidence of other variables out of this subset is known. Under the assumption of NPBN and Gaussian copula, exact inference for the model can be performed. To find a given conditional distribution is a matter of computing the conditional distribution on the standard normal variables $Y_1, \cdots, Y_n$. The univariate margins are later transformed back to their original units using their inverse distribution function [5].

Suppose the model includes variables $X_1, \cdots, X_n$ and large enough data samples. We aim to get the belief of $X_n$ given the evidence of variables $X_1, \cdots, X_{n-1}$. The general procedure to get the inference result follows as:

1. Construct dependence structure of variables;
2. Calculate each variable’s marginal distribution $F_1, \cdots, F_n$. The margin can be empirical distribution or user defined distribution.
3. Transform $X_1, \cdots, X_n$ to the standard normal variables $Y_1, \cdots, Y_n$;

4. To each arc of this BN assign the (un)conditional correlation, by computing the sample version rank correlation of $Y_1, \cdots, Y_n$;

5. Compute the correlation matrix $R$;

6. Find out the quantiles of evidence from $F_1, \cdots, F_{n-1}$;

7. Transform the quantiles to standard normal versions;

8. Compute the conditional distribution of $Y_n$, given normal transformed quantiles;

9. Transform $Y_n$ back to original units $X_n$ using inverse of $F_n$.

In case of absence of data, both margins and (conditional) rank correlations need to be assigned by experts.

### 2.4 Copula testing

One of the appealing features of copulas, is that they allow for a "decoupling" of the univariate margins and their dependence structure. Copulas have been widely applied in various areas such as finance, engineering, actuarial sciences and so on. However, a fundamental limitation of the copula approach is that there are in principle an infinite number of possible copulas (Genest and Mackay 1986, Genest 1987, Genest and Rivest 1993, Nelson 1998). Up to now, quite some evidence (D.Kurowicka and R.M.Cooke 2004, A.M.Hanea 2006) shows nice properties of the Gaussian copula, which makes it attractive for application. It is indeed sometimes more advantageous to prefer a simple copula to one that fits the sample data better, but still one needs to be aware of the choice of the copula. It might bring a very wrong result to embrace blindly the Gaussian copula hypothesis to model, if it does not reflect the dependence structure reasonably. Consequently, one wants to know whether the Gaussian copula provides a sufficiently good statistical approximation of the unknown true copula [14].

Lately, several Goodness-of-fit (GOF) copula tests are presented in the literature. Prior to the use of GOF tests, various information criterions were employed, such as Akaike’s information Criterion (AIC). These tests do not provide us with any
understanding of the size of the decision rule employed, nor its power. Hence, GOF tests are preferred. In this report we will discuss three different GOF tests. The first one we call bivariate Gaussian copula test (Y.Malevergne, D.Sornette 2003), the second one the blanket test (C.Genest et.al. 2009) and the third on the D-test (A.M.Hanea, D.kuwicka and R.M.Cooke 2006).

In the following, we denote the null hypothesis as:

$$H_0: \text{The dependence between variables can be described by the Gaussian copula.}$$

i.e. $C_m \in C_0$ ($C_m$ is the empirical copula, $C_0$ is the Gaussian copula with parameter $\theta$)

Before entering into the detailed presentation of each testing method, some general considerations are given below.

2.4.1 General considerations

Recall the definition of bivariate Gaussian copula in section 2.2.2 but extend to n-dimension, the n-dimensional Gaussian copula is the copula derived from the multivariate Gaussian distribution. Let $\Phi$ denote the standard cumulative normal distribution and $\Phi_{\Sigma,n}$ the n-dimensional normal distribution with correlation matrix $\Sigma$. Then the Gaussian n-copula with correlation matrix $\Sigma$ is:

$$C_{\Sigma}(u_1, \ldots, u_n) = \Phi_{\Sigma,n}(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_n))$$

(2.32)

whose density is:

$$c_{\Sigma}(u_1, \ldots, u_n) = \frac{\partial C_{\Sigma}(u_1, \ldots, u_n)}{\partial u_1 \cdots \partial u_n} = \frac{1}{\sqrt{\det \Sigma}} \exp\left(-\frac{1}{2} y(u)^T(\Sigma^{-1} - I)y(u)\right)$$

(2.33)

with

$$y(u) = (y_1(u), \ldots, y_n(u)) \text{ and } y_k(u) = \Phi^{-1}(u_k), k = 1, \ldots, n$$

(2.34)

The Gaussian copula is completely determined by the knowledge of the correlation matrix $\Sigma$, which can be easily estimated from sample data.

Suppose the sample set involves an $m \times n$ data set \{\(x_{ij}\) \(i = 1, \ldots, m; j = 1, \ldots, n\)}, nonparametric estimation of the margins can be derived by using the most natural empirical counterpart:
\[
\hat{F}_j(t) = \frac{1}{m} \sum_{i=1}^{m} \mathbf{1}(x_{ij} \leq t)
\]

(2.35)

And the rank of \(x_{ij}\) among \(x_{1j}, \cdots, x_{mj}\) can be estimated with \(R_{ij} = m \hat{F}_j(x_{ij})\).

When testing the hypothesis \(H_0\), it is preferred to use empirical margins rather than parametric modelling of margin. Such a step would prevent the testing from noise brought by margin fitting.

For GOF testing, given that the underlying copula \(C\) is invariant by continuous, strictly increasing transformations of the components of one random vector, it appears that the most reasonable option for testing \(H_0\) consists of basing the inference on the ranks. They can be viewed as functions of the collection \(U_1 = (U_{11}, \cdots, U_{1n}), \cdots, U_m = (U_{m1}, \cdots, U_{mn})\) of pseudo-observations deduced from the ranks, \(U_{ij} = \frac{R_{ij}}{m+1} = \frac{m \hat{F}(x_{ij})}{m+1}\), where the scaling factor \(\frac{m}{m+1}\) is only introduced to avoid potential problems with \(C_{\theta}\) blowing up at the boundary of \([0,1]^m\).

The information contained in pseudo-observations \(U_1, \cdots, U_m\) is conveniently summarized by the associated empirical distribution, viz.

\[
C_m(u) = \frac{1}{m} \sum_{i=1}^{m} \mathbf{1}(U_{i1} \leq u_1, \cdots, U_{in} \leq u_n), \ u = (u_1, \cdots, u_n) \in [0,1]^n
\]

(2.36)

This function is usually called the 'empirical copula', though it is neither a copula nor exactly the same (except asymptotically) as originally defined by Deheuvels (1979).

Besides the definitions presented up to now, we introduce one proposition that will be used for deriving the test statistics in next sections.

**Proposition 2.2** Assuming that the \(n\)-dimensional random vector \(x = (x_1, \cdots, x_n)\), with distribution function \(F\) and marginal \(F_i\), satisfies the null hypothesis \(H_0\), then the variable

\[
z^2 = \sum_{i=1}^{n} \Phi^{-1}(F_i(x_i))(\Sigma^{-1})_{ij}\Phi(F_j(x_j))
\]

where the entries of matrix \(\Sigma\) is consisted by

\[
\Sigma_{ij} = \text{cov} \left[ \Phi^{-1}(F_i(x_i)), \Phi^{-1}(F_j(x_j)) \right]
\]

follows a \(\chi^2\)-distribution with \(n\) degrees of freedom.
Proof: See proof in [13], page 236.

Other background considerations needed for each method presented next will be investigated separately.

2.4.2 Bivariate Gaussian copula test

As will be seen from this section, at first glance, the bivariate Gaussian copula test seems only valid for models with two random variables. Nevertheless, the restriction to pairs of random variables here is not crucial for the reason that the testing procedure exposed in the following can be applied to any number of variables by means of combination. For simplicity purpose, only one pair test is presented. Moreover, testing the Gaussian copula hypothesis for two random variables is not restrictive because it gives useful information for a large number of dependent variables constituting a model. Indeed, assume that each pair \((X_1, X_2), (X_2, X_3)\) and \((X_3, X_1)\) has a Gaussian copula, and in addition that the copula of triplet \((X_1, X_2, X_3)\) is elliptical. Then, the triplet \((X_1, X_2, X_3)\) also has a Gaussian copula. This result generalizes to an arbitrary number of random variables [13]. For more details refer to [13], we will just sketch roughly here\(^3\).

Consider two random variables \((n = 2)\) with sample size \(m\), \(X_1 = (x_{11}, \ldots, x_{1m})\), \(X_2 = (x_{21}, \ldots, x_{2m})\). Then the cumulative marginal distribution \(\hat{F}_i, i = 1, 2\) of each variable and estimated Gaussian transform \(\hat{y}_i\) can be derived by using formula 2.38. The sample correlation matrix \(\Sigma\) is estimated by the expression 2.37. This is the Pearson's product moment correlation coefficient of the Gaussian transforms.

\[
\Sigma = \frac{1}{m} \sum_{j=1}^{m} \hat{y}(j) \cdot \hat{y}(j)^t, \hat{y}(j) = (\hat{y}_{1j}, \hat{y}_{2j})
\] (2.37)

\[
\hat{y}_{ik} = \Phi^{-1}(\hat{F}_i(x_{ik})), i = 1, 2; k = 1, \ldots, m
\] (2.38)

which allows us to calculate the variable

---

\(^3\)An elliptical distribution (and an elliptical copula) is fully determined by the knowledge of its mean, its shape (or covariance) matrix, and the generator of its type. Once the distributions of every pair of random variables \((X_i, X_j), i, j \in [1, \ldots, n]\), are known, the type of the generator is fixed and the mean and the shape matrix of the joined distribution of \((X_1, \ldots, X_n)\) can be constructed.
\[ z_k^2 = \sum_{i,j=1}^{2} \hat{y}_{ik}(\Sigma^{-1})_{ij} \hat{y}_{jk} \] (2.39)

\( z^2 \) should be distributed according to a \( \chi^2 \)-distribution if the Gaussian copula hypothesis is correct as proposition 2.2 shows.

The common way of comparing the empirical and theoretical distribution is to measure the distance between these two distributions and to perform the Kolmogorov test or the Anderson-Darling test. Two additional measures which are defined as averages of the Kolmogorov and Anderson-Darling distance respectively.

Kolmogorov distance:

\[ d_1 = \max_z |F_{z^2}(z^2) - F_{\chi^2}(z^2)| \] (2.40)

Average Kolmogorov distance:

\[ d_2 = \int |F_{z^2}(z^2) - F_{\chi^2}(z^2)|dF_{\chi^2}(z^2) \] (2.41)

Anderson-Darling distance:

\[ d_3 = \max_z \frac{|F_{z^2}(z^2) - F_{\chi^2}(z^2)|}{\sqrt{F_{\chi^2}(z^2)[1 - F_{\chi^2}(z^2)]}} \] (2.42)

Average Anderson-Darling distance:

\[ d_4 = \int \frac{|F_{z^2}(z^2) - F_{\chi^2}(z^2)|}{\sqrt{F_{\chi^2}(z^2)[1 - F_{\chi^2}(z^2)]}}dF_{\chi^2}(z^2) \] (2.43)

The Kolmogorov distances are more sensitive to the deviation occurring in the bulk of the distributions, while Anderson-Darling distances are more accurate in the tails of the distributions [13].

The procedure suggested by [13] page 237, in order to test the Gaussian copula assumption for a given data set, reads:

1. Generate the standard Gaussian transforms \( \hat{y}_i \) 2.38 from the original sample data;

2. Estimate the correlation matrix \( \Sigma \) 2.37 of the Gaussian variables, which allows us
to compute the variables $z^2$ 2.39 and then measure the distances $d_1, d_2, d_3$ and $d_4$, of its estimated distribution to the $\chi^2$-distribution;

3. Given this correlation matrix $\Sigma$, generate numerically Gaussian random vectors with zero mean and the same correlation coefficient matrix $\Sigma$;

4. For this series of Gaussian vectors, estimate the corresponding correlation matrix $\tilde{\Sigma}$;

5. With each of the synthetic Gaussian vectors, associate the corresponding realization of the random variables $z^2$, called $\tilde{z}^2$

6. Construct the empirical distribution for the variables $\tilde{z}^2$ and measure the distances $d_1^*, d_2^*, d_3^*$ and $d_4^*$, between this empirical distribution and the $\chi^2$-distribution;

7. Repeat the steps (3) – (6) a large number of times $N$ to obtain an accurate estimate of the cumulative distribution of distances $d_1^*, d_2^*, d_3^*$ and $d_4^*$, between the distribution of the synthetic Gaussian variables and the theoretical $\chi^2$-distribution. This cumulative distribution represents the test statistic, which allow us to reject or not the null hypothesis $H_0$ at a given significance level;

8. Compute the P-value by $P_{d_i} = \frac{1}{N} \sum_{j=1}^{N} 1(d_i > d_{ij}^*), i = 1, 2, 3, 4$.

Given significance level $\alpha$, if the P-value is bigger than $\alpha$, then the Gaussian copula hypothesis cannot be reject. Otherwise it reject the hypothesis $H_0$.

### 2.4.3 Blanket test

The GOF test 'Blanket test' is applicable to all copula structures and requiring no strategic choice for their use, compared to the bivariate Gaussian copula test presented in previous section and some other categorization methods. Blanket test does not have the restriction of model’s dimension and class.

Given the empirical copula is entirely nonparametric, $C_m$ in equation 2.36 is arguably the most objective benchmark for testing hypothesis $H_0$ (C.Genest, B.Remillard, D.Beaudoin 2007). Suppose we want to apply the parametric copula
family $C_\theta$ with parameter $\theta$. The most natural choice of $\theta$ is the estimated one which maximizes the log pseudo-likelihood

$$\ell(\theta) = \sum_{i=1}^{m} \log \left( c_\theta(\hat{F}_1(x_{i1}), \cdots, \hat{F}_1(x_{in})) \right)$$

(2.44)

where $\hat{F}_j$ is the empirical counterpart as presented in equation 2.35. Thus, the distance between $C_m$ and estimated copula $C_\theta$ of the underlying copula is proposed to be a measure for testing the null hypothesis. One such distance is presented in equation 2.45

$$d_m = \sqrt{m(C_m - C_\theta)}$$

(2.45)

GOF tests based on the empirical process $d_m$ are briefly considered by Fermanian (2005), who comments that they "seem to be unpractical, except by bootstrapping" (C.Genest, B.Remillard, D.Beaudoin 2009). In particular, two versions of statistic are taken into account, they are:

Carmer-von Mises statistic:

$$S_m = \int_{[0,1]^n} d_m(u)^2 dC_m(u)$$

(2.46)

Kolmogorov-Smirnov statistic:

$$T_m = sup_{u \in [0,1]^n} |d_m(u)|$$

(2.47)

Large values of these statistics lead to the rejection of $H_0$, and an approximate P-value can be deduced from their limiting distributions. However, these two limiting distributions depend on the family of copula under the composite null hypothesis, and on parameter $\theta$. Consequently, the test statistics and P-value can only be obtained via specially adapted Monte Carlo methods, that is, a parametric bootstrap procedure is applied. As with the bivariate Gaussian copula testing method presented in previous section, 10,000 synthetic repetitions are recommended. Genest and Remillar show that the tests based on $S_m$ and $T_m$ are consistent.

We summarize the following procedure leads to an approximate P-value for the test based on $S_m$. 
1. Compute the empirical copula $C_m$, choose parametric copula family $C_\theta$ and estimate the parameter $\theta$;

2. Generate a random sample $V_1, \cdots, V_k$ from distribution $C_\theta$;

3. Approximate $C_\theta$ by $B_k = \frac{1}{k} \sum_{i=1}^k \mathbf{1}(V \leq v), v \in [0,1]^m$, $k \geq m$;

4. Approximate $S_m$ by $S_m = \sum_{i=1}^m (C_m(V_i) - B_k(V_i))^2$;

5. Step into bootstrap with step size $N$. For each iteration step $s$, generate a random sample $V_1^*, \cdots, V_k^*$ from distribution $C_\theta$. Based on new data sample $V_1^*, \cdots, V_k^*$, repeat step (1)--(4) to get $S_{sk}^*$. Then vector $S_{1k}^*, \cdots, S_{Nk}^*$ for null statistics' null distribution is generated.

6. Compute the P-value by $P = \frac{1}{N} \sum_{s=1}^N \mathbf{1}(S_{sk}^* > S_m)$

Generally, in GOF test the greater the sample size, the better. Large data sets help in the reliability of the parametric bootstrap procedures used to approximate the statistics’ null distribution. Also, in order to get efficient double bootstrap, the number $k$ of repetitions must be substantially larger than the sample size $m$.

2.4.4 D-test

D-test is a Gaussian copula validation (only for Gaussian) tool employed in UNINET. UNINET is a software which allows for quantification of discrete and continuous BNs. We use UNINET as the tool for learning the structure of BNs and generating reference results for other methods. This learning structure from an ordinal multivariate data set does not make any assumptions about marginal distribution of variables.

As we have seen, hypothesis testing requires an overall measure of multivariate dependence on which statistical test can be based. In UNINET, such a measure as the determinant of the correlation matrix with $n$ variables, $D_{1,\ldots,n}$, is employed. Hence, this test method is called D-test.

**Theorem 2.2** Let $D$ be the determinant of a $n$-dimensional correlation matrix
(\(D > 0\)). For any partial correlation BN specification

\[
D = \prod (1 - \rho_{ij}^2)
\]  

(2.48)

where \(\rho_{ij;D_{ij}}\) is the partial correlation associated with the arc between \(i\) and \(j\), with conditioning set \(D_{ij}\), and the product is taken over all arcs in the BN.

**Proposition 2.3** \(D_{1,\ldots,n} = \text{Det}(\Sigma)\), with \(\Sigma\) the correlation matrix of \(X_1,\ldots,X_n\) satisfies:

- \(0 \leq D_{1,\ldots,n} \leq 1\);
- \(\forall i, D_i = 1\);
- \(\forall \pi \in n!, D_{1,\ldots,n} = D_{\pi(1),\ldots,\pi(n)}\);
- \(K,J \subseteq \{1,\ldots,n\}, L/K = \emptyset, X_K \perp X_J \Rightarrow D_{K,J} = D_K D_J\);
- \(\perp\{1,\ldots,n\} \Rightarrow D_{1,\ldots,n} = 1\);
- \(X_1 \in l(X_2,\ldots,X_n) \Rightarrow D_{1,\ldots,n} = 0\);
- \(D_{1,\ldots,n} = 0, D_{2,\ldots,n} > 0 \Rightarrow X_1 \in l(X_2,\ldots,X_n)\).

where \(l(X_1,\ldots,X_n)\) denotes the linear span of the variables \((X_1,\ldots,X_n)\), \(\perp\{1,\ldots,n\}\) says that the variables \((X_1,\ldots,X_n)\) are independent.

**Proof:** See proof in [8].

On the basis of the above background theory, it is reasonable to choose the determinant of the correlation matrix to be the measure of multivariate dependence.

In the following a brief description of D-test is presented (see more detail in [8]). Suppose we have a multivariate data set, we may define:

- DER = the determinant of the empirical rank correlation matrix;
- DNR = the determinant of the rank correlation matrix obtained by transforming the univariate distributions to standard normals, and then transforming the product moment correlations to rank correlations using Pearson’s transformation;
- DBBN = the determinant of the rank correlation matrix of a BN using the Gaussian copula.
DNR assumes the Gaussian copula, thus differs from the DER, if the empirical copula is not Gaussian. Set the bootstrap and generate the sample data for many times, a distribution of DNR will be gotten. Given significance level $\alpha = 0.05$, if DER locates in the 90% confidence level of DNR, then the Gaussian copula hypothesis $H_0$ cannot be rejected. If the value of DNR falls in the confidence interval of DBBN, then the BN model structure represent the data fairly.

In UNINET, simply clicking the ‘Validation of Gaussian copula’ option and choosing the bootstrap/sample size, it will automatically be done.
CHAPTER III
TRAFFIC LOADS IN BRIDGES

3.1 Problem Statement

It is very important not only in the design of bridges, but also in the reliability assessment of existing structures to understand the dynamic and static responses of bridges under traffic flow. The problem has been investigated extensively in the past years (see for example [16]). In short the problem consists of investigating the possible consequences of very heavy loaded vehicles under unusual traffic configurations, which place biggest burdens in infrastructure and safety in general. Thus, of importance for this problem is vehicles’ distribution over a particular bridge given likely maximal bending moment of the bridge. A visual representation of the problem is presented below, problem statement and brief description of the data will be presented in the following.

Figure 3.1: A visual bridge model

Figure 3.1 represents a L meters long bridge model with two lanes, fast and slow with respect to traffic. In our research, 100 and 50 meters long bridge models are taken into consideration. For simplicity purpose, only one traffic direction is taken into
CHAPTER III : TRAFFIC LOADS IN BRIDGES

(a) A bridge on A16 road near Rotterdam (b) Maximal bending moment curves for an extreme case

Figure 3.2: A maximal bending moment example on a Dutch bridge

consideration. We consider the static situation, from the point of view of equivalent bending moment at the mid-span of the bridge.

In physical terms, gravity causes the force acting on bridge beams and surface when vehicles pass by. From figure 3.2 one can observe how the weights and positions of axles (we say these balls represent axles) affect bridge’s bending moment qualitatively. The maximal bending moment curves are generated from an extreme case of a short bridge with unequal loads. The three lightly dotted parabolas are maximum moment curves under each of the loads taken separately. The three heavily dotted curves are curves of maximum moment under each of the loads, for the three loads passing over the bridge, at the given distances, from left to right. The heavy continuous line gives the curve for the reverse direction of passage of the loads. As might be expected, the more, heavier, closer to the center, of loads, the bigger the bending moment would be.

People might not be satisfied with the qualitative conclusion, but require a quantitative representation. In real application, very great accuracy in drawing this kind of curves or the determined bending moment is unnecessary because of the dynamic traffic flow. Many literatures specializing in bridge analysis employ grillage and finite element method[19]. In grillage method the deck is represented by an equivalent grillage of beams. The finite element method is a numerical method solving complicated structural engineering problems. It predicts the bridge behavior under

---

the truck axle loading. Both methods rely much on physics. However very little may be said about uncertain traffic flow. In our research we motivate us to start with a specific data driven model - Non-parametric Bayesian Networks under the Gaussian copula assumption.

The variables involved include loads (kN), number of axles and maximal bending moment (kNm) per day. These variables represent a complex Multidimensional Distribution. One appealing feature of BNs would be, for example that it is possible to say something about the distribution of bending moment with the evidence of configuration of loads and axles (conditionalization). This is of interests since it is useful in applications such as the design of new bridges, the rating and fatigue life assessments of existing bridges and the control of overweight vehicle or traffic flow. Another interesting representation for practitioners is the configuration of axles leading to maximal bending moments.

To make it more clear, we shortly present two questions of interest here:

1. Predict the configuration of vehicles, given know information of the maximal bending moment;
2. Predict the value of bending moment, when the vehicles’ configuration is obtained.

With these ideas in mind, the problem is to develop an accurate prediction algorithm to try to answer problems similar to the ones discussed above.

The research was conducted on the basis of data simulated from WIM systems’ record at location RW-16L, a highway in the Netherlands (see section 3.2.1). The model needs to be re-quantified when applied to other locations. This is because different materials, structures and traffic configurations will influence the correspondence relationship between bending moment and vehicle configuration.

3.2 Data Preparation
3.2.1 Data simulation

The data sets used for model are not entirely collected from measurement. Rather, it has been generated by Monte-Carlo simulation [17]. This simulation includes
two main modeling components. First, a CNBN is quantified with data formed by the Weight-in-motion (WIM) system of the Netherlands. The WIM system plays an important role in collecting data in order to determine traffic loads and has been operated by related government institutions in the Netherlands. It allows trucks to be weighed in the traffic flow without disruption of the traffic flow. For a recent overview of WIM systems see [36]. For each vehicle, variables recorded by the system in the Netherlands include vehicle speed, number of axles, inter-axle-separation and axle weight. These data or variables in general follow a complex multidimensional probability distribution. A CNBN was employed to capture the dependence structure.

This model is shown in figure 3.3. The model represents the joint distribution of individual axle loads for vehicles with 2 up to 11 axles in highways 04, 12, 15 and 16 in the right and left direction. The model is extensively discussed in [17].

The second modeling tool corresponds to time series. WIM system’s data is also used in the quantification of the time series model for vehicle separation. This step is quite important because of its effect on bending moment. The time series model described in [18] chapter 8 is used. Let the time series for separation be denoted \( S_1, S_2, \cdots \) or \( \{ S_t : t = 1, 2, \cdots \} \). Suppose \( S_t \sim F \), where \( F \) is a continuous univariate cumulative distribution function (cdf) with density \( f \). Then \( F_{XY}(x,y) = C(F_X(x), F_Y(y)) \) is a bivariate distribution with univariate margins both equal to \( F \), where \( C \) denotes the copula of variables \( X \) and \( Y \). Let \( C_{Y|X}(v|u) \) denote the conditional distribution of the copula. The transition distribution of \( \{ S_t \} \) is:
\[ H(s_t|s_{t-1}) = P(S_t \leq s_t|S_{t-1} = s_{t-1}) = C_{Y|X}(F(y_t)|F(y_{t-1})) \]  

(3.1)

The conditional copula \( C_{Y|X} \) in this case is modeled with auto correlations \( r(S_t, S_{t-1}) = 0.2, 0.3, 0.4, 0.8 \), of which 0.2 is calculated from data, and the other three parameters are set to look at the sensitivity of \( r(S_t, S_{t-1}) \) to the inference results. A realization of \( S_t \), given \( L = 100 \text{ meters} \) and \( r(S_t, S_{t-1}) = 0.2 \), for 208 vehicles is given in Figure 3.4:

![Figure 3.4: A realization of \( S_t \) with \( r(S_t, S_{t-1}) = 0.2 \)](image)

The two models briefly described above are combined in a simulation procedure described next. The data used corresponds to highway 16 in the left direction for the month of April 2009. The purpose is to investigate the equivalent uniform distribution load (EUDL). The Monte-Carlo simulation process consists of three steps in general, which are:

1. Generate randomly vehicles representing the average number of vehicles in a day, which is 7937 in this project. That is, those observed in highway 16L for a typical month such as April 2009. We will consider two lanes bridges. Important characteristics of each vehicle in these lanes are:
   - The number of axles of each vehicle, in each of the two lanes
   - Individual axle weight
   - The inter-axial separation per vehicle
   - The separation of each vehicle with respect to the vehicle right behind it
2. Once these vehicles are generated, we position the two lanes in the beginning of
the bridge and let them advance in the bridge with a step of 1(m);

3. Individual axle loads at step $k$ are denoted as $w_j$ and the distance from the
beginning of the bridge to the position of the $j$-th axle is denoted as $d_j$. At every
step $k$ we compute the EUDL as follows:

$$c_j = \begin{cases} 
\frac{d_j}{L}, & \text{for } L - \frac{d_j}{2} \\
\frac{L - d_j}{2}, & \text{for } d_j \geq \frac{L}{2}
\end{cases}$$

$$q_k = \frac{\left(\sum_j c_j w_j\right)}{L^2/8}$$

(3.2)

where $q_k$ is the quantity of interest at every iteration step $k$.

The time interval is 6 years with 360 days per year. In general we are interested
in recording maximum occurrences in a time interval. Thus here we record the
maximal bending moment per day, and the corresponding vehicle configuration. The
vehicle configuration refers to the number of axles, their individual weights and their
position over the bridge at the moment where the maximal bending moment is
observed in the middle of the span.

Figure 3.5: A realization of the simulation

Figure 3.5 shows a sample from the realization of simulation. Corresponding
to the $8^{th}$ largest maximal bending moment configuration. There are 9 axles in the first/slow lane and 5 in the second/fast one. Most axles (8 from the slow lane and all 5 from the fast lane) locate in the middle of the bridge at this moment. According to figure 3.5, a new vehicle is coming in as shown by the axle load observed close to the beginning of the bridge. All axle loads show individual load between 50($kN$) and 150($kN$) each.

The simulation records:

1. Maximal bending moment per day of the fast lane, slow lane and the whole bridge.
2. The number of axles with corresponding to the maximum bending moment, in each lane and the whole bridge.
3. Individual axle’s position and weight.

These data recorded by the simulation procedure will be used further in our analysis, not in its raw form but already summarized. We describe in next section how the data is processed for its final use in our BN model.

### 3.2.2 Data processing

We consider the strategy of separating the bridge into $N$ sections. We categorize the axles into each section according to their positions, and then sum the weights and number of axles locating in the same section. In this research, only the cases with $N = 1$ and $N = 3$ are investigated. For a 100 meter bridge the case $N = 1$ means we consider the whole bridge as a whole and $N = 3$ means we consider 3 sections of $100/3(m)$ each. This method could be extended into high $N$ values if needed, in case of longer bridges. The idea can be easily obtained by looking at the following example. Suppose we have the simulation data set as following:

<table>
<thead>
<tr>
<th>Positions(m)</th>
<th>Slow lane</th>
<th>Fast lane</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positions(m)</td>
<td>$94.9$</td>
<td>$85.1$</td>
</tr>
<tr>
<td>Loads(kN)</td>
<td>$56.9$</td>
<td>$30.4$</td>
</tr>
</tbody>
</table>

**Table 3.1:** One simulation sample of axles on the bridge

Table 3.1 shows the position of each axle corresponding to the observed maximum bending moment of the day with respect to the beginning of the bridge in
each lane. The loads of each axle are also shown. If we choose \( N = 3 \), then the whole bridge is separated into 3 sections, where the position intervals are \([0, 33.33], [33.33, 66.67] \) and \([66.67, 100]\) approximately. Looking at axles’ positions in the slow lane in table 3.1, the first two axles are categorized into section 3, i.e. \([66.67, 100]\), while the left three are in the second section \([33.33, 66.67]\). Thus, the number of axles in the third section in slow lane is 2, and the loads sum is 87.3\((kN)\), while the number of axles in the second section in slow lane is 3, and the total loads is 292.3\((kN)\). Note that in this case there is no axle that locates in the first section in the slow lane, meaning that both number of axles and weights are zero in this case. Actually this scenario turns out to be a big challenge when applying BNs because of its high frequency of occurrence. We will discuss further on this issue in the implementation section.

After accomplishing the data categorizing, new variables are constructed. The notations and definitions of variables involved in case \( N = 1 \), are as follows.

<table>
<thead>
<tr>
<th>Name of variables</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>( W_f ) (kN)</td>
<td>the total weights of axles in the fast lane</td>
</tr>
<tr>
<td>( W_s ) (kN)</td>
<td>the total weights of axles in the slow lane</td>
</tr>
<tr>
<td>( C_f )</td>
<td>the total number of axles in the fast lane</td>
</tr>
<tr>
<td>( C_s )</td>
<td>the total number of axles in the slow lane</td>
</tr>
<tr>
<td>( q_q )</td>
<td>the maximum bending moment per day over the whole bridge</td>
</tr>
</tbody>
</table>

Table 3.2: Notations and definitions of variables involved in case \( N = 1 \)

Note that the maximum bending moments regarding to individual lanes are not taken into consideration here. These might be important for individual components of bridges with certain characteristics. The models to be discussed here may be extended when investigating such bridge configurations would be of importance. One can apply the same mechanical model for the maximum bending moment per day over the fast lane \( q_qf \) or the slow lane \( q_qs \), by simply replacing \( q_q \) with \( q_qf \) or \( q_qs \), the other four variables would remain the same.

Plotting the frequency histograms for selected variables in the data set, we get figure 3.6 with the values on the x-axis and number of samples on the y-axis.

From Figure 3.6, we infer that \( W_s \) cumulates much more at zero than \( W_f \), which centers around 1000\((kN)\). This phenomenon comes from the sparser vehicles
distribution in the slow lane. Also, $W_f$ and $C_f$ behave quite consistently, i.e. high weights come together with a high number of axles. This can be observed from the frequency histogram of average weight of axles for each sample in figure 3.7. We can observe that the most average weights locate around 100, and declines at both sides. No average value bigger than 140 is observed in this data set. However, one may speculate that average weight of each axle would tend to behave consistently due to technological constraints (individual axles are designed to handle certain weights).

**Figure 3.6:** Histogram of $W_f$, $W_s$, $C_f$ and $q_q$

Case 2: $N = 3$

Variables in case 2 have the same properties as those in case 1.
CHAPTER III: TRAFFIC LOADS IN BRIDGES

<table>
<thead>
<tr>
<th>Name of variables</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wf1 (kN)</td>
<td>the total weights of axles in the first section of the fast lane</td>
</tr>
<tr>
<td>Wf2 (kN)</td>
<td>the total weights of axles in the second section of the fast lane</td>
</tr>
<tr>
<td>Wf3 (kN)</td>
<td>the total weights of axles in the third section of the fast lane</td>
</tr>
<tr>
<td>Ws1 (kN)</td>
<td>the total weights of axles in the first section of the slow lane</td>
</tr>
<tr>
<td>Ws2 (kN)</td>
<td>the total weights of axles in the second section of the slow lane</td>
</tr>
<tr>
<td>Ws3 (kN)</td>
<td>the total weights of axles in the third section of the slow lane</td>
</tr>
<tr>
<td>Cf1</td>
<td>the total number of axles in the first section of the fast lane</td>
</tr>
<tr>
<td>Cf2</td>
<td>the total number of axles in the second section of the fast lane</td>
</tr>
<tr>
<td>Cf3</td>
<td>the total number of axles in the third section of the fast lane</td>
</tr>
<tr>
<td>Cs1</td>
<td>the total number of axles in the first section of the slow lane</td>
</tr>
<tr>
<td>Cs2</td>
<td>the total number of axles in the second section of the slow lane</td>
</tr>
<tr>
<td>Cs3</td>
<td>the total number of axles in the third section of the slow lane</td>
</tr>
<tr>
<td>qq (kNm)</td>
<td>the maximal bending moment of the whole bridge</td>
</tr>
</tbody>
</table>

Table 3.3: Notations and definitions of variables involved in case \( N = 3 \)

3.3 Model Construction

3.3.1 Learning the structure of BN from data

Data mining is the process of extracting and analysing information from a large database. A method for mining ordinal multivariate data using CNBN with Gaussian copula avoids the computational complexity often encountered in this type of problems. An ordinal multivariate data set is one in which the numerical ordering of values for each variables is meaningful. A database containing bending moment, axles’ loads and axel number at each time point on a bridge is ordinal. However, any conditionalization result from mining is meaningless if the model does not represent the data reasonably. Hanea [5] introduced an algorithm of learning the structure of a CNBN with the Gaussian copula from an ordinal multivariate data set, that may contain a large number of variables with no assumptions about the marginal distribution of the variables involved. Hence it will not raise problems thanks to its allowing for quantification of mixed discrete and continuous BNs and its free of marginal distribution constraints.

The learning algorithm has been implemented into the software UNINET. UNINET has a friendly interface and fast simulation. Furthermore, a limited version of
UNINET is available \(^5\). In our research, all structure learning of BNs are accomplished in UNINET.

Citing the definitions of DER, DNR and DBBN in section 2.4.4, a rough statistical test for the suitability of DNR for representing DER is to obtain the sampling distribution of DNR and check whether DER is within the 90% (also can be 95% or 99%, decided by users) central confidence band of DNR. This statistical test gives an indication whether the joint normal copula adequately represents the original data, so-called Gaussian copula validation. It compares the dependence structure of the original data with the dependence structure of the normal data. Another test to check whether the current model is a good representation of the normal data, one needs to obtain the sampling distribution of DBBN and check whether DNR locates within the 90% central confidence band of DBBN. It compares the dependence structure of the model with the dependence structure of the normal data. This test is named as model validation.

The procedure for learning structure of a BN from data is not fully automated; the general steps that we have used are as follows:

1. Import database to UNINET;
2. Construct a saturated BN on chosen variables;
3. Execute the model validation and Gaussian copula validation, if both tests are successful, then stop, else continue with the following steps;
4. Compute the conditional rank correlation coefficients, remove the arcs whose assigned coefficients are below 0.1 (selected by users);
5. Repeat step 3 and 4 until both validation tests are successful.

It is quite possible that one never obtain both a successful model validation and Gaussian copula validation. In this case, the general rule to choose the model is that this model meets following two criteria:

- this model is successful in model validation;

\(^5\)www.lighttwist.net
• the DER of this model is the closest to the central confidence band of DNR.

When applying Gaussian copula approach to BN quantification, we assign the each arc of BN with a conditional rank correlation coefficient. When data is available these coefficients may be directly estimated from data. If we start with a unsaturated BN, we add arcs between variables only if the rank correlation between those two variables is large enough. We also remove arcs from the BN, which correspond to small rank correlations.

In general, there is no ‘best’ model. Two models with different structure may give equal results in the sense of quality. As long as the model can represent the data fairly to some extent, it could be a choice in practice. Also, some small influences may be included because the user wants to see these influences, even though they are small.

3.3.2 Ordinal bridge data mining with UNINET

We illustrate our algorithm for learning a BN from data using an ordinal multivariate bridge data set. The data are gathered from simulation of WIM over three years, with parameters $L = 100(m)$ and $r(S_t, S_{t-1}) = 0.2$, and processed under procedures that we briefly introduced in section 3.2.3. The database contains daily maximal bending moment, and corresponding loads and number of axles quantities. Besides maximal bending moment $qq$, case $N = 1$ contains four more variables $Wf$, $Ws$, $Cf$ and $Cs$. For case $N = 3$, it includes twelve more variables, which are $Wf_1$, $Wf_2$, $Wf_3$, $Ws_1$, $Ws_2$, $Ws_3$, $Cf_1$, $Cf_2$, $Cf_3$, $Cs_1$, $Cs_2$, $Cs_3$. Since we have daily data over the course of three years, the data set will contain 2160 samples forming the joint distribution.

The goal is to build a BN that captures the dependence structure between these variables, using the approach presented in the previous section. Both cases $N = 1$ and $N = 3$ are investigated.

We take the one dimensional marginal distributions directly from the data, and model the dependence with the joint normal copula. The hypothesis that the dependence structure in the data is that of a joint normal copula can be tested by Gaussian(normal) copula validation. Once we have a suitable BN, we are aiming to find out the relationship between maximal bending moment ‘qq’ and the other
variables. We start with model construction by selecting reasonable variables to be contained in model. For this purpose, we compute the empirical rank correlation matrix.

Figure 3.9: BN on 13 nodes with no arcs of case $N = 3$

Since we are aiming to look for the dependence relationship between qq and the other variables, we observe the empirical correlation matrix and pick out the highly correlated variable pairs.

Figure 3.8 presents a BN with no arcs on the left, and on the right is the empirical correlation matrix of the 5 variables in the BN at left hand side. One can notice that the variables pair Ws and Cs are highly correlated, with correlation value 0.995\(^6\). This indicates that Ws and Cs provide almost the same information to

---

\(^{6}\)The standard to distinguish a high correlation and a low correlation is determined by users. In our research, we set it as 0.95
determine $qq$. From physical point of view, it is the axle loads rather than the quantity of axles that determine the most bending moment at the middle of the deck. Furthermore, the quantity of axles can be derived from axle loads. Zero loads indicates no axles pass by, higher loads increases the possibility of denser axles. Hence in the model of case $N=1$, we exclude variable $Cs$ and include $Wf$, $Ws$, $Cf$, $qq$.

Figure 3.10 shows the empirical correlation matrix of case $N=3$. The correlation value of variable pairs $(Wf1,Cf1)$, $(Wf3,Cf3)$, $(Ws1,Cs1)$, $(Ws2,Cs2)$ and $(Ws3,Cs3)$ are above 0.95, thus we exclude variables $Cf1$, $Cf3$, $Cs1$, $Cs2$ and $Cs3$ from the BN.

With the remaining variables, we start by considering these variables as being completely dependent, that is, a saturated graph. The discussion in the previous section recommends the choice of the determinant of the rank correlation matrix as an overall dependence measure. The saturated BN induces a joint distribution whose rank determinant is equal to DNR, since the BN uses the Gaussian copula. Further we will remove those arcs that are associated with very small (close to zero) correlations, such that the value of DNR stays inside the confidence interval for DBBN, and the value of DER stays within the central confidence band of DNR. Figure 3.11 and 3.12 present the saturated BN model of case $N=1$ and $N=3$ correspondingly. UNINET calculates from data the (conditional) rank correlations that are assigned to the arcs of the BN.
Figure 3.11: Saturated BN model of case $N = 1$

Figure 3.12: Saturated BN model of case $N = 3$
Following the model construction procedure described in the previous section, we execute the model validation and Gaussian copula validation in UNINET. For case $N = 1$, the determinant of the normal rank correlation matrix $DNR = 0.12133$ falls between the 0.55 and 0.6 quantiles of DBBN, the determinant of the rank correlation matrix of this BN using Gaussian copula. Meanwhile, the determinant of the empirical rank correlation matrix $DER = 0.12775$ is within the 90% central confidence band of DNR, specifically, within the 0.8 – 0.85 quantiles of the distribution of DNR. In this case, both tests indicate that BN from figure 3.13 represents fairly the data.

Figure 3.13 presents the validation results of saturated BN model for case $N = 1$. $DNR$ locates within the 90% confidence interval of DBBN. However, UNINET calculates the DER of this BN model as 0.055251. It falls outside of the required central confidence band, since the 5% quantile of DNR is 0.059339. We have to face the fact that the saturated BN model of case $N = 3$ fails in Gaussian copula validation.

Following BN model construction step 4, we remove the arcs whose corresponding (conditional) rank correlation is lower (in absolute value) than the value 0.1, which is specified by users. UNINET estimates the (conditional) rank correlations from data. In figure 3.12 we underline the coefficients whose absolute values are below 0.1 with red. A semi-saturated BN is generated by deleting the corresponding arcs, as shown in figure 3.15. Figure 3.16 presents the validation results. The determinant of
the rank correlation matrix based on the new BN differs from DNR, as this BN is semi-saturated. In the new model the 90% central confidence interval is [0.058443, 0.076336]. Although the semi-saturated BN model still fails in Gaussian copula validation, we notice that DER is closer to the 5% quantile of DNR than the saturated BN model. Hence the new model improves in representing data.

One option would be to keep removing/adding arcs to the BN in order to comply with the hypothesis test. Nevertheless, we stay with this model to consider all arcs whose associated (conditional) rank correlations are above 0.01 in absolute value. Moreover, we want to use other hypothesis testing procedures suggested in the literature on this data set.

In the following section, we employ the ‘Bivariate Gaussian copula test’ and ‘Blanket test’ to analyze the suitability of Gaussian copula applying in semi-saturated model.

It is worth re-mentioning that the BN structure learned from the data set will not be unique. Adding/deleting different arcs from BN may provide a different suitable structure.

3.4 Gaussian copula Testing Experiment

Gaussian copula is a choice for computational convenience. Nevertheless, the
Figure 3.15: Semi-saturated BN model of case $N = 3$, after removing arcs whose absolute (conditional) rank correlation below 0.1

Gaussian copula does not always represent the data dependence structure the best as one would desire. From the D-test presented in section 3.3 we can conclude that the Gaussian copula hypothesis cannot be rejected for model 2. However it would cast doubts for the case of model 3. In order to further investigate whether the Gaussian copula fairly represents the data under given BN structure, we tried to implement Bivariate Gaussian copula test and Blanket test on model 2 and model 3.

The general steps of Bivariate test and Blanket test have been described in section 2.4. In both tests, the parameter (covariance/correlation matrix for Gaussian copula) for the assumed parametric copula is estimated from data. Under given BN structure, the correlation matrix calculated from BNs might be different from the correlation matrix estimated from data. Since we are in BN world now, the correlation matrix or parameter estimated from BNs is applied when generating the random samples.
3.4: GAUSSIAN COPULA TESTING EXPERIMENT

Figure 3.16: Validation results of semi-saturated BN model of case $N = 3$

For Blanket test, when computing the empirical copula (see equation 2.36) the $u = (u_1, \cdots, u_n)$ are randomly chosen from $[0,1]^n$, where $n$ indicates the dimension or the number of nodes in BN. This random choice would cause uncertainty to P-values. In order to weaken this uncertainty, we run the Blanket test 100 times and take the mean of the results as the final P-value.

The bootstraps for both Bivariate test and Blanket test are settled as 10000, and the significance level is 0.05.

Table 3.4 presents the Bivariate Gaussian copula test for model 2 in case $N=1$. From this table we can see that two distance measures, Kolmogorov distance $d_1$ and Anderson-Darling distance $d_3$, show that all P-values, though the other two distance measures give low P-values even zeros. The Average Kolmogorov distance $d_2$ and Average Anderson-Darling distance $d_4$ are not common use in statistics. Thereby, to some extent we can conclude that the Gaussian copula hypothesis cannot be rejected based on Bivariate Gaussian copula test.

By running the Blanket test for 100 times with $\text{bootstrap} = 10000$ for each running, and taking the mean of the results, we obtain the P-value for model 2 of case $N=1$ as: $P = 0.29$, which is bigger than significance level 0.05. Hence again, the Gaussian copula hypothesis cannot be rejected based on Blanket test.

Table 3.5 shows the Bivariate Gaussian copula test results for model 3 of case
CHAPTER III: TRAFFIC LOADS IN BRIDGES

Table 3.4: Bivariate Gaussian copula testing results for model 2 in case N=1 (10000 bootstrap steps)

<table>
<thead>
<tr>
<th></th>
<th>$\rho$</th>
<th>$d_1%$</th>
<th>$d_2%$</th>
<th>$d_3%$</th>
<th>$d_4%$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$W_f,W_s$</td>
<td>-0.64</td>
<td>97.2%</td>
<td>0.0%</td>
<td>57.5%</td>
<td>0.0%</td>
</tr>
<tr>
<td>$W_f,C_f$</td>
<td>0.77</td>
<td>92.2%</td>
<td>0.6%</td>
<td>58.8%</td>
<td>0.3%</td>
</tr>
<tr>
<td>$W_f,q_q$</td>
<td>0.49</td>
<td>91.7%</td>
<td>0.0%</td>
<td>65.8%</td>
<td>0.0%</td>
</tr>
<tr>
<td>$W_s,C_f$</td>
<td>-0.45</td>
<td>71.1%</td>
<td>0.0%</td>
<td>46.8%</td>
<td>0.1%</td>
</tr>
<tr>
<td>$W_s,q_q$</td>
<td>-0.03</td>
<td>65.2%</td>
<td>0.0%</td>
<td>55.1%</td>
<td>0.0%</td>
</tr>
<tr>
<td>$C_f,q_q$</td>
<td>0.29</td>
<td>65.4%</td>
<td>0.3%</td>
<td>62.4%</td>
<td>0.1%</td>
</tr>
</tbody>
</table>

N=3. Remember that based on D-test, the Gaussian copula assumption is rejected. However, from table 3.5 we see that the P-values calculated on the basis of $d_1$ and $d_3$ are higher than significance level 0.05, indicating that the Gaussian copula hypothesis cannot be rejected. The P-values from $d_2$ and $d_4$ columns are all zeros, which give an opposite result.

The Blanket test result for model 3 of case N=3 is: $P = 0.17$. The Gaussian copula hypothesis cannot be rejected by Blanket test, given significance level equals 0.05.

As a conclusion, for both model 2 and model 3, the Gaussian copula assumption is acceptable in capturing dependence relations on basis of given data sets.

3.5 Marginal Distribution Fitting

The copula approach separates the effect of dependence from the effect of marginal distributions in a joint distribution. The choice of margins plays an important role in determining the accuracy of inference results. The most common situation is to apply empirical or a parametric distribution to fit the observations of a variable. Empirical distribution represents the configuration of data truly, but is restricted by the samples. Any value that falls out of the samples cannot be used in the inference procedure. An alternative method is fitting a parametric distribution.

The principle behind fitting parametric distributions to data is to find the type of distribution (normal, lognormal, mixture-normal, etc) and the value of the parameters (mean, variance, etc) that resembles the data best in some sense. Usually we do not know that the data comes from any specific type of distribution, though we can often guess a family that could resemble the data. Figure 3.6 presents the
### 3.5: MARGINAL DISTRIBUTION FITTING

<table>
<thead>
<tr>
<th></th>
<th>$\hat{p}$</th>
<th>$d_1%$</th>
<th>$d_2%$</th>
<th>$d_3%$</th>
<th>$d_4%$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$W_f, C_f$</td>
<td>0.88</td>
<td>99.6%</td>
<td>0.0%</td>
<td>50.4%</td>
<td>0.0%</td>
</tr>
<tr>
<td>$W_f, W_s$</td>
<td>-0.89</td>
<td>95.6%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>$W_f, W_f$</td>
<td>-0.27</td>
<td>95.8%</td>
<td>0.0%</td>
<td>50.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>$W_f, W_s$</td>
<td>-0.23</td>
<td>97.4%</td>
<td>0.0%</td>
<td>49.2%</td>
<td>0.0%</td>
</tr>
<tr>
<td>$W_f, W_s$</td>
<td>-0.35</td>
<td>85.8%</td>
<td>0.0%</td>
<td>44.1%</td>
<td>0.0%</td>
</tr>
<tr>
<td>$W_f, q$</td>
<td>-0.3</td>
<td>84.8%</td>
<td>0.0%</td>
<td>43.9%</td>
<td>0.0%</td>
</tr>
<tr>
<td>$C_f, W_f$</td>
<td>0.34</td>
<td>99.9%</td>
<td>0.0%</td>
<td>52.4%</td>
<td>0.0%</td>
</tr>
<tr>
<td>$C_f, W_s$</td>
<td>-0.79</td>
<td>98.9%</td>
<td>0.0%</td>
<td>53.3%</td>
<td>0.0%</td>
</tr>
<tr>
<td>$C_f, W_f$</td>
<td>-0.24</td>
<td>97.1%</td>
<td>0.0%</td>
<td>48.6%</td>
<td>0.0%</td>
</tr>
<tr>
<td>$C_f, W_f$</td>
<td>-0.21</td>
<td>98.5%</td>
<td>0.0%</td>
<td>50.8%</td>
<td>0.0%</td>
</tr>
<tr>
<td>$C_f, W_s$</td>
<td>-0.31</td>
<td>92.3%</td>
<td>0.0%</td>
<td>43.1%</td>
<td>0.0%</td>
</tr>
<tr>
<td>$C_f, W_s$</td>
<td>-0.27</td>
<td>87.2%</td>
<td>0.0%</td>
<td>44.2%</td>
<td>0.0%</td>
</tr>
<tr>
<td>$C_f, q$</td>
<td>0.25</td>
<td>98.9%</td>
<td>0.0%</td>
<td>53.3%</td>
<td>0.0%</td>
</tr>
<tr>
<td>$W_s, W_f$</td>
<td>0.12</td>
<td>93.5%</td>
<td>0.0%</td>
<td>46.5%</td>
<td>0.0%</td>
</tr>
<tr>
<td>$W_s, W_f$</td>
<td>0.21</td>
<td>95.5%</td>
<td>0.0%</td>
<td>48.4%</td>
<td>0.0%</td>
</tr>
<tr>
<td>$W_s, W_s$</td>
<td>0.28</td>
<td>97.8%</td>
<td>0.0%</td>
<td>45.6%</td>
<td>0.0%</td>
</tr>
<tr>
<td>$W_s, q$</td>
<td>0.23</td>
<td>84.2%</td>
<td>0.0%</td>
<td>44.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>$W_s, q$</td>
<td>-0.03</td>
<td>99.3%</td>
<td>0.0%</td>
<td>50.1%</td>
<td>0.0%</td>
</tr>
<tr>
<td>$W_f, W_f$</td>
<td>0.06</td>
<td>97.9%</td>
<td>0.0%</td>
<td>49.6%</td>
<td>0.0%</td>
</tr>
<tr>
<td>$W_f, W_s$</td>
<td>0.12</td>
<td>85.6%</td>
<td>0.0%</td>
<td>41.8%</td>
<td>0.0%</td>
</tr>
<tr>
<td>$W_f, W_s$</td>
<td>0.11</td>
<td>83.6%</td>
<td>0.0%</td>
<td>44.1%</td>
<td>0.0%</td>
</tr>
<tr>
<td>$W_f, q$</td>
<td>0.07</td>
<td>95.7%</td>
<td>0.0%</td>
<td>48.4%</td>
<td>0.0%</td>
</tr>
<tr>
<td>$W_f, q$</td>
<td>0.08</td>
<td>88.9%</td>
<td>0.0%</td>
<td>43.6%</td>
<td>0.0%</td>
</tr>
<tr>
<td>$W_f, W_f$</td>
<td>0.07</td>
<td>83.1%</td>
<td>0.0%</td>
<td>43.9%</td>
<td>0.0%</td>
</tr>
<tr>
<td>$W_f, q$</td>
<td>0.23</td>
<td>96.9%</td>
<td>0.0%</td>
<td>50.4%</td>
<td>0.0%</td>
</tr>
<tr>
<td>$W_s, W_s$</td>
<td>0.27</td>
<td>80.4%</td>
<td>0.0%</td>
<td>42.4%</td>
<td>0.0%</td>
</tr>
<tr>
<td>$W_s, q$</td>
<td>0.09</td>
<td>84.5%</td>
<td>0.0%</td>
<td>44.4%</td>
<td>0.0%</td>
</tr>
<tr>
<td>$W_s, q$</td>
<td>0.13</td>
<td>82.7%</td>
<td>0.0%</td>
<td>42.6%</td>
<td>0.0%</td>
</tr>
</tbody>
</table>

Table 3.5: Bivariate Gaussian copula testing results for model 3 in case $N=3$ (10000 bootstrap steps)

 histograms of variables $W_f$, $W_s$, $C_f$ and $q$ from case $N = 1$. One notices that each distribution will not be characterized fully by a Gaussian or normal parametric distribution. As shown in figure 3.17, the fitted normal to $W_f$ does not reflect the high density in interval $[1000,1200]$ of $W_f$, while the fitted distribution to $W_s$ fails to capture the high mass at zero. Graphically the density of, especially $W_s$, has more than one peaks/modes and a huge mass at zero. In order to reflect these special features, mixture-normal distribution is a good choice to fit the data.

Mathematically, a mixture distribution is the probability distribution of a random variable whose values can be interpreted as being derived in a simple way
(a) Density fitting

(b) Cumulative distribution fitting

Figure 3.17: Normal distribution fitting to variables Wf and Ws
(usually linear combination) from an underlying set of other random variables. In case where each of the underlying random variables follows a normal distribution, it leads to a mixture-normal distribution. A mixture-normal distribution is a convex combination, i.e. a weighted sum with non-negative weights that sum to 1, of normal distributions who have different parameters (mean and standard deviation). The individual distributions combined to form the mixture distribution are called the mixture components, and the weights associated with each component are named the mixture weights. The number of components is restricted to being finite.

Suppose the density functions of each component are know as \( \phi(\mu_1, \sigma_1), \ldots, \phi(\mu_n, \sigma_n) \), the cumulative distributions as \( \Phi(\mu_1, \sigma_1), \ldots, \Phi(\mu_n, \sigma_n) \), the mixture-normal distribution can be represented by writing either density \( \phi_m \) or the cumulative distribution \( \Phi_m \):

\[
\phi_m = \sum_{i=1}^{n} w_i \phi(\mu_i, \sigma_i) \quad (3.3)
\]

\[
\Phi_m = \sum_{i=1}^{n} w_i \Phi(\mu_i, \sigma_i) \quad (3.4)
\]

where \( w_1, \ldots, w_n \) are mixture weights such that \( w_i \geq 0 \) and \( \sum w_i = 1 \). The mean and standard deviation parameters of each component refers to \( \mu_i \) and \( \sigma_i \) correspondingly.

Fitting a mixture-normal model to data we try fitting with a specified number of components to data by using \texttt{gmdistribution} method in Matlab. This method uses an expectation maximization (EM) algorithm, which assigns posterior probabilities\(^7\) to each component density with respect to each observation, and employs an iterative algorithm that converges to a local optimum. After fitting successfully, it returns a set of parameters including components’ mean, standard deviation and weights. The negative log-likelihood (NlogL) is also computed. We use these negative log-likelihoods (NlogL) for data from a mixture-normal fitting to determine an "appropriate" number of components for a model based on the criteria that lower NlogL implies better fitting.

\(^7\)The posterior probabilities for each point indicate that each data point has some probability of belonging to each subpopulation
For the purpose of saving computation time, we bound the number of components between 2 and 25 in experiment. Figure 3.18 shows the NlogL value given by gmdistribution mixture-normal fitting method, with respect to different choices of the number of the components. The blue points indicate the minimal value of NlogL. This will be our choice for the number of components. Apparently, when fitting mixture-normal model to data, the selected number of the components for variable Wf is 24, 24 for Ws, 12 for Cf and 19 for qq. However, it is worth mentioning that as the components number bound increases, the selected number might change.

**Figure 3.18:** NlogL line with respect to the number of components, for variables Wf, Ws, Cf and qq

**Figure 3.19:** Goodness-of-fitting comparison between normal fitting and mixture-normal fitting to Wf, by compare survival function
Figure 3.19 presents the goodness-of-fitting comparison between normal fitting and the 24-mixed (optimal number for Wf) normal fitting to Wf. The survival function of mixed-normal fitting distribution follows the trace of the empirical distribution all the way. It significantly improves the fitting accuracy, compared to the normal distribution. In the following section, we use the mixture-normal model to the margins with the selected number of components.

<table>
<thead>
<tr>
<th>Variables</th>
<th>Wf2</th>
<th>Cf2</th>
<th>Ws2</th>
<th>Wf1</th>
<th>Wf3</th>
<th>Ws1</th>
<th>Ws3</th>
<th>qq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Optimal number</td>
<td>24</td>
<td>19</td>
<td>24</td>
<td>25</td>
<td>21</td>
<td>23</td>
<td>24</td>
<td>19</td>
</tr>
</tbody>
</table>

Table 3.6: Optimal number of components for variables in case $N = 3$, chosen from interval $[2, 25]$

### 3.6 Inference

So far we have developed the CNBN models for data set of case $N = 1$ and $N = 3$. This data set is simulated from WIM model with parameters $L = 100(m)$, $r(S_t, S_{t-1}) = 0.2$ and sample size 2160, and processed by categorizing traffic loads (in terms of weights) and traffic density (in terms of the number of axles) into $N$ sections of the hypothetical bridge. The CNBN model involves probabilistic nodes whose marginal distribution can be fitted by parametric distributions. The influences between variables are quantified from data. Given the static CNBN model, we are ready to make inferences. Section 3.6.1 describes the comparison of influence results between different fitted margin but using the same model structure.

The separations between vehicles are important during the simulation process. This is due to the fact that, one may speculate that the closer the vehicles are to each other, the larger the bending moment of the bridge will be in the middle of the span [17]. The coefficient $r$ denotes the parameter of conditional copula which describes the dependence structure between the transition distribution of $\{S_t\}$, as equation 3.1 shows. $S_t$ denotes the time series for separation. In section 3.6.2 we investigate the models with different time separation coefficients $r = 0.2, 0.3, 0.4, 0.8$ and bridge length parameters $L = 100(m), 50(m)$.

#### 3.6.1 One bridge section case
A quantified CNBN provides a compact representation of a high dimensional probability distribution over the set of variables. The key feature of CNBN encoded with Gaussian copula is its efficient inference. Generally, influence in BNs world is the process of deriving a conditional probability distribution given known information.

From the physical point of view, bridge bending moment exists when a solid object is subjected to forces that make it bend, in our research the forces come from traffic loads. As section 3.2.3 states, we are aiming to achieve the distribution of bending moment with the evidence of configuration of traffic loads and axles, and the other way around. We investigate in two models, one for case $N = 1$ and the other for case $N = 3$, as described in section 3.3.2. Hence two scenarios are taken into consideration:

1. Get the configuration of weights and number of axles, given $qq$ is known;
2. Get the distribution of $qq$, given one or more weights and number axles variables are known.

![Figure 3.20: Inference results implemented in UNINET, given $qq = 28(kNm)$, case $N = 1$](image)

We construct the saturated BN model, in which all variables are dependent with each other, for case $N = 1$. In section 3.4 we obtain both positive Gaussian copula and model validation results, hence this model represents the data reasonably. The
values associated with arcs are (conditional) rank correlations learned from data using UNINET. Suppose we got the measure of \( qq \) as \( 28(kNm) \) from observations, we update this information in the static BN model. Figure 3.7 presents the inference results of \( W_f, W_s \) and \( C_f \) given evidence \( qq = 28(kNm) \), implemented in UNINET. Note that the evidence of \( qq \) locates closely to its right bound, as shown in the lower left histogram of \( qq \). Within the boxes with red line edges, the gray bins represent variables’ original margins (unconditional univariate distributions), while the dark black bins stand for the distribution after information updating. One can notice that the distribution of \( W_f \) and \( C_f \) move to the right significantly, specially indicated by the increase of the mean values shown at the bottom of the boxes. The change to \( W_s \) cannot be observed directly from the histograms, but we can see the decrease in mean, from \( 152(kN) \) to \( 144(kN) \). Since the number of vehicles per day is a constant during data simulation procedure (see section 3.2.1), the increase of \( C_f \) results in a decrease of \( C_s \), thereby here comes the decrease of \( W_s \). Also, the evidence reduces the uncertainties of \( W_f, W_s \) and \( C_f \), as the more information we gain, the better we recognize the event.

As explained previously, we are interested in conditionalizing on out of sample values. Hence it is needed to fit a parametric margin to data. In order to get an idea of the effect of using such margin, we first investigate "in sample" values. Table 3.7 presents the influence results of models with different margin fitting approach. The mean values of the conditional distribution are employed as the prediction approximation. To compare the inference accuracy, we treat the prediction results from UNINET as the benchmark, and define the relative difference \( Rd_N \) for normal margin fitting model, \( Rd_M \) for mixture-normal margin fitting model, as:

\[
Rd_i = \frac{|Measure_i - Measure_U|}{Measure_U} \times 100
\]

where \( i \) refers to Normal (\( Rd_N \)) or Mixed Normal (\( Rd_M \)). The measures can be mean, median, standard deviation etc., and the subscripts \( N \) refers to model with normal margin fitting, \( M \) to mixture-normal margin fitting model, \( U \) to the empirical margin model implemented in UNINET.

We can notice from the table 3.7 that most of \( Rd_M \) are smaller than \( Rd_N \) except
the $Rd_M$ for the measure of mean of $Wf$. It might be the over-fitting or the fat tails which results in the worse performance of mixture-normal model in $Wf$ prediction. Nevertheless, from the analysis of $Rd$ values we can conclude that the mixture-normal model seems to be a more suitable model in performing prediction, compared to normal margin fitting model in general. Surely it also depends on the real needs of application, if one specifically want to get an accurate prediction of $Wf$ rather than others, normal margin fitting model might be more appropriate.

Recall that one of the advantages of fitting margins with parametric distributions is the probability to observe "out of sample" values. Any value who falls out of the corresponding variable’s bounds cannot be updated in UNINET, thus for instance the extreme scenario as $qq = 30(kNm)$ can be investigated with a suitable parametric margin fitting models but not in UNINET. Information updating on extreme values, specially extremely large values in traffic loads investigation on bridge, is of great interests. As a reference, we would like to update the evidence of $qq = 30(kNm)$ to see how the $Wf$, $Ws$ and $Cf$ will be.

### Table 3.7: The comparison of inference results given $qq = 28(kNm)$ for case $N = 1$, among different margin fitting models

<table>
<thead>
<tr>
<th></th>
<th>UNINET</th>
<th>Normal</th>
<th>Mixture-normal</th>
<th>$Rd_N$</th>
<th>$Rd_M$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Wf$-mean</td>
<td>1392.7</td>
<td>1254.8</td>
<td>1152.4</td>
<td>9.9</td>
<td>17.2</td>
</tr>
<tr>
<td>$Wf$-std</td>
<td>206</td>
<td>260.5</td>
<td>230.9</td>
<td>26.5</td>
<td>12.1</td>
</tr>
<tr>
<td>$Ws$-mean</td>
<td>144.3</td>
<td>136.8</td>
<td>142.9</td>
<td>5.2</td>
<td>1</td>
</tr>
<tr>
<td>$Ws$-std</td>
<td>265</td>
<td>271.2</td>
<td>262.3</td>
<td>2.3</td>
<td>1</td>
</tr>
<tr>
<td>$Cf$-mean</td>
<td>13.1</td>
<td>14.4</td>
<td>12.9</td>
<td>10</td>
<td>1.5</td>
</tr>
<tr>
<td>$Cf$-std</td>
<td>3.5</td>
<td>2.8</td>
<td>3.5</td>
<td>20</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 3.7 presents measures of interests for variables $Wf$, $Ws$ and $Cf$, respectively. The $Rd_N$ and $Rd_M$ are the range of the $Wf$ in normal and mixture-normal distributions, respectively.
conditionalized on $qq = 30(kNm)$, where the slash represent unmeasurable quantities. We expect to find out the basic rule that higher traffic loads result in bigger bending moment. Comparing the inference results in tables 3.8 and 3.7, we conclude that the mixture-normal margin fitting model confirms the basic rule but the normal margin fitting model does not, since all statistical measures in table 3.8 decrease with respect to those in table 3.7. This reinforces the idea of applying mixture-normal margin fitting to the one dimensional margins rather than normal fitting.

Another interesting representation for practitioners is the prediction of bending moment given known traffic loads and distribution. This is of interests because it is meaningful in applications such as the design of new bridges, the fatigue life assessments of existing bridges and the control of traffic flow over the bridge. The bridge needs to be redesigned or maintained, or the traffic needs to be controlled, if the conditionalized bending moment tends to be close or even bigger than the tolerated maximal bending moment of the bridge. As we stated in section 3.2.3, we introduce the evidence of $W_f = 1835.8(kN)$, $W_s = 1440(kN)$ and $C_f = 25$ to predict the bending moment $qq$, as an instance.

From figure 3.21 one notices that all evidence reach the right bound of the corresponding variables, resulting in the maximal value of $qq = 28.2$ with no uncertainty. Not only the evidence but also the prediction is not allowed to fall out of the values observed in the sample even when we are supposed to obtain a higher than 28.2 as prediction of qq given above mentioned evidence. The prediction results of mixture-normal and normal margin fitting models are presented in table 3.21:

<table>
<thead>
<tr>
<th></th>
<th>UNINET</th>
<th>Normal</th>
<th>Mixture-normal</th>
<th>$Rd_N$</th>
<th>$Rd_M$</th>
</tr>
</thead>
<tbody>
<tr>
<td>qq-mean</td>
<td>28.2</td>
<td>23.5</td>
<td>25.4</td>
<td>16.7</td>
<td>10</td>
</tr>
<tr>
<td>qq-std</td>
<td>0</td>
<td>1.6</td>
<td>2.8</td>
<td>inf</td>
<td>inf</td>
</tr>
</tbody>
</table>

**Table 3.9:** Inference results given $W_f = 1835.8(kN)$, $W_s = 1440(kN)$ and $C_f = 25$, for case $N = 1$

Apparently both normal and mixture-normal margin fitting models underestimated the bending moment. When the empirical margins are used, the cumulative probability of these extreme large values reaches closer to 1 than the case of in normal and mixed-normal, thus the normal inverse used in conditionalization are
CHAPTER III : TRAFFIC LOADS IN BRIDGES

Figure 3.21: Inference results implemented in UNINET, given $W_f = 1835\,(kN)$, $W_s = 1440\,(kN)$ and $C_f = 25$, for case $N = 1$

larger which eventually results in a larger approximation of $qq$. Getting rid of over-fitting or underestimation could be a direction to go further in future research.

3.6.2 Three bridge sections case

In the following we come to a more complicated model, case $N = 3$. In case $N = 3$, the bridge is divided into 3 sections, and the number of axles and axle loads are categorized into each section based on their positions. Variables $W_f^1, W_f^2, W_f^3, W_s^1, W_s^2, W_s^3, C_f^1, C_f^2, C_f^3, C_s^1, C_s^2, C_s^3$ and $qq$ are generated, more details about data processing see section 3.2.2. By learning the structure from data, the dimension of model reduces to 8 rather than the original 13 variables. For ease of read we will repeat the notations. Recall the definitions we have for variables in case $N = 3$:

In the semi-saturated model for case $N = 3$, as shown in figure 3.15, only 8 variables remain in the BN model. The reason behind is that according to the simulation described in section 3.2.1, the maximal bending moment is most of the
### Table 3.10: Notations and definitions of variables involved in case $N = 3$

<table>
<thead>
<tr>
<th>Name of variables</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$W_{f1}$ (kN)</td>
<td>the total weights of axles in the first section of the fast lane</td>
</tr>
<tr>
<td>$W_{f2}$ (kN)</td>
<td>the total weights of axles in the second section of the fast lane</td>
</tr>
<tr>
<td>$W_{f3}$ (kN)</td>
<td>the total weights of axles in the third section of the fast lane</td>
</tr>
<tr>
<td>$W_{s1}$ (kN)</td>
<td>the total weights of axles in the first section of the slow lane</td>
</tr>
<tr>
<td>$W_{s2}$ (kN)</td>
<td>the total weights of axles in the second section of the slow lane</td>
</tr>
<tr>
<td>$W_{s3}$ (kN)</td>
<td>the total weights of axles in the third section of the slow lane</td>
</tr>
<tr>
<td>$C_{f1}$</td>
<td>the total number of axles in the first section of the fast lane</td>
</tr>
<tr>
<td>$C_{f2}$</td>
<td>the total number of axles in the second section of the fast lane</td>
</tr>
<tr>
<td>$C_{f3}$</td>
<td>the total number of axles in the third section of the fast lane</td>
</tr>
<tr>
<td>$C_{s1}$</td>
<td>the total number of axles in the first section of the slow lane</td>
</tr>
<tr>
<td>$C_{s2}$</td>
<td>the total number of axles in the second section of the slow lane</td>
</tr>
<tr>
<td>$C_{s3}$</td>
<td>the total number of axles in the third section of the slow lane</td>
</tr>
<tr>
<td>$qq$ (kNm)</td>
<td>the maximal bending moment of the whole bridge</td>
</tr>
</tbody>
</table>

Times realized by axles positioned in the mid section of the fast lane of the bridge. For inference instances, three situations are considered:

- Conduct inference given $qq = 28(kNm)$
- Conduct inference given $qq = 30(kNm)$
- Conduct inference given $W_{f2} = 1369(kN)$

The inference results implemented in UNINET is shown in figure 3.11, where the histograms outside of the grey circle are unconditional ones, while those inside are conditional distributions. One can notice that all variables move to the right after conditionalization on $qq = 28(kNm)$, except $W_{f2}$. This is due to the choice of the right bound of $qq$. Physically it means that at the middle of the span in the fast lane, the traffic loads and bending moments have positive correlation. The reason for the decrease of $W_{s2}$ is from the fixed number of vehicles per day, thereby a certain bound of the total traffic loads, as the average loads of individual axle falls in a bounded interval (see figure 3.7).

Table 3.11 presents the statistical approximation mean and standard deviation of each variables’ conditional distribution given $qq = 28(kNm)$. The comparison inference
Figure 3.22: Inference results implemented in UNINET, given $qq = 28(kNm)$ for case $N = 3$

results between mixture-normal model and normal model are shown in figure 3.23. Here and forth, mixture-normal model refers to the BN model with mixture-normal margin fitting, and normal model refers to the BN model with normal margin fitting. As shown in figure 3.23(d), all red diamonds and stars positions are above the blue ones except for variable $W_s2$. It indicates that mixture-normal model performs better than normal model for most variables of interest. The choice of model depends on practitioners’ needs, and also one may need to investigate other statistical measures such as median, mode etc..

When it comes to updating the extreme case $qq = 30(kNm)$ information in BN
3.6 : INFERENCES

(a) Mean of three models

(b) Standard deviation of three models

(c) Error bar of three models, error refers to standard deviation

(d) Relative difference of mixture-normal model and normal model

Figure 3.23: Inference results comparison between mixture-normal model and normal model, given $qq = 28(kNm)$, $N = 3$, with UNINET result as the benchmark
CHAPTER III : TRAFFIC LOADS IN BRIDGES

<table>
<thead>
<tr>
<th></th>
<th>UNINET</th>
<th>Normal</th>
<th>Mixture-normal</th>
<th>$Rd_N$</th>
<th>$Rd_M$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wf2-mean</td>
<td>1159.8</td>
<td>1583.5</td>
<td>1151.7</td>
<td>36.5</td>
<td>0.7</td>
</tr>
<tr>
<td>Wf2-std</td>
<td>85.4</td>
<td>228.9</td>
<td>76</td>
<td>168.1</td>
<td>11</td>
</tr>
<tr>
<td>Cf2-mean</td>
<td>9.9</td>
<td>11.1</td>
<td>10.3</td>
<td>12.5</td>
<td>3.6</td>
</tr>
<tr>
<td>Cf2-std</td>
<td>1.5</td>
<td>2</td>
<td>2.1</td>
<td>32.3</td>
<td>35.7</td>
</tr>
<tr>
<td>Ws2-mean</td>
<td>92.3</td>
<td>93.7</td>
<td>100.5</td>
<td>1.5</td>
<td>8.9</td>
</tr>
<tr>
<td>Ws2-std</td>
<td>195</td>
<td>211.1</td>
<td>195.2</td>
<td>8.3</td>
<td>0.01</td>
</tr>
<tr>
<td>Wf1-mean</td>
<td>62.5</td>
<td>75.5</td>
<td>68.4</td>
<td>20.8</td>
<td>9.4</td>
</tr>
<tr>
<td>Wf1-std</td>
<td>127</td>
<td>113.1</td>
<td>133.3</td>
<td>11</td>
<td>5</td>
</tr>
<tr>
<td>Wf3-mean</td>
<td>88.3</td>
<td>121.2</td>
<td>94.3</td>
<td>37.2</td>
<td>6.8</td>
</tr>
<tr>
<td>Wf3-std</td>
<td>148</td>
<td>115.6</td>
<td>148.5</td>
<td>21.9</td>
<td>0.3</td>
</tr>
<tr>
<td>Ws1-mean</td>
<td>25.7</td>
<td>30.4</td>
<td>19.2</td>
<td>18.4</td>
<td>25.2</td>
</tr>
<tr>
<td>Ws1-std</td>
<td>82.7</td>
<td>71.5</td>
<td>69</td>
<td>13.5</td>
<td>16.7</td>
</tr>
<tr>
<td>Ws3-mean</td>
<td>32.1</td>
<td>38.9</td>
<td>34.9</td>
<td>21.1</td>
<td>8.9</td>
</tr>
<tr>
<td>Ws3-std</td>
<td>95.2</td>
<td>74.7</td>
<td>100.4</td>
<td>21.5</td>
<td>5.5</td>
</tr>
</tbody>
</table>

Table 3.11: Inference results given $qq = 28(kNm)$, for case $N = 3$

model, we expect to see an increase more or less of the mean of those conditional variables, compare to those in case of conditionalizing on $qq = 28(kNm)$. Since $qq = 30(kNm)$ is out of interval boundary of $qq$ in this data set, inference is not allowed to implemented in UNINET, thus both UNINET results and relative difference are missing, as shown in table 3.12. Based on mixture-normal model, note that the mean of $Ws2$ distribution decreases with respect to table 3.11 from the $Ws2$ (conditionalizing on $qq = 28(kNm)$), but all others increase as we expected. Consequently, we can conclude that $Wf2$ plays a more important role in determining bending moment than $Ws2$. Given this information, in real application such as designers should take this conclusion into consideration, and for traffic control purpose, one should pay more attention to control the traffic flow in the middle of the span in fast lane, and so on.

One of the nice features of BN model is its flexibility in inference tasks, for instance diagnostic task and prediction task. If the reasoning is done "bottom up" (in terms of the directionality of arcs), the BN is used for diagnosis, whereas if it is done "top-down", the BN serves for prediction[5]. In case $N = 1$ both diagnosis and prediction are accomplished by taking information updating on basis of known information $qq$ (diagnosis), and $Wf$ (prediction). Table 3.13 presents the prediction results given $Wf2 = 1369(kN)$. There appears negative numbers in the mean
Table 3.12: Inference results given \( qg = 30(kNm) \), for case \( N = 3 \)

<table>
<thead>
<tr>
<th></th>
<th>UNINET</th>
<th>Normal</th>
<th>Mixture-normal</th>
<th>( R_d_N )</th>
<th>( R_d_M )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wf2-mean</td>
<td>/</td>
<td>1539.8</td>
<td>1223.3</td>
<td>/</td>
<td>/</td>
</tr>
<tr>
<td>Wf2-std</td>
<td>/</td>
<td>225.5</td>
<td>42.3</td>
<td>/</td>
<td>/</td>
</tr>
<tr>
<td>Cf2-mean</td>
<td>/</td>
<td>11</td>
<td>11.5</td>
<td>/</td>
<td>/</td>
</tr>
<tr>
<td>Cf2-std</td>
<td>/</td>
<td>2</td>
<td>2</td>
<td>/</td>
<td>/</td>
</tr>
<tr>
<td>Ws2-mean</td>
<td>/</td>
<td>94.2</td>
<td>96</td>
<td>/</td>
<td>/</td>
</tr>
<tr>
<td>Ws2-std</td>
<td>/</td>
<td>201.5</td>
<td>204.4</td>
<td>/</td>
<td>/</td>
</tr>
<tr>
<td>Wf1-mean</td>
<td>/</td>
<td>78.4</td>
<td>83.4</td>
<td>/</td>
<td>/</td>
</tr>
<tr>
<td>Wf1-std</td>
<td>/</td>
<td>114.1</td>
<td>145.9</td>
<td>/</td>
<td>/</td>
</tr>
<tr>
<td>Wf3-mean</td>
<td>/</td>
<td>119.7</td>
<td>127.9</td>
<td>/</td>
<td>/</td>
</tr>
<tr>
<td>Wf3-std</td>
<td>/</td>
<td>121.9</td>
<td>173.5</td>
<td>/</td>
<td>/</td>
</tr>
<tr>
<td>Ws1-mean</td>
<td>/</td>
<td>32.8</td>
<td>32.4</td>
<td>/</td>
<td>/</td>
</tr>
<tr>
<td>Ws1-std</td>
<td>/</td>
<td>72.4</td>
<td>94.1</td>
<td>/</td>
<td>/</td>
</tr>
<tr>
<td>Ws3-mean</td>
<td>/</td>
<td>34.2</td>
<td>30.8</td>
<td>/</td>
<td>/</td>
</tr>
<tr>
<td>Ws3-std</td>
<td>/</td>
<td>77.1</td>
<td>93.7</td>
<td>/</td>
<td>/</td>
</tr>
</tbody>
</table>

prediction of \( Ws_2, Wf_1, Ws_1 \) and \( Ws_3 \), which are unrealistic since both traffic loads and number of axles cannot be negative in reality. The negative mean prediction comes from the fact that the normal and mixture-normal margins may take values in the whole real line. The (conditional) dependence among variables makes it impossible to correct those negative means to zero without changing other variables’ mean values. But for inference in UNINET, mean prediction of \( Ws_2 \) is zero, for the reason that the prediction cannot fall out of corresponding variable’s interval and, in this data set the minimal value of \( Ws_2 \) is zero. An alternative way is to treat variables with negative means as zero, and then re-conditionalize on these known information.

We have already concluded that mixture-normal model shows more accuracy prediction and diagnosis than normal model on the basis of previous case analysis. In table 3.13, only one variable \( Ws_2 \)’s mean prediction turns out to be negative in mixture-normal model, thus we re-conditionalise on \( Ws_2 = 0(kN) \). Results are presented in table 3.14.

Table 3.14 provides the information of mean prediction and diagnosis, conditionalized on \( Wf_2 = 1369(kN) \) and \( Ws_2 = 0(kN) \), and the summarised version of these information are shown in figure 3.24. Note that there still exists unrealistic negative numbers in predictions of normal model, but not in mixture-normal model results any more. The mixture-normal model keeps its higher accuracy quality in
(a) Mean of three models
(b) Standard deviation of three models
(c) Error bar of three models, error refers to standard deviation
(d) Relative difference of mixture-normal model and normal model

Figure 3.24: Inference results comparison between mixture-normal model and normal model, given $W_f 2 = 1369(kN)$ and $W_s 2 = 0(kN)$, with UNINET result as the benchmark.
Table 3.13: Inference results given $Wf_2 = 1369(kNm)$, for case $N = 3$

<table>
<thead>
<tr>
<th></th>
<th>UNINET</th>
<th>Normal</th>
<th>Mixture-normal</th>
<th>$Rd_N$</th>
<th>$Rd_M$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cf2-mean</td>
<td>13.3</td>
<td>11</td>
<td>14.4</td>
<td>17.4</td>
<td>8.3</td>
</tr>
<tr>
<td>Cf2-std</td>
<td>0.9</td>
<td>1.5</td>
<td>1.3</td>
<td>61.6</td>
<td>43.3</td>
</tr>
<tr>
<td>Ws2-mean</td>
<td>0</td>
<td>-98.2</td>
<td>-2.1</td>
<td>/</td>
<td>/</td>
</tr>
<tr>
<td>Ws2-std</td>
<td>0</td>
<td>177</td>
<td>0.9</td>
<td>/</td>
<td>/</td>
</tr>
<tr>
<td>Wf1-mean</td>
<td>2.95</td>
<td>-4.6</td>
<td>6.3</td>
<td>256.6</td>
<td>112.6</td>
</tr>
<tr>
<td>Wf1-std</td>
<td>26.2</td>
<td>108.5</td>
<td>42.1</td>
<td>314.3</td>
<td>60.6</td>
</tr>
<tr>
<td>Wf3-mean</td>
<td>5.4</td>
<td>16.1</td>
<td>9.1</td>
<td>197.4</td>
<td>67.7</td>
</tr>
<tr>
<td>Wf3-std</td>
<td>35.6</td>
<td>114</td>
<td>49.1</td>
<td>220.2</td>
<td>37.8</td>
</tr>
<tr>
<td>Ws1-mean</td>
<td>0.6</td>
<td>-11.4</td>
<td>0.2</td>
<td>1997</td>
<td>66.6</td>
</tr>
<tr>
<td>Ws1-std</td>
<td>11.2</td>
<td>72.5</td>
<td>14.4</td>
<td>547.3</td>
<td>28.6</td>
</tr>
<tr>
<td>Ws3-mean</td>
<td>1.1</td>
<td>-7.7</td>
<td>1.5</td>
<td>799.4</td>
<td>37.8</td>
</tr>
<tr>
<td>Ws3-std</td>
<td>15.7</td>
<td>72.2</td>
<td>25.9</td>
<td>360</td>
<td>64.9</td>
</tr>
<tr>
<td>qq-mean</td>
<td>23.8</td>
<td>21.1</td>
<td>24</td>
<td>11.5</td>
<td>0.9</td>
</tr>
<tr>
<td>qq-std</td>
<td>2.4</td>
<td>1.4</td>
<td>2.5</td>
<td>40.1</td>
<td>2.3</td>
</tr>
</tbody>
</table>

In conclusion, mixture-normal distribution is a suitable margin fitting choice because of its acceptable accuracy in prediction and diagnosis, with UNINET results as the benchmark. Furthermore, re-conditionalization needs to be employed when unrealistic values appear in statistical measures after information updating. Again all above analysis is conducted on the basis of data described at the beginning of section 3.6. For different data set, the BN model structure and the choice of margin fitting distribution may change as well, but, the general idea for the mode is the same.

3.6 Investigation of data sets generated with different parameters

The assessment of extreme traffic load effects induced in bridges typically requires two models as suggested by (Allaix 2007). The first model called traffic load model concerns the properties of the single vehicle, while the second model named as traffic flow model, describes the parameters of the traffic circulation, like the velocity, the inter-convoy distance and the intensity of the vehicle passages (see more details in
<table>
<thead>
<tr>
<th></th>
<th>UNINET</th>
<th>Normal</th>
<th>Mixture-normal</th>
<th>$Rd_N$</th>
<th>$Rd_M$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cf2-mean</td>
<td>13.4</td>
<td>10.7</td>
<td>13.7</td>
<td>20</td>
<td>2</td>
</tr>
<tr>
<td>Cf2-std</td>
<td>0.8</td>
<td>1.4</td>
<td>1.3</td>
<td>69.1</td>
<td>58.3</td>
</tr>
<tr>
<td>Wf1-mean</td>
<td>3.2</td>
<td>-4</td>
<td>1.5</td>
<td>222.8</td>
<td>54.3</td>
</tr>
<tr>
<td>Wf1-std</td>
<td>27.6</td>
<td>110.8</td>
<td>21.8</td>
<td>301.4</td>
<td>21.2</td>
</tr>
<tr>
<td>Wf3-mean</td>
<td>5.4</td>
<td>7.7</td>
<td>5.8</td>
<td>42.9</td>
<td>7.2</td>
</tr>
<tr>
<td>Wf3-std</td>
<td>35.6</td>
<td>117.6</td>
<td>39</td>
<td>230.5</td>
<td>9.6</td>
</tr>
<tr>
<td>Ws1-mean</td>
<td>0.4</td>
<td>-8.3</td>
<td>1.1</td>
<td>2113.8</td>
<td>170.1</td>
</tr>
<tr>
<td>Ws1-std</td>
<td>8.9</td>
<td>71.7</td>
<td>22.8</td>
<td>703.5</td>
<td>155.6</td>
</tr>
<tr>
<td>Ws3-mean</td>
<td>0.8</td>
<td>-9.8</td>
<td>0.9</td>
<td>1295</td>
<td>9.3</td>
</tr>
<tr>
<td>Ws3-std</td>
<td>13.6</td>
<td>74.1</td>
<td>21.8</td>
<td>445</td>
<td>60.3</td>
</tr>
<tr>
<td>qq-mean</td>
<td>23.1</td>
<td>21.3</td>
<td>25.9</td>
<td>7.9</td>
<td>12.1</td>
</tr>
<tr>
<td>qq-std</td>
<td>2.1</td>
<td>1.4</td>
<td>2.4</td>
<td>33.9</td>
<td>16.6</td>
</tr>
</tbody>
</table>

Table 3.14: Inference results given $Wf_2 = 11369(kNm)$, $Ws_2 = 0$, for case $N = 3$

The parameters are estimated from the real recorded data, and further employed in simulating data corresponding to the same location where the original data is recorded.

These parameters, such as length of bridge, auto correlation coefficients $r(S_t, S_{t-1})$, will eventually influence the loads distribution and the maximal bending moment per day ($qq$). For instance, high auto correlation $r(S_t, S_{t-1})$ reflects the high intensity of vehicles laying on bridge. Furthermore, the closer the vehicles are to each other, the larger the bending moment of the bridge will be in the middle of the span, theoretically. In this section we are aiming to investigate the performance of $qq$, with respect to different parameters, the length of bridge $L$ and the auto correlation $r(S_t, S_{t-1})$.

The data sets shown in table 3.15 are simulated on the basis of the same location $RW16 – L$, and share the same average number of vehicles per day in both lanes 7937. Nevertheless, they are generated with different bridge length $L$ and auto correlation $r(S_t, S_{t-1})$. The names of the data sets contain the information of bridge length and auto correlation parameters. For instance “DataL50r0.2” means this data set is generated on the basis of 50 meters long bridge and the auto correlation $r(S_t, S_{t-1}) = 0.2$. The same idea applies to the other data sets. From table 3.15 we see that generally the increase in auto correlation results in an increase of bending moment $qq$, and 100 meters long bridge
3.6 : INVESTIGATION OF DATA SETS GENERATED WITH DIFFERENT PARAMETERS

has significantly lower bending moment than 50 meters long bridge. Note that data set “DataL50r0.2” turns out to be unusual as it gives outstanding value of maximal \( qq \). The data set “DataL50r0.4” presents unexpected low values in both maximal and mean of \( qq \). This comes from the stochastic characteristic of traffic and the uncertainty of traffic configuration.

<table>
<thead>
<tr>
<th>Data sets</th>
<th>Sample size</th>
<th>Maximal ( qq ) (kNm)</th>
<th>Mean of ( qq ) (kNm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DataL50r0.2</td>
<td>1760</td>
<td>55.94</td>
<td>34.51</td>
</tr>
<tr>
<td>DataL50r0.3</td>
<td>1828</td>
<td>50.95</td>
<td>34.41</td>
</tr>
<tr>
<td>DataL50r0.4</td>
<td>1851</td>
<td>48.24</td>
<td>34.34</td>
</tr>
<tr>
<td>DataL50r0.8</td>
<td>1873</td>
<td>52.68</td>
<td>34.46</td>
</tr>
<tr>
<td>DataL100r0.2</td>
<td>2160</td>
<td>28.15</td>
<td>19.57</td>
</tr>
<tr>
<td>DataL100r0.3</td>
<td>1315</td>
<td>27.08</td>
<td>19.65</td>
</tr>
<tr>
<td>DataL100r0.4</td>
<td>1507</td>
<td>30.32</td>
<td>19.67</td>
</tr>
<tr>
<td>DataL100r0.8</td>
<td>1220</td>
<td>31.03</td>
<td>19.95</td>
</tr>
</tbody>
</table>

Table 3.15: Sample size, maximal ‘\( qq \)’ and mean value of ‘\( qq \)’, for data sets with different length and correlation parameters \( r(S_t, S_{t-1}) \)

Figure 3.25: The color map of \( qq \) from different data sets shown in table 3.15

Figure 3.25 presents the color map of \( qq \) from eight data sets with parameters shown in table 3.15. The yellow-green represents the measure around 30\((kNm)\) and the light blue states for 20 approximately. Thus graphically, the maximal bending moment per day \( qq \) decreases significantly when the length of bridge \( L \) extends from 50\((m)\) to 100\((m)\). The reason behind is that the pressure on the deck caused by traffic
loads is dispersed by bridge components, thereby the longer bridge is, the less pressure each components bearing, which eventually results a lower bending moment. Consequently, the length of bridge parameter $L$ plays a important role in determining bending moment, and should be a key factor in consideration when designing or assessing the bridge for safety purpose.

**Figure 3.26:** The percentage of $qq$ falls in intervals. Data sets are generated with $L = 50(m)$, but different auto correlation coefficients

**Figure 3.27:** The percentage of $qq$ falls in intervals. Data sets are generated with $L = 100(m)$, but different auto correlation coefficients

For the sensitivity analysis of the auto correlations to the bending moment, we categorize $qq$ into $5(kNm)$ length intervals, and then compare the percentages of $qq$ in each interval under condition of the same length bridge. The bar graphs 3.26 and 3.27
show the comparison of bending moment configuration with respect to data sets with different auto correlation \(r(S_t, S_{t-1})\). Note that in these two figures the auto correlations are shortly presented as \(r\). For 50\((m)\) length bridge, there is no big difference of percentages among four data sets in the first five intervals. However, in the extreme large bending moment intervals, it provides a evidence that \(qq\) from data set with 0.8 auto correlation has more samples than the other data sets. It indicates that higher auto correlation leads to more possibilities in deducing large bending moment. For 100\((m)\) length bridge, we can notice that the positive influence of auto correlation to the bending moment is more obvious, as shown in figure 3.27, i.e. the red bar stays the highest in the last three intervals.

<table>
<thead>
<tr>
<th>Observations</th>
<th>(qq = 28(kNm))</th>
<th>(qq = 50(kNm))</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\text{Measures of } W_f) (kN)</td>
<td>(\text{Mean})</td>
<td>(\text{Percentiles' value})</td>
</tr>
<tr>
<td>Data(L_{50}r_{0.2})</td>
<td>586.0</td>
<td>[223.5 1005.7]</td>
</tr>
<tr>
<td>Data(L_{50}r_{0.3})</td>
<td>598.5</td>
<td>[195.0 1007.5]</td>
</tr>
<tr>
<td>Data(L_{50}r_{0.4})</td>
<td>550.2</td>
<td>[179.7 989.1]</td>
</tr>
<tr>
<td>Data(L_{50}r_{0.8})</td>
<td>608.1</td>
<td>[114.1 1012.3]</td>
</tr>
<tr>
<td>Data(L_{100}r_{0.2})</td>
<td>1364.5</td>
<td>[1074.0 1683.4]</td>
</tr>
<tr>
<td>Data(L_{100}r_{0.3})</td>
<td>1620.0</td>
<td>[1269.4 1926.3]</td>
</tr>
<tr>
<td>Data(L_{100}r_{0.4})</td>
<td>1404.0</td>
<td>[1059.6 1800.0]</td>
</tr>
<tr>
<td>Data(L_{100}r_{0.8})</td>
<td>1480.3</td>
<td>[1039.1 2023.0]</td>
</tr>
</tbody>
</table>

**Table 3.16:** The mean and [0.05 0.95] percentiles’ values of ‘\(W_f\)’, conditionalized on observed ‘\(qq\)’

<table>
<thead>
<tr>
<th>Observations</th>
<th>(W_f = 1369(kN))</th>
<th>(W_f = 1799(kN))</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\text{Measures of } qq) (kNm)</td>
<td>(\text{Mean})</td>
<td>(\text{Percentiles' value})</td>
</tr>
<tr>
<td>Data(L_{50}r_{0.2})</td>
<td>50.8</td>
<td>[42.1 55.6]</td>
</tr>
<tr>
<td>Data(L_{50}r_{0.3})</td>
<td>50.9</td>
<td>[42.8 55.6]</td>
</tr>
<tr>
<td>Data(L_{50}r_{0.4})</td>
<td>42.7</td>
<td>[38.5 49.6]</td>
</tr>
<tr>
<td>Data(L_{50}r_{0.8})</td>
<td>41.7</td>
<td>[36.4 51.4]</td>
</tr>
<tr>
<td>Data(L_{100}r_{0.2})</td>
<td>20.9</td>
<td>[18.0 24.1]</td>
</tr>
<tr>
<td>Data(L_{100}r_{0.3})</td>
<td>20.9</td>
<td>[18.0 24.1]</td>
</tr>
<tr>
<td>Data(L_{100}r_{0.4})</td>
<td>20.8</td>
<td>[18.1 23.6]</td>
</tr>
<tr>
<td>Data(L_{100}r_{0.8})</td>
<td>20.7</td>
<td>[17.7 23.8]</td>
</tr>
</tbody>
</table>

**Table 3.17:** The mean and [0.05 0.95] percentiles’ values of ‘\(qq\)’, conditionalized on observed ‘\(W_f\)’

In order to investigate the sensitivity of bridge length \(L\) and auto correlation parameter \(r(S_t,S_{t-1})\) to the inference results, we experiment the inferences by
conditionalizing on qq and Wf respectively (for simplicity purpose, only case \( N = 1 \) is experimented). As previous section concludes, given a data set, it is the Wf that determines qq most. Thereby, the conditionlized value of Wf given known qq and the conditionalized value of qq given known Wf are presented.

Table 3.16 shows the mean and [0.050.95] percentiles of Wf, conditionlized on \( qq = 28(kNm) \) and \( qq = 50(kNm) \) respectively. Staying in the same length bridge, we find out that Wf tends to increases when the auto correlation grows. The special case lies in the third row, where the data set is characterized with bridge length \( L = 50 \) meters, and auto correlation \( r(S_t,S_{t-1}) = 0.4 \). The reason behind is that the conditionalized observation falls farer away from the upper bound of the samples of qq. Another abnormal situation happened in the case with data set "DataL100r0.3". We can see unusual large values of Wf, resulted by the low value of maximal qq.

Table 3.17 presents the inference results of qq, by updating the information of \( Wf = 1369(kNm) \) and \( Wf = 1799(kNm) \) respectively. Wf shows a decreasing tendency when the auto correlation increases. This is because the qq values for high auto correlation cases concentrate more in the right interval, as shown in Figures 3.26, 3.27 and Table 3.15.

In conclusion, the bending moment is sensitive to both parameters: the length of bridge \( L \) and the auto correlation \( r(S_t,S_{t-1}) \). The former one has negative and significant influence, while the latter is weaker in determining bending moment, but positive.
CHAPTER IV
TRAFFIC INFORMATION MODELLING

4.1 Problem Statement

Traffic situations can be measured with variables as traffic speed, traffic intensity or traffic density. If we are able to predict any one of these three variables, then the traffic congestion is also predictable\cite{22}. Being notified for possible congestions, travellers can choose another route to arrive their destinations rather than continuing through the original route. Specially, the prediction of speed makes it possible to calculate the travel time, which further provides travellers with important information to arrange the schedule and determine the route from origin to destination. Of course all these possible outcomes should be predicted with some time in advance, otherwise it adds little value to users. To achieve an useful traffic prediction, we are aiming to find out a method to predict speed accurately in an acceptable time horizon. In this thesis project we only deal with predicting speed of vehicles for Dutch freeways.

There exists numerous instantaneous, model or data driven traffic prediction methods over the world. In our research we choose CNBNs for traffic prediction, because traffic is dynamic, time dependent, and CNBNs has certain advantages compare to other existing methods. More details about these advantages will be discussed later. Van den Haak (2006) presents a discrete Bayesian Network method in modelling traffic information in his Master thesis at TNO. The prediction in his thesis report were done on a small section (about 4 kilo meters long) of the motorway A4 in Netherlands. The results were promising, but both the prediction time horizon and the setting were not satisfying from the mobility point of view. Later on, Morales Napoles et.al. provide a Continuous Dynamic Non-parametric Bayesian Network (CDNBNs)
approach in predicting on the same single location and time horizon as in van den Haak (2006)[22]. We inherit CDNBNs method in our application, but the aim is much more ambitious as we try to predict the traffic situation on the whole motorway A10 (Ring Amsterdam with 30 km of road) with prediction time horizon 30 minutes. Further more, we bring traffic flow theory into the BN model construction, which is a new feature as well. The extension technique from one location to the whole A10 ring introduced in this report, provides a entrance to draw a traffic movement picture of the whole freeway network in the Nehterlands.

Traffic situations are correlated to many variables like environment, incidents, road conditions, etc. Also, traffic characteristics differ for different countries because of regulations, economics and social differences between people. All these elements make the field of traffic research challenging and unable to be solved completely in a short period. Nevertheless, our research in predicting the whole A10 ring traffic situation by using BNs can be further applied in predicting travel time and traffic flow management.

4.2 Traffic Modelling

4.2.1 Traffic theory

Research on the subject of traffic flow modelling started some fifty years ago, when Lighthill and Whitham (1955) presented a model based on the analogy of vehicles in traffic flow and particles in a fluid [27]. Most analytical traffic modelling techniques are conducted on the basis of variables which capture the traffic characteristics. Generally, these variables can be categorized into two subgroups: the vehicular variables (spacing, headway, vehicle speed) and stream variables (volume/intensity, concentration/density, mean speed), and can be clearly illustrated via a time-space diagram of the trajectories of the vehicles constituting a traffic stream [25]. In order to get a overall impression of basic traffic theory, we introduce these fundamental elements and then present the mathematical relationship between them.

For traffic stream calculations, two measures - spacing and time-headway between successive vehicles, are of fundamental importance. The spacing is simply the
distance between successive traffic units, typically measured from back bumper to front bumper. Time-headway is the time between the arrival of successive traffic units at a specified point. Nevertheless, the most important variables in traffic analysis are: speed, intensity and density. Before entering into the definitions and mathematical analysis of these variables, we introduce a useful graphical tool called trajectories image which are typically plotted on a time-space diagram.

Traffic units are commonly constrained to move along a certain path, which is defined as the trajectories. Thus, the relevant aspects of their motion can often be described in cartesian coordinates of time \( t \) and space \( x \). Figure 4.1 presents a trajectories plot in time-space diagram. The t-axis represents the time extension measured in hour, while the x-axis represents the space movement measured in kilometer, and the irregular curves are trajectories for individual vehicles.

**Figure 4.1:** Trajectories example in time-space diagram [26]

Traffic density, also referred to as traffic concentration \( k \) of the traffic stream, is defined as the average number of vehicles occupying a unite length of roadway at a given time period or the ratio of the number of vehicles appearing on a photograph to the length of the roadway segment. The formal definition for density is:

\[
k = \frac{m}{X}
\]

where \( m \) denotes the number of vehicles and \( X \) denotes the unit length of the road section [24].

The number of vehicles counted at the point of observation (point along a
roadway or traffic lane) divided by the total observation time (equivalent hourly rate at which vehicles pass a point on a freeway lane during a time period) is defined as the traffic volume or traffic intensity $q$, and measured in vehicles per unit time. That is, volume is the total number of vehicles that pass a point on a freeway during a given time interval. The formal definition for traffic volume is:

$$q = \frac{n}{T}$$  \hspace{1cm} (4.2)

where $n$ is the number of vehicles passing the cross-section of a road, and $T$ is the time period. Normally, the unit of intensity is expressed in vehicles/hour. The same as density, the intensity can refer to a specific lane or a road section with lanes. Intensity can be aggregated in lanes or time pieces if needed.

The third basic measurement of traffic is that of speed or mean speed. Speed of travel is simple and widely used measure of the quality of traffic flow. In the time-space diagram the speed of a vehicle at any time is the slope of the line. These speeds are known as spot speeds. Formally the spot speed at time $i$ can be expressed as:

$$V_i = \frac{dx_i}{dt_i}$$  \hspace{1cm} (4.3)

Spot speed is the instantaneous speed of a vehicle as it passes a specified point along a road. The average of a series of measures of spot speeds can be expressed as time-mean speed or space-mean speed. The time-mean speed $V_t$ is the arithmetic mean of spot speeds of all vehicles passing a point during a specified interval of time

$$V_t = \frac{1}{N} \sum_{i=1}^{N} V_i$$  \hspace{1cm} (4.4)

where $V_i$ refers to the observed spot speed of $i$–th vehicle, and $N$ is the number of vehicles observed.

The space-mean speed is calculated on the basis of the average travel time it takes $N$ vehicles to traverse a length of roadway D. Mathematically, space-mean speed $V_s$ is defined as the harmonic mean of spot speeds:
In other words space-mean speed is the average of vehicles speeds weighted according to how long they remain on the section of road.

The speed is measured in kilometer/hour, intensity in vehicles/hour and density in vehicles/kilometer. Note that the units balance represents a three-dimensional relationship between the most important basic traffic stream variables - intensity, mean speed and density as

\[ q = k \cdot V \]  

(4.6)

Generally it is incorrect to attempt to compute the value of one of these three variables by varying another while holding the third constant. Greenshields (1935) [37] was able to develop a model of uninterrupted traffic flow that predicts and explains the trends that are observed in real traffic flows, which reduces the above-mentioned traffic fundamental variables relationship from three dimensions to two dimension. The assumption of uninterrupted traffic stream is not really a constraint for our research, since we are focusing on A10 ring freeway in the Netherlands and the freeway is characterized by few interruptions.

Within Greenshield’s model, speed and density are linearly related. Inserting this speed-density relationship into density-intensity-speed fundamental equation 4.6 yields a second order equation between density and intensity, speed and intensity. A graphical view of these three relationship is shown in figure 4.2.

Greenshield model is not perfect but simple and powerful. From his model, we can conclude the followings. When the density on the freeway is zero, the intensity is also zero as there are no vehicles on the roadway, and as the density increases, the intensity also increases. However, when the density reaches its maximum, generally referred to as the jam density \( k_j \), the intensity must be zero, as vehicles will tend to line up to the end. It follows, therefore, that as density increases from zero, the intensity will also initially increases from zero to a maximum value. Further continuous increase in density, however, will result in continuous reduction of the intensity, which
CHAPTER IV : TRAFFIC INFORMATION MODELLING

(a) Density $k$ VS Speed $V$
(b) Density $k$ VS Intensity $q$
(c) Speed $V$ VS Intensity $q$

Figure 4.2: Two dimension relationship among density, intensity and speed, within Greenshield model

will eventually be zero when the density is the maximal jam density \[25\].

Greenshield model provides us with an opportunity to reduce the number of variables in model construction, because of the linearity relationship between density and speed. More details will be discussed in section 4.4.2.

4.2.2 Early works

Under the assumption that individual road users are rational decision makers who base their choices (e.g. route and departure time) on minimizing their expected costs, usually in terms of travel time and reliability, providing travellers, road managers, transportation companies with traffic information allows them to make more informed decisions, yielding not only economic benefits for individuals, but potentially also more stable and less congested traffic conditions for all road users \[28\]. For these potential benefits, a lot of work has been done in the field of traffic information simulation and prediction. Generally, these work can be categorized into instantaneous, model and data driven methods. The instantaneous approach only takes historical or current data into account and travel time are often predicted with an instantaneous approach, that is, with given data this approach should be able to yield immediate results if the number of necessary computations are optimized \[28\]. The model approach, often taken by Civil Engineers, is conducted on the basis of engineers’ understanding of individual vehicle characteristic and extensive knowledge of traffic stream. Last but not least, we refer to the data driven model, which benefits from the huge data set collected by different systems.
As the name suggests, data driven method performs the prediction by inductive techniques. It considers traffic process and, tries to correlate observed measurements to the historical and current traffic data to model the traffic information. Given enough data available, data driven model is a good choice as not only it deals with the feedback automatically in the data, but also it provides observable traffic patterns. Supported by Smart Mobility (TNO) with abundant traffic data, our research thereby will focus on data driven approach to predict traffic situation. More interpretation about data acquisition typically for this research will be presented in section 3.1.

During the past years, an enormous amount of data driven approaches can be found in literatures. It is impossible to list them all here. We list the main categories of existing data driven approaches here [22]:

- ARIMA models
- linear time series models
- non-linear time series models
- support vector regression models
- feed-forward neural networks
- recurrent neural networks
- Bayesian networks

Sun et al. [29] proposed a Bayesian Network model in 2004 to forecast traffic flow with incomplete data. This model creates a Bayesian Network, in which the nodes represent the intersections where travellers can change directions and the links represent travel direction. It incorporates intensity values from road links at different time instants in the past to be the margins of the nodes. After creating the Bayesian Network, Sun et al. performed a Principal Component Analysis to reduce the number of dimensions in which the data is present [22] [29]. A principal component is a linear combination of variables chosen such that data originating from different classes is separated as far as possible [22]. For next step after reducing the data dimension, Sun
et al. employed a Gaussian Mixture Model (GMM) to approximate the joint probability
distribution of the set variables of the Bayesian Network. A Gaussian Mixture Model is
a parametric probability function represented as a weighted sum of Gaussian
component densities. The parameters are estimated from training data using the
iterative Expectation-Maximization algorithm.

The Bayesian Network method proposed by Sun et al. [29] can be used for a
short-term traffic information forecasting, ranging from five minutes to half a hour. The
intensity prediction of this method fails to recognize some peaks which is, of course, a
defect since intensity peaks are very important to road users since it often refers to
traffic congestion.

to the Bayesian Network. Spatio-temporal means the common-sense background
information such as people’s long term activity laws. They derived the correlation
between road sections by using Pearson correlation coefficient. The input data for their
Bayesian Network originates from UTC/SCOOT, which is a system of the Traffic
Management Bureau of Beijing. No details about the performance measurements are
shown in their report.

van den Haak 2010 [22] presents a discrete Bayesian Network approach in
modelling traffic information in his master thesis. Differ from the method described in
[29], van den Haak introduced speed and density measures rather than intensity into
the model. This model analysis is conducted on the basis of Dutch traffic data collected
on the freeway A4 in the Netherlands by the MONItoring CAsco (MONICA) system
[28]. The data collection for the road network in the Netherlands is managed by the
Dutch Ministry of Transport, Public Works and Water Management with MONICA.
van den Haak presents the model shown in figure 4.3. The density and speed historical
information at three different locations selected from A4 road are taken into account to
predict the speed for a upstream (follow the traffic direction) location A with
prediction time horizon T (see figure 4.3 (a)). Figure 4.3 (b) presents the Bayesian
Model structure.

It is possible to use both upstream and downstream traffic information in the
model. van den Haak [22] concluded that the model which only takes downstream
Figure 4.3: A demonstration of Bayesian Model from [22] - A schematic overview

traffic information into account performs the best in his specific data set. This discrete BN model is robust and able to predict traffic congestions for over a time horizon of 5 minutes very promisingly and for a prediction horizon for 120 minutes but not very accurately.

To overcome the low accuracy in long time horizon prediction, van den Haak [22] generated a hybrid model that combines the historical modelling with the Bayesian Network. The historical modelling tries to find the best fitting model by calculating and comparing the distances from the current data to the models in a model database. Then the prediction model is able to adapt itself to the new data continuously based on the historical data. Therefore, the hybrid prediction model based on the historical data first calculates raw initial estimates of all input nodes for the Bayesian Network. After this the Bayesian Network is able to use its estimated conditional probabilities tables to smooth the initial historical based prediction to a more correct traffic prediction.

The prediction performance of the hybrid model is significantly improved for a long prediction time horizon such as 150 minutes, thus is convincing. However, the discrete BN model does have some shortcomings as we mentioned in section 2.2:

- Computational constraints: adding more nodes or more discretization classes for continuous variables will yield an exponential increase in computation time or in the quantification of the required conditional probability tables;
it may suffer from excessive assessment burden in data-sparse environment. Using higher number of classes to discretize the continuous data requires more data points to be sure that every class is represented in the total dataset.

To get rid of above mentioned potential constraints embedded in discrete BN models, Morales Napoles et al. [23] introduced an alternative approach named as Continuous Dynamic Non-parametric Bayesian Networks (CDNBNs) to predict traffic information. Different from discrete BNs, all variables in CDNBNs are continuous and, the Non-parametric refers to the fact that the 1-d margins may be non-parametric distributions. Details about CDNBNs have been described in section 2.2. The third feature - Dynamic - refers to the fact that we model a discrete-time stochastic process. A dynamic BN is represented by a collection of static Bayesian Networks, one at each time-slice, together with arcs between nodes in different time-slices [23]. Thus, a CDNBN is a non-parametric Bayesian Network in which all nodes are continuous, and some arcs represent the (conditional) rank correlation of variables in time. Morales Napoles et al. investigated the same location on A4 freeway in the Netherlands and used the same data set as van den Haak did in [22]. One example of CDNBN is shown in figure 4.4, where $V_i$ and $D_i$ represents the speed and density at location $i$ respectively, and each block shows a static BN in a time-slice.

![CDNBN Diagram](image)

**Figure 4.4:** One example dynamic BNs in [23]

The CDNBN approach presented by Morales Napoles et al. does improve the computation efficiency in quantifying the high dimensional BNs, and it provides convincing results up to prediction time horizon 30 minutes. However, there are a couple of points need to be further improved or developed for this method. First, the
Gaussian copula might not be the better choice for describing dependencies than other copulas. Thus, whether the Gaussian copula sufficiently represents the data needs to be validated. Another issue is both BN models in [22] and [23] focus on one location prediction but didn’t extend the model to the one road section network, though it is very important in real applications.

4.2.3 Problem statement

Recommended by the mobility team in TNO, we focus on modelling the traffic information on highway A10 ring around Amsterdam in the Netherlands, an important and challenging enough work for the project that represents a realistic part of the Dutch road network.

The freeway A10 ring Amsterdam has a length of 32km. There are four other freeways connected to A10: freeway A8 at interchange Coenplein, A4 at interchange De Nieuwe Meer, A2 at interchange Amstel, and A1 at interchange Watergraafsmeer. Besides, most of city routes of Amsterdam take exit from A10 towards the city route in the center. Different from the model location in [22] [23], A10 is an open test road section, which makes the prediction more complicated. Figure 4.5 (b) shows the whole A10 ring network. We will not model all locations on this ring, but choose the points existing in data set instead. The main reasons for choose A10 ring were:

1. It is a relatively small motorway but quite challenging to predict;

2. It is a very important motorway for the Dutch network as its traffic flows (vehicles/hour) are high;

3. Many TNO Mobility projects are already running on this highway.

We present a prediction location example as shown in figure 4.5 (c). Three locations A, B and C are taken into account for the model. The most interesting value to predict is the mean speed, as from this we can compute the travel time and determine the traffic situation for road users. Both upstream and downstream information are important in prediction, as in [22] it states that the model which only takes downstream traffic information into account performs the best. This corresponds
CHAPTER IV : TRAFFIC INFORMATION MODELLING

(a) Route information on A10 ring

(b) Overview experiment road section, A10 ring in the Netherlands

c) Location example on A10 ring

Figure 4.5: Schematic overview of experiment road section A10 ring, and demonstration of prediction location B
with the findings of Treiber and Helbing in [31] in which it is stated that traffic information propagates following the driven direction with a constant speed of 80km/h and against the driving direction with 15km/h if the traffic is in a congested state. Thus, we are aiming to predict the vehicle mean speed at location B, on the basis of historical and current traffic data at locations A, B and C.

Note that the distances between A and B, B and C are not specified in figure 4.5 (c). The choice of A and C when we are predicting mean speed at B will be discussed in section (Model construction). With the same methodology applying to other locations, we are able to capture the traffic situation for the whole A10 network.

4.3 The Data

This subchapter describes how traffic data for this thesis is acquired and prepared. Since we use the same dataset as it in [22], we cite the most important acquisition descriptions and for more details please refer to [22] chapter 4. Data acquisition and pre-processing are important in any traffic data related project. The objective of the work presented in this section is to illustrate how traffic data can be collected and processed such that it can be used for future modelling directly. The data preparation consists of two steps: data acquisition and data processing. Besides, traffic has certain patterns. Finding out these patterns helps reducing the complexity of modelling and improving the prediction accuracy.

4.3.1 Data acquisition

Traffic flow parameters such as intensity, speed can be measured by sensors locating along the highway. The sensors can be inductive loops, cameras, pneumatic tube and so on. Pneumatic tube, which basically consists of a hollow tube that is put across the road, has the longest history in acquiring traffic data. It delivers quite accurate results, but it is not very reliable as the tubes easily get damaged [24]. Cameras records and delivers not only the traffic flow measurements but also the images. These images can be used for visual inspection by the operators of the traffic control center in order to detect and diagnoses incidents. However, cameras are more expensive though it also extracts accurate information [32].
The most widely spread highway traffic sensor technology is the inductive loop detectors (ILDs). They are wire loops buried in pavement, connected to an inductive loop detector in the controller cabinet. When a vehicle passes over the loop, the loop inductance changes, and the detector sends an impulse to the controller. The inductive loop can be categorized into two types: single and dual ILDs. The single inductive loop detectors can only do a single measurement on a passing vehicle, thus unable to measure the speed directly. It estimates the speed based on the recorded distance and time between two measuring locations when a vehicle passes by [22].

In the Netherlands, dual ILDs are settled along the highways. Dual ILDs differ itself from single ILDs by consisting of two consecutive single loop detectors placed a short distance apart, thereby the speed can be easily calculated and the measurement accuracy is improved comparing to single ILDs technique. The data collection for the road network in the Netherlands is managed by the Dutch Ministry of Transport, Public Works and Water Management with the MONItoring CAso (MONICA) system [28]. Figure 4.6 presents an example of dual ILDs setting and its loop configuration applied in the Dutch highways.

In general, there is an ILD every 500 meter at the Dutch highways [28]. The MONICA system collects the real time traffic data, which measures point mean speed (kilometer/hour), point mean intensities (vehicles/hour) and contains the messages on Dynamic Route Information Panels (DRIPs), from the equipped dual ILDs. Possible lane closures, temporal speed limits, etc. are provided by DRIPs, as figure 4.7 shows.
In this research, all data are collected by MONICA system on A10 ring Amsterdam in 2011. Theoretically, the raw data should consist of point mean speeds and intensities, together with corresponding hectometers and time for each measurement. However, in reality the raw data are usually incomplete due to the cases of missing data. In order to draw a complete graph of traffic information we have to execute the data process before getting into modelling.

4.3.2 Data processing

Sometimes the sensors or the communications are out of order or fail to produce accurate values due to unexpected situations such as power failures, interferences, servicing of the devices, works on the road surface, incidents and accidents, etc. If the sensor was not available in the first place, the measurement is labeled as unavailable, using a sentinel element. In our data set the sentinel element '0' was used; this value is likely to occur in reality as in case of heavy traffic jam but seldom, thus we replace the missing data by zeros.

On average, the MONICA system has a missing value percentage of 12%. In a significant number of cases there are even occasions in which over 20% of the measurements are either missing or corrupt [28]. Here corrupt data refers to the case that the sensor value does not comply with the basic assumptions, for instance the non-negativeness of speeds and intensities. The data missing occurs more frequently
in case of traffic congestion, as the traffic flow could be too low and only few cars are passing the detector.

In [22], the author introduced two simple approaches for filling in missing data. The first one is taking just the mean value to replace the missing data but it disturbs the traffic harmonic process. The second simple approach is to use linear interpolation with a nearest neighbor technique. The term ‘neighbor’ here denotes measurements close by in time or location. With this approach it would be difficult to get a good result when dealing with large time or space missing data gap[22].

Another interpolation-based method to estimated the missing values using historical data was presented in [32]. In this approach, an assumption that the evolution of traffic patterns on a given day of the week is the same as the evolution of the traffic pattern on the corresponding day in a reference week was made. The missing data at time step $k$ is calculated by timing the reference data at time step $k$ with the factor of data at time step $k - 1$ divided by reference data at time step $k - 1$. The main advantage of this method is that its computational complexity is low so that it can also be used in on-line real-time applications. But it meets the same defect as in mean approach that the traffic might be not smooth.

A state-of-art algorithm for filling in these missing values is the Treiber and Helbing filter [31]. This method uses a spatio-temporal lowpass filter, i.e. only components of low frequency can pass the filter, while high-frequency contributions are considered as fluctuations and smoothed out. One particular feature of our filter is that it is nonlinear and adaptive to the traffic situation in distinguishing free and congested traffic. It has following three important properties [31]:

1. In case of free traffic: perturbations move essentially in the direction of traffic flow with constant characteristic speed $C_{\text{free}} = 80\text{km/h}$;

2. In case of congested traffic, perturbations propagate against the direction of traffic flow with a characteristic and remarkably constant velocity $C_{\text{cong}} \approx -15\text{km/h}$;

3. The filter should smooth out all high-frequency fluctuations.

In our research, we use the data where the Treiber and Helbing filter was
applied for filling the missing data, as it has been implemented by the faculty of Civil Engineering at Delft University of Technology and optimized by the Netherlands Organization for Applied Science (TNO). Figure 4.8 presents an example of comparison between speed plots of raw data and Treiber and Helbing filtered data. From the speed colorbar, we can see that the green part represents the regular traffic flow, while the red part represents the potential or actual traffic congestion. Note that traffic information are smoothed over the missing values areas and the key characteristics such as congestion location and time seem to be kept. van den Haak [22] stated that the red part (the traffic congestions) seem to propagate against the direction of the traffic flow in time. This observation happens in this case as well.

Bayesian Networks can also be an approach to filter the missing data because of its predicting and diagnoses property. We will discuss more about this in later subchapters.

The MONICA system measures speed and intensity per lane. The road network in the Netherlands usually consists of roads with 2, 3 or 4 lanes. It will be a very complicated work if we investigate all lanes. As a simplified approach, we aggregate the speed and intensity measured at each lane, and then propagate the analysis. A convenient overview of aggregation methods can be found in [22]. In this research, the harmonic mean method (equation 4.5) is used to calculate the aggregated mean speed.
4.3.3 Preliminary Analysis

When investigating the traffic patterns, one may find out that the traffic data are characterised by certain homogenous subsets [22]. On highways, the vehicles follow a regular speed in case of free traffic, thus there is nothing special to tell. However, during potential congestion time period the mean speed might be characterised by observing characteristics such as working days or weekends, seasons, etc.

In [22], the author analyzed the basic traffic pattern, the further clusters and the homogeneity of the traffic patterns. This analysis was conducted on basis of the traffic data set S collected on the A4 left side around hectometer 23.0 for the whole year - 365 days - of 2009. The possible subsets are listed as following:

- weekend: yes - no
- day of the week: Mon - Tue - Wed - Thu - Fri - Sat - Sun
- holiday: yes - no
- public holiday: yes - no
- summer: yes - no
- rain: no - little - medium - heavy - extreme
- temperature: low - medium - high.

Assume that $V(t, d)$ is the average vehicle speed as function of the time $t = t_i$, where $1 \leq t \leq N$, and for a specific day $d$, where $1 \leq d \leq M$. Here $M$ denotes the number of days in set S, and $N$ is the number of time splits corresponding with that day. In this case, the average vehicle speed graph of the data set $\bar{V}(t = t_i)$ is

$$\bar{V}(t = t_i) = \frac{1}{M} \sum_{d=1}^{M} V(t_i, d) \quad (4.7)$$

To compute the standard deviation $\sigma_i$ at time $t_i$, we compute

$$\sigma_{t_i} = \sqrt{\frac{1}{M - 1} \sum_{d=1}^{M} (V(t_i, d) - \bar{V}(t_i))^2} \quad (4.8)$$
The average standard deviation $\sigma_{total}$ average over $i$ becomes:

$$\sigma_{total} = \frac{1}{N} \sum_{i=1}^{N} \sigma_{i}$$

(4.9)

The starting point of the traffic patterns analysis in [22] is the data set $S$ with 365 morning rush hours from 6am to 11am in 2009. The average standard deviation was employed to analyze the homogeneous property among different subsets. High standard deviation indicates that the corresponding subset has large variation in speed, while lower $\sigma_{total}$ represents more homogeneity. The same analysis applies to intensities and densities. General conclusions of the analysis in [22] can be summarized as:

1. Both weekend and week days clusters have lower standard deviation than the total set. The variance in weekend speed decreases enormously, while in week days it is still high. Thus, it is needed to find homogeneous subgroups for week days;

2. Clustering the data for rain or season seem to have little effect on the standard deviation decreasing;

3. The cluster of holidays or public holidays turn out having homogeneity.

4. The cluster for Mondays, Tuesdays and Wednesdays have comparable standard deviations for speed, whereas Thursdays have a higher standard deviation. The standard deviation for Friday is lower than the standard deviation for the other working days of the week.

In order to verify whether homogeneous property presented above are applicable in our data set, we try the experiment at one location chosen from A10 ring Amsterdam in the Netherlands. This ring is a very important freeway section for the Dutch network as its traffic flows are high. The data set prepared for this experiment consisted of speed and intensity values in year 2011. These values were collected at hectometer 28.07(km) of A10 ring (see figure 4.9) measured per minute. In this data set, the hectometers increase monotonically when the traffic direction is clockwise or right side. The hectometer starts at 1.0(km) and ends at 33.1(km). Note that the locations with hectometers 1.0(km) and 33.1(km) overlap with each other. If the traffic
move along the counterclockwise direction, the hectometer decreases monotonically from 33.1\((km)\) to 1.0\((km)\). Treiber and Helbing filter was used to filter the missing data (see section 4.3.2).

Our experiment differs itself from the data analysis presented in [22] in two parts: first we take the traffic information for all day into investigation rather than only including information from morning rush hours. The reason behind is that we found that potential traffic congestion exists not only in the morning but also in the afternoon and, in this project we are aiming to model traffic for the whole time horizon; Secondly we reduced the investigation dimensions to speed only because we expect that intensity and density would be consistent with the speed property.

It becomes clear that the weekend cluster has a significant different pattern, comparing to the week days cluster, as show in figure 4.10 (a). The total set locates in between as it is a combination of weekend and week day subsets. Thereby it is needed to distinguish with week day traffic and weekend traffic. Figure 4.10 (b) tells us that the public holiday traffic are less possible to have congestion than Non public holiday or the total set. Unlike the results in [22], the mean speed for the days from Monday to Friday are quite consistent as subfigure (c) presents. Separating public holiday from total set or week days seems having little effect on changing the patterns. The reason is that the number of public holiday is too low to bring influence to the total set. In year 2011, there was 12 public holidays in the Netherlands, which is negligible comparing
(a) Mean speed for Total set, Weekdays, Weekend days  

(b) Mean speed for Total set, Public holiday set, Non public holiday set  

(c) Mean speed for Weekdays, Public holiday in week days, Non public holidays in week days  

(d) Mean speed for days of the week  

Figure 4.10: Mean speed plots for the total set, weekdays, weekend days, days of the week, public holidays, Non public holidays, public holiday in week days, Non public holidays in week days, in 2011
(a) Mean speed for Total set, Public holiday set, Non public holiday set
(b) Mean speed for Weekdays, Public holiday in week days, Non public holidays in week days

Figure 4.11: Mean plots for speed for the total set, public holidays, Non public holidays, week days, public holiday in week days, Non public holidays in week days, with high percentage of public holidays in both total set and week day set, in 2011

to 365 days in the whole year somehow. Figure 4.10 (d) tells us that the mean speed from different week days does not show much difference, neither between the days from weekend.

Then one questions arises: what if the percentage of public holidays in total set or week days is big? Will the effect of holiday separation become large? To answer this, we select 5 random days from total set and week days set and combine them with 2 public holidays. Then we repeat the average speed plot.

Figure 4.11 presents a situation with high percentage (around 28.6%) of public holiday in total set and week day set. Note that for total set or week day set, the separation of public holiday does not show much the traffic patterns’ difference. Nevertheless, if we want to model the traffic on holidays, it is suggested to choose public day cluster as the data source for the purpose of getting more accurate results.

It is worthy to mention that the results of this experiment are only based on the data set collected at the A10 ring at hectometer 28.07(km) in 2011. The application in different locations may bring very different observations. For instance in hot tourist places, the traffic on weekend or holidays may have more traffic congestions than working days.
4.4 Model Preparation

In the previous section we have discussed data acquisition and processing to achieve a centralized and complete database. With this complete data source, we are able to investigate the traffic prediction using BNs. Both the data processing and the modelling introduce the flexibility into the system. The data flow is summarized in Figure 4.12. First, the raw data are collected by the sensors. Next, the missing data are filtered by using processing models. The filtered complete data set then can be furthered used for prediction, where the prediction models are applied.

![Figure 4.12: Schematic view of the data flows between the sensors and the database and the applications.](image)

In this research, we are aiming to predict the traffic speed in a certain time horizon $T$, based on the complete historical and current traffic data, at A10 ring network in the Netherlands. We have present some earlier data driven traffic modelling works in section 4.2.2. In our research we show the use of Continuous Dynamic Non-parametric Bayesian Networks (CDNBNs) to predict traffic speed.

For the construction of the BN, variables should be selected first and the structure should be learned from data. These two steps are presented in section 4.4.1 and 4.4.2. Further, we assume Gaussian copula for capturing the dependence relationships among variables.

4.4.1 Nodes selection

The general BN model which we proposed at the beginning consists of nodes
CHAPTER IV : TRAFFIC INFORMATION MODELLING

(a) Traffic propagation in time and space

(b) A coincidence of location A and C example

Figure 4.13: Application of Traffic Flow theory at A10 ring Amsterdam

\[ V_{k,-i} \] and \[ D_{k,-i} \]. \[ V_{k,-i} \] denotes the speed at location \( k \), \( i \) time steps before current time, and \[ D_{k,-i} \] refers to the density measure. In [22][23], it was decided to take speed and density measures of 3 locations (the locations denoted by A, B and C, see figure 4.3(a)) into account. We inherit the same idea of locations, for the sake of complexity and computation power. In this section we try to solve two problems: how to choose locations taken into model? and whether both speed and density measures should be involved in the BN model simultaneously?

From the Traffic Flow theory, we know that traffic can be decomposed into two main states: the free flow and the congestion state. In the free flow state drivers can freely drive with their desired speed, unless there are speed limits. And in the congested state drivers have to slow down the speed due to the dense traffic. Any given traffic state can be approximated by a linear combination of these two basic states. As we discussed in section 4.2.1, perturbations move in the direction of traffic flow with constant characteristic speed \( c_{free} = 80(km/h) \) in case of free flow traffic, while in case of congested traffic, perturbations propagate against the direction of traffic with constant characteristic speed \( c_{cong} = -15(km/h) \). These two flow propagation properties are useful in determining locations selected for modelling.

Figure 4.13 (a) presents an example of traffic propagation in time and space, given traffic direction from A towards D. From Traffic Flow theory we know that the state at location B at time \( t_B \) is a linear combination of state at location A at time \( t_A \) in the past that propagates along the free flow line and the state at location D at time \( t_D \)
in the past that propagates along the congested line. Note that the free flow line has positive slope 80 and the slope of the congested line equals to −15. The number of time horizon units occupied by these two lines are determined by the distance between A and B, and B and D.

On the other hand, suppose we want to predict the speed at location B at time $t_B$ with known traffic information up to one unit time horizon before $t_B$, i.e. up to $t_A$. Following Traffic Flow theory we are able to find out the two components which form the combination at B - they are location A and C. For instance, assume we are aiming to predict the speed at location B with prediction time horizon 30 minutes, then we can get: $d(A, B) = \frac{30}{60} \times 80 = 40(km)$ and $d(B, C) = \frac{30}{60} \times 80 = 7.5(km)$. Note that the distance between A and B falls out of the scope of A10 ring, since the freeway A10 ring Amsterdam has a length of 32km in total.

Based on Traffic Flow theory, the distances $d(A, B)$ and $d(B, C)$ increase as the prediction time horizon increases. Our experimental road sector A10 ring is approximated by the circumference of a circle, thus as prediction time horizon increases the locations A and C get closer and closer and coincide somewhere when the prediction time horizon reaches certain point. Suppose we are observing traffic flow in the clockwise direction, then the shortest prediction time horizon resulting in the coincidence of A and C would be:

$$t_{\text{coincide}}^1 = \frac{32}{80 + 15} = 0.337(h) \approx 20.21(minutes)$$ (4.10)

where the superscript 1 denotes the first time. Similarly, we can get the predicting time horizon which leads to the $i$-th coincidence of locations A and C:

$$t_{\text{coincide}}^i = \frac{32i}{80 + 15} = 0.337i(h) \approx 20.21 \times i(minutes)$$ (4.11)

From figure 4.14 we can observe that when the distance between A and C is close to zero, the correlation of $V_A$ (speed at location A) and $V_C$ (speed at location C) is close to 1. The correlation decreases dramatically as distances grows at first, then becomes stable when the distances are above 5km. The Pearson correlation is 1 in the case of a perfect positive (increasing) linear relationship as discussed in section 2.1.1,
and high value of the correlation indicates high degree of linear dependence between variables. We emphasise again here that we are aiming to predict the speed at location B given traffic information obtained at A and C. In case of perfect positive speed or density correlation between A and C, then it does not add value to gain more accurate prediction than only using information at one location A or C.

Consequently, we have to avoid the situation that the selected locations A and C are too close as figure 4.13(b) shows. In our research, we set \(d(A, B) = 6\text{km},\ d(B, C) = 1\text{km}\) in case of distance between theoretically selected location A and C is less than 7\text{km}.

We come back to the issue that \(d(A, B)\) or \(d(B, C)\) is bigger than the length of A10 ring Amsterdam. In this case, we assume that the vehicles keep running along the ring. Thereby, the perturbation will be propagated along the ring circle by circle. Under this assumption, the selection of locations A and C can be determined by the remainder of the theoretical distances divided by propagation speed.

In general, in our research four main steps should be followed to select locations A and C when predicting speed at location B:

1. Traffic flows from A towards B towards C (see figure 4.13(a));

2. Calculate \(d^T(A, B) = 80\times T,\ d^T(B, C) = 15\times T\), \(T\) denotes the predicting time horizon measured in hour; if \(d(A, B) < 32\text{km}, d(B, C) < 32\text{km}\), then \(d(A, B) = d^T(A, B)\),
\( d(B,C) = d^T(B,C); \)

3. If \( d^T(A,B) > 32(\text{km}) \), then recalculate \( d(A,B) = 32 - \text{mod}(d^T(A,B),32) \); if \( d(B,C) > 32(\text{km}) \), then recalculate \( d(B,C) = 32 - \text{mod}(d^T(B,C),32) \); \( \text{mod}(d^T(A,B),32) \) denotes the remainder of \( \frac{d^T(A,B)}{32} \) and \( \text{mod}(d^T(B,C),32) \) denotes the remainder of \( \frac{d^T(B,C)}{32} \);

4. Calculate \( d(A,C) = 32 - d(A,B) - d(A,C) \); if \( d(A,C) < 7(\text{km}) \), then \( d(A,B), d(B,C) \) are re-defined as \( d(A,B) = 6(\text{km}) \) and \( d(B,C) = 1(\text{km}) \).

4.4.2 Model construction and validation

In the earlier works [22][23], both density and speed measurements are taken into traffic prediction models. Greenshield [33] developed the first and most simple relation between speed and density on freeways specifically. Within Greenshield model, speed and density are linearly related. Inserting this speed-density relationship into density-intensity-speed fundamental equation yields a second order relation between intensity and speed. We have discussed details about these traffic basic in section 4.2.1.

In order to verify whether the Greenshield model can be applied in our data set, we take the speed and intensity data collected at A10 ring 28.08 hectometers, on 4th of March in 2011. We use the 720 observations measured per minute from 12 : 00 up to 24 : 00. Figure 4.15 presents the Density-Speed and Intensity-Speed plots. Clearly the density and speed have a linear relation, and the intensity and speed have a parabola trajectory.

In figure 4.15 only the data at one location at A10 ring in a certain time period are presented. But we assume the observation - Density-Speed has a linear relation - applies to the data sets. Linear relation between two variables leads the absolute value of the Pearson correlation of them to be 1. In this case, we claim that only one variable needs to be taken into BN model rather than both, since taking both would not add more information to predict the speed. Consequently, the variables will be considered in BN model reduce to speed only.

In [23], several network structures and their results were discussed. We employ one time step inhomogeneous model and two time steps inhomogeneous mode in our research. For one time step model, only the current speed information will be used to
predict with prediction time horizon $T$, while for the two time steps model, the speed information collected at present and $T$ time horizon before will be used to predict the speed $T$ time horizon into the future. Figure 4.16 presents a demonstration of these two BN types. VB denotes the current speed at location B; VB-T denotes the speed at location B but $T$ time units before; VBT denotes the speed at location B in a $T$ time horizon into the future. The same applies to VA, VC, VA-T and VC-T. By inhomogeneous we mean the dependence (correlation matrix) and margins vary in time rather than being constant. This can easily be seen in figure 4.17, where we map Pearson correlation between VA and VBT, VB and VBT, and VC and VBT, based on the speed data collected on the 4th of March in 2011 at A10 ring 28.08 hectometers.

As a data driven approach, it is recommended to learn the BN structure from
data. Hanea [20] introduced an algorithm of learning the structure of a NPBN with Gaussian copula from an ordinal multivariate data set. This algorithm is fully implemented in the software UNINET.

In this research, we will use the filtered data set described in section 4.3.2 for the year 2011, for the locations at A10 ring Amsterdam network. Data is available for hectometers from 1.4(km) up to 33.08(km) (see figure 4.5(a)), and every location in between is separated by 0.1km distance, with clockwise traffic direction. The data set has one minute intervals and around 1% missing data. We disregard the missing data since their percentage is very low. It is impossible and not necessary to present the BN structure learning procedures for models corresponding to different locations and days. So we take the day - 4th of March (Friday) in 2011, the location - hectometer 28.08 at A10 ring Amsterdam, the prediction time horizon $T = 10(min)$, as our testing demonstration.

The train data set was generated by randomly selecting 51 days speed data from the whole year. For one time step model, the speed data at location location A - hectometer 14.82, location B - hectometer 28.08 and location C - hectometer 31.38 are included. VBT - speed at location B but $T = 10(min)$ in the future is a shifted version of the VB. This arrangement leaves $73440 \times 4$ samples in the train data. For each minute, we compute the correlation matrix with observations one hour ahead of all selected days within the training data. For instance, at 00 : 01 hours we compute the correlation matrix (according to the BN of interest) with observations on all selected days in the

Figure 4.17: A demonstration: Pearson correlation vary in time
training data up to 01:01. Thus the data for computing each correlation matrix has 3060 samples.

The test data set includes one day’s speed data arranged by 4 columns of VA, VB, VC and VBT as well. VA, VB and VC are treated as observations, and VBT is treated as the real data. The test data set contains 1440 × 4 samples.

Following the BN structure learning from data procedure presented in section 3.3.1, we obtain the model which represents the train data fairly. Figure 4.18 shows the model structure. The corresponding validation was conducted on the basis of 3660 × 4 samples from 16:00 to 17:00 on day 4th of March, 2011. The validation results are shown in Figure 4.19. Note that the determinant of the normal rank correlation matrix $DNR = 0.63$ falls between the $[0.6, 0.65]$ quantiles of the distribution of the determinant of the rank correlation matrix of this BN using Gaussian copula, that is DBBN. Meanwhile, the determinant of the empirical rank correlation matrix $DER = 0.6$ is within the 90% central confidence band of $DNR$, specifically, falls between the $[0.35, 0.4]$ quantiles of the distribution of $DNR$. These two results mean that both the model and Gaussian copula represent the data reasonably in sense of given significant level.

![Figure 4.18: Model I: One time step model at time 16:00, 4th of March (Friday) in 2011, at hectometer 28.08 on A10 ring](image)

For two time steps model, we have 7 variables - VA-T, VB-T, VC-T, VA, VB, VC and VBT - by hand (in UNINET, VA-T, VB-T, VC-T are not valid, so we use VA_T, VB_T, VC_T instead). Similarly, we start learning structure from data with the saturated model, as shown in figure 4.20. Figure 4.21 presents the model and Gaussian copula
4.5 Implementation with Filtered Data

In this chapter, a case study is conducted to explore Bayesian Networks for traffic prediction purpose. This study focuses on predicting traffic average speed over validation respectively, from which we can obtain that this complete model is suitable for capturing dependence structure based on the present data set.
the whole A10 ring Amsterdam, with data collected by MONICA system in 2011. The Bayesian Networks structure employed in the following experiments have been discussed and presented in section 4.4.2. The nodes in BNs represent traffic measurements, specifically the traffic speed, for locations close to a location of interest.

Based on Traffic Flow theory, the traffic speed measurements could be taken from different locations and modeled in a Bayesian Network. The BN models are implemented and tested for evaluation in an experiment. The speed data in the Bayesian Network is obtained from the MONICA ILD system, as explained earlier in section 4.3.1. The raw MONICA data is given in average values per minute. Since ILD’s are not always reliable, there exists a certain amount of missing data, as discussed in section 4.3. The Treiber and Helbing filter was introduced to fill the missing data, based on the theory that a traffic congestion propagates upstreams with a speed of $15\text{ km/h}$.

We are not only interested in predicting traffic delays, i.e. traffic congestions, but also the traffic situation of the whole network. Hence it is decided to concentrate on afternoon rush hours and the whole day of the whole network.

The year 2011 has 365 days, leading to 525600 samples for each variable since the data are measured with one minute interval. It would be unfair to train and test the models on the same sets. The basic traffic patterns differ in weekend and working
days, or the day of week and so on. Also a too large train data set will increase the computational time significantly. Thereby it is decided to take 51 days for training and one day for testing. The choice of the training days is determined by the day of interest in testing. If the testing day is a working day, the training days should be randomly chosen from the worded days. On the other hand, if the testing day is a Saturday or Sunday, then the training days come from weekends set. This train data selecting method avoids the inhomogeneities of traffic pattern not only between train data set and test data, but in train data set itself at best. It might not be a sufficiently representative set because of the sample size, but it enhance the algorithm efficiency definitely. Efficiency is of importance because in real application, it would be meaningless if the algorithm costs, for example one hour to get a half hour time horizon prediction.

The models are trained on the train data set and tested on the test data. The performance of the models need to be computed for comparison purpose. Section 4.5.1 introduces several types of performance measure which will be further used.

4.5.1 Performance measures

There are a large number of accuracy measures available for the evaluation of the performance for the predictions. Tayman et al. [34] state that any summary measure of error should meet five highly desirable criteria:

1. Validity
2. Reliability
3. Easy to interpret
4. Clarity of presentation
5. Support of statistical evaluation

There are two common error measures widely used in application: Mean Absolute Percentage Error (MAPE) and the symmetric Mean Absolute Percentage Error (sMAPE). The MAPE can be described in mathematical form as follows:
MAPE = \frac{1}{N} \sum_{i=1}^{N} \left| \frac{A_i - P_i}{A_i} \right| \quad (4.12)

where $A_i$ denote the actual value and $P_i$ is the prediction value at time interval $i$. When MAPE reaches zero, it means the prediction value equals to the real value. While the MAPE is getting high, it means the prediction accuracy decreases. MAPE has some shortcomings. First, it lacks the validity criterion for evaluating time series since it has a tendency to overestimate the error, claimed by Tayman et. al [34]. Secondly, there is no restriction in MAPE’s upper bound, which will influence the average and makes these values hard to compare when some MAPE values are very high [22]. Furthermore, MAPE fails to deal with the situation that the real values are zeros. In this case, we shift the real value to 1 to avoid this defect.

To cope with the shortcomings along with MAPE, sMAPE was introduced in [35]. The mathematical form of sMAPE is defined as:

$$sMAPE = \frac{1}{N} \sum_{i=1}^{N} \left| \frac{A_i - P_i}{A_i + P_i} \right|$$

where $A_i$ is the actual value and $P_i$ is the prediction value at time interval $i$. The sMAPE is a more convenient form for the reason that it has the lower and upper bound, that is 0% and 100%, which makes it very convenient during the comparison of different prediction models. One problem with sMAPE is that it is not as symmetric as it sounds, since over-predictions and under-predictions are not treated equally [23]. We use these measures since they have been used in the investigation of traffic problems before.

Traffic congestion is one of the most important factors in describing traffic situation on freeways, since during free flow period the traffic speed would not change too much except the speed limit control but traffic congestion brings unexpected delay to road users. In order to analyze the model performance in congestion situation, it is needed to have a clear understanding of the definition of traffic congestion. One widely used definition for traffic congestion is given by Verkeers Informatie Dienst (Traffic Information Service) \(^8\): slow traffic, where the average vehicle speed for a road section of 2(km) is below 50(km/h). In our case, we use a slightly different definition

\(^8\)More details can be found at www.vid.nl
for traffic congestion. This is due to the fact that we use 1-minute aggregates of vehicle speeds at specific locations.

We define a traffic congestion to be a data point below 50(km/h), meaning that the average vehicle speed was below 50(km/h) for one minute.

In order to investigate the accuracy of models in predicting traffic congestion, we introduce two more error measures specially serving for traffic congestion accuracy analysis. These two measures are false positive (FP) and false negative (FN) introduced in [22]. False positives are data points where a predictor indicates a congestion, while there is none. False negatives are data points where a predictor indicates no congestion, while there is one. Thus, high FN or FP indicates that the corresponding model is inaccurate in predicting the congestions.

The mathematical formulas for FP and FN are:

\[
FP = \frac{1}{N} \sum_{i=1}^{N} 1(V_{i}^{pred} \leq 50 \text{ and } V_{i}^{real} > 50)
\]

\[
FN = \frac{1}{N} \sum_{i=1}^{N} 1(V_{i}^{pred} > 50 \text{ and } V_{i}^{real} \leq 50)
\]

where \(V_{i}^{pred}\) denotes the predicted speed at time point \(i\), and \(V_{i}^{real}\) is the real speed at time point \(i\).

For comparison purpose, we introduce two simple models besides the BN model. One is the so-called Naive model. The Naive model simply treats the current speed as the prediction speed with a given time horizon into the future. In mathematical form, the Naive prediction model can be described as:

\[
V_{i+T} = V_{i}
\]

where \(T\) denotes the prediction time horizon. Naive model is a powerful predictor because in most phases, the traffic are stable on the freeways. The vehicles tend to keep moving with sort of constant (not changing dramatically) speed within a short time period. The error increases if the prediction time horizon increases, which is a drawback of the Naive model. Another of its powerful properties is that the efficiency
because there is almost no computation time needed for the Naive model.

Another comparable model we refer to as historical model, or H model. This H model differs itself from the historical model introduced in [22] as it is simpler and more efficient. The H-model predicts the speed at time $t$ by taking the mean of speeds at the same time $t$ of the days selected in the train data at the same location. Mathematically, it can be described as:

$$V_{i,T} = \frac{1}{M} \sum_{i=1}^{M} V_{i,T}$$

(4.17)

where $M$ is the number of days selected in the train data, and $V_{i,T}$ denotes the speed measured at time $t + T$ on the $i$th day of the selected days in the train data. The H-model does not use any intelligence or any information from the test day. Thus the performance of H-model is independent of prediction time horizon. For long term prediction, the H-model might be more accuracy than Naive and BN models.

### 4.5.2 Prediction at one location

In order to obtain predictions by using BNs with Gaussian copula, the empirical distributions of the variables need to be estimated first. VBT has almost the same empirical distribution as VB since it is just a shift of VB. Next these variables are transformed to standard normals. The (conditional) rank correlations of the transformed variables are computed corresponding to each arc in the BN (see figure 4.18 and 4.20). Note that both models shown in figure 4.18 and 4.20 are saturated graph models, thus the order of the arcs has no influence on the evaluation of the correlation matrix. However, for un-saturated BN models, in most cases the order of the nodes does matter with correlation matrix calculation.

We perform the prediction with clockwise traffic direction. In this research, the statistical measures, mean and median, will be employed to present the possible predictions.

For diverse tests purpose, we choose 3 test days: the 4th of March, Friday, the 6th of June, Tuesday and the 10th of November, Thursday in 2011. Two locations are
also selected: hectometer 28.08\((km)\) and hectometer 16.3\((km)\). These two locations are interesting as the latter one lies in the vicinity of main exits which may result in congested and turbulent traffic states, and the first one lies in the middle of two main exits. The experimental prediction time horizon is set as \(T = 30\) minutes.

First we take an investigation of error measures’ performance of the three models over the selected testing data sets. The MAPE of the BN Model I (represented as BN I), BN Model II (represented as BN II), the Naive model and H model are presented in table 4.1. We can observe from the table that the difference among these models are marginal. BN I model performs stably better than BN II model. It is hard to conclude a best model as the MAPE performance of each model is not stable. Because of the shortcomings of MAPE measure as described in section 4.5.1, it is difficult to compare these MAPE values, and the conclusions should be taken with care.

Table 4.2 presents the results of sMAPE for all models of interests over the testing data sets. It is more convenient to compare sMAPE values as they range from 0\% to 100\%. In general, BN models perform better than Naive model since all sMAPE of BN models values are lower. We observe that the H model is comparable with BN models in three tests, which is not a surprise since the H model do capture the general traffic patterns. Naive model is expected to lose the accuracy when prediction time increases because of its late predicting.

It is also interesting to observe how good the models perform on the other two error measures: false positive (FP) and false negative (FN). A situation of FN is undesirable, as the unexpected traffic congestion brings inconvenience to road users. The requirements of the accuracy of FP could be less, as drivers might be happy to see that a predicted delay turns out to be absent. However, if high FP happen frequently it may result in a uneconomic traffic distribution since drivers need to re-route because of the incorrect prediction.

Table 4.3 presents the FP performance results for all models over the testing afternoon rush hours period. Clearly BN models perform the best in all testing cases except the last testing data set. It seems that the BN models have a tendency to predict free flow more often. The H model performs better than the Naive model, but worse than BN models generally.
The performance of false negatives for all models are listed in table 4.4. We have realized false negatives are undesirable for road users thus we are trying to find a model with lowest value of FN. From table 4.4 we observe that no model outperforms in all testing cases. For instance: for the testing data set on 4-th of March at hectometer 28.08(km), H model using statistic measure median has a false negative error of 0.019, while the Naive model has 0.054. This means the H model using median predicts much better than Naive model in congestion situations. Another example is the prediction on data set on 10-th of November at hectometer 28.08. BN I median has false negative value of 0.015, while the H mode using mean has a higher FN value of 0.057. This comparison indicates that the BN model performs much better the Historical model in congestion situations. However, we cannot conclude that BN model performs better than Naive or H model from table 4.4.

From Table 4.1 to 4.4, we can roughly conclude that BN I model predict better than BN II model in case of predicting traffic speed on the above mentioned testing days. Both BN models and H model performs better than the Naive model in most cases, but it is not clear which one is better between BN I model and H model. For H model itself, the median measure performs better in prediction than mean. Together with the reason that BN I model is simpler thus more efficient than the BN II model, in the following we only present BN I model instead of showing both.

<table>
<thead>
<tr>
<th>Day</th>
<th>04-03-2011</th>
<th>06-06-2011</th>
<th>10-11-2011</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hectometer (km)</td>
<td>16.3 28.08</td>
<td>16.3 28.08</td>
<td>16.3 28.8</td>
</tr>
<tr>
<td>Naive</td>
<td>0.147 0.170</td>
<td>0.143 0.240</td>
<td>0.196 0.138</td>
</tr>
<tr>
<td>BN model I mean</td>
<td>0.135 0.162</td>
<td>0.129 0.251</td>
<td>0.204 0.120</td>
</tr>
<tr>
<td>BN model I median</td>
<td>0.136 0.162</td>
<td>0.173 0.248</td>
<td>0.211 0.106</td>
</tr>
<tr>
<td>BN model II mean</td>
<td>0.152 0.164</td>
<td>0.136 0.260</td>
<td>0.210 0.119</td>
</tr>
<tr>
<td>BN model II median</td>
<td>0.143 0.168</td>
<td>0.169 0.266</td>
<td>0.219 0.110</td>
</tr>
<tr>
<td>Historical mean</td>
<td>0.138 0.137</td>
<td>0.182 0.260</td>
<td>0.219 0.138</td>
</tr>
<tr>
<td>Historical median</td>
<td>0.120 0.128</td>
<td>0.182 0.249</td>
<td>0.178 0.107</td>
</tr>
</tbody>
</table>

Table 4.1: MAPE performance measures for chosen models, experimented during afternoon rush hours on testing days and locations, with prediction time horizon 30 minutes

More interesting to see is how good the prediction models perform with different prediction time horizon, and testing not only during afternoon rush hours but the whole
4.5 : IMPLEMENTATION WITH FILTERED DATA

(a) MAPE on 4th of March, at hectometer 16.3 (km).

(b) MAPE on 4th of March, at hectometer 28.08 (km).

(c) MAPE on 6th of June, at hectometer 16.3 (km).

(d) MAPE on 6th of June, at hectometer 28.08 (km).

(e) MAPE on 10th of November, at hectometer 16.3 (km).

(f) MAPE on 10th of November, at hectometer 28.08 (km).

Figure 4.22: MAPE performance measures for chosen models, experimented on testing days and locations; BN model refers to BN model I.
(a) sMAPE on 4th of March, at hectometer 16.3(km)
(b) sMAPE on 4th of March, at hectometer 28.08(km)
(c) sMAPE on 6th of June, at hectometer 16.3(km)
(d) sMAPE on 6th of June, at hectometer 28.08(km)
(e) sMAPE on 10th of November, at hectometer 16.3(km)
(f) sMAPE on 10th of November, at hectometer 28.08(km)

Figure 4.23: sMAPE performance measures for chosen models, experimented on testing days and locations; BN model refers to BN model I
4.5: IMPLEMENTATION WITH FILTERED DATA

(a) FN on 4th of March, at hectometer 16.3 km
(b) FN on 4th of March, at hectometer 28.08 km
(c) FN on 6th of June, at hectometer 16.3 km
(d) FN on 6th of June, at hectometer 28.08 km
(e) FN on 10th of November, at hectometer 16.3 km
(f) FN on 10th of November, at hectometer 28.08 km

Figure 4.24: FN performance measures for chosen models, experimented on testing days and locations; BN model refers to BN model I
Figure 4.25: FP performance measures for chosen models, experimented on testing days and locations; BN model refers to BN model I
Day | 04-03-2011 | 06-06-2011 | 10-11-2011
--- | --- | --- | ---
Hectometer (km) | 16.3 28.08 | 16.3 28.08 | 16.3 28.8
Naive | 0.063 0.071 | 0.072 0.079 | 0.065 0.060
BN model I mean | 0.049 0.064 | 0.058 0.072 | 0.048 0.052
BN model I median | 0.043 0.061 | 0.079 0.069 | 0.042 0.041
BN model II mean | 0.044 0.069 | 0.079 0.071 | 0.041 0.051
BN model II median | 0.051 0.068 | 0.081 0.083 | 0.043 0.047
Historical mean | 0.052 0.069 | 0.073 0.072 | 0.055 0.058
Historical median | 0.043 0.055 | 0.073 0.074 | 0.041 0.044

Table 4.2: sMAPE performance measures for chosen models, experimented during afternoon rush hours on testing days and locations, with prediction time horizon 30 minutes

Day | 04-03-2011 | 06-06-2011 | 10-11-2011
--- | --- | --- | ---
Hectometer (km) | 16.3 28.08 | 16.3 28.08 | 16.3 28.8
Naive | 0.048 0.053 | 0.063 0.029 | 0.042 0.021
BN model I mean | 0 0.027 | 0.006 0 | 0 0.012
BN model I median | 0.004 0.031 | 0.021 0.003 | 0.008 0.021
BN model II mean | 0 0.028 | 0.008 0 | 0 0.015
BN model II median | 0.005 0.031 | 0.013 0.003 | 0 0.025
Historical mean | 0.048 0.053 | 0.010 0.005 | 0.003 0.005
Historical median | 0.045 0.057 | 0.009 0.002 | 0.001 0.002

Table 4.3: FP performance measures for chosen models, experimented during afternoon rush hours on testing days and locations, with prediction time horizon 30 minutes

day. For this purpose, we experimented the prediction on the whole selected days and locations. The error measure performance are presented in Figure 4.22, 4.23, 4.24 and 4.25. Based on the above analysis, only the results of BN I mean, BN I median, H median and the Naive model are presented.

Figure 4.22(a) shows the error measure MAPE performance on the testing day of 4-th of March at hectometer 16.3(km). The figure shows that the BN model performs consistently better than the Naive model for all prediction time horizons. When the prediction time horizons are less than 18 minutes, BN model perform better than H median but worse in case of prediction time horizons are big than 20 minutes. This should not come as a surprise since BN model tend to lose the accuracy when the prediction time horizon grows, while the H model has no relation with the information on the testing day. From sMAPE performance shown in Figure 4.23(a) we can get the
<table>
<thead>
<tr>
<th>Day</th>
<th>04-03-2011</th>
<th>06-06-2011</th>
<th>10-11-2011</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hectometer (km)</td>
<td>16.3 28.08</td>
<td>16.3 28.08</td>
<td>16.3 28.8</td>
</tr>
<tr>
<td>Naive</td>
<td>0.047 0.054</td>
<td>0.053 0.044</td>
<td>0.042 0.021</td>
</tr>
<tr>
<td>BN model I mean</td>
<td>0.073 0.038</td>
<td>0.103 0.078</td>
<td>0.072 0.019</td>
</tr>
<tr>
<td>BN model I median</td>
<td>0.058 0.023</td>
<td>0.103 0.061</td>
<td>0.064 0.015</td>
</tr>
<tr>
<td>BN model II mean</td>
<td>0.075 0.033</td>
<td>0.103 0.098</td>
<td>0.075 0.023</td>
</tr>
<tr>
<td>BN model II median</td>
<td>0.064 0.038</td>
<td>0.102 0.086</td>
<td>0.070 0.018</td>
</tr>
<tr>
<td>Historical mean</td>
<td>0.033 0.049</td>
<td>0.101 0.099</td>
<td>0.057 0.057</td>
</tr>
<tr>
<td>Historical median</td>
<td>0.025 0.019</td>
<td>0.101 0.078</td>
<td>0.059 0.056</td>
</tr>
</tbody>
</table>

Table 4.4: FN performance measures for chosen models, experimented during afternoon rush hours on testing days and locations, with prediction time horizon 30 minutes.

same conclusion.

Figure 4.22(b) present the error measure MAPE performance on the testing day of 4-th of March at hectometer 28.08(km). From this figure we can observe that H median outperforms among all models, and BN model performs better than the Naive model if the prediction time horizon is longer than 20 minutes. The same applied to sMAPE as shown in Figure 4.23(b).

From Figure 4.22(c) it becomes clear that BN I mean outperforms all the other models when the prediction time horizons are bigger than 12 minutes and less than one hour. However, sMAPE performance shows a slight difference as shown in Figure 4.23(c), in which the H model performs better than BN I mean when the prediction time horizon is above 52 minutes. It indicates that the traffic situation for this typical testing day is out of the general traffic patterns. Many reasons might cause the unexpected traffic pattern, such as public holiday, bad weather, accidents and so on.

In Figure 4.22(d) BN model tends to perform better than the Naive model when the prediction time horizon is longer than 42 minutes approximately. From sMAPE performance point of view, this tendency turns out to be much earlier. When the prediction time horizon has not arrived 20 minutes, the BN model shows a better prediction results than the Naive model based on sMAPE as shown in Figure 4.23(d). Both MAPE and sMAPE performance shows that the H model starts to outstand among all models after the prediction time horizon arrives at 30 minutes around.

For the remaining two testing data sets for a regular Tuesday in November shown
in Figure 4.22(e)(f) and Figure 4.23(e)(f), it is clear that BN model tends to predict better than the Naive model when the prediction time horizon grows. This is reasonable since the Naive model simply takes the current speed as the prediction, thereby as prediction time horizon grows, the delay increases which brings bigger error.

Figure 4.24 presents the false negative performance for all testing data sets. From FN point of view, BN I model does not show a outstanding performance in prediction. The possible reason could be that BN I model sometimes fails to capture the congestion but Naive model always can as long as the prediction time horizon is shorter than the congestion lasting time. Nevertheless, BN I model performs much better the Naive model, as shown in Figure 4.25. It indicates that BN I model is able to predict the transformation from congestion to free flow better.

In conclusion, BN I model could be a choice when the prediction time horizon falls in time interval [20, 40] minutes. If the prediction time horizon is lower than 20 minutes, the Naive model is recommended. And if the prediction time horizon is above 40 minutes, H model could be used for prediction.

4.5.3 Prediction over the network

In section 4.5.2 the prediction analysis for a single location is presented. It is hard and not sufficient to conclude an "optimal" model only based on the models’ performance at a single location. More important, for road users as drivers and road managers, the traffic information supplied about one or a few location points does not help too much as the vehicles are moving along the road rather than staying at one location. Thereby, an overview description of traffic situation into the future on a piece of road or a network is needed to help the driver to prepare for the potential coming traffic congestion. This type of information will also support the road managers in traffic control. All these elements motivates us to generate an algorithm to achieve the prediction of a road network, for example A10 ring.

Figure 4.26(a) presents a traffic information overview about A10 ring Amsterdam with clockwise traffic direction, on 3rd of March in 2011. The X axis presents time in minutes and the Y axis presents the locations from hectometers 1.4(km) to 33.08(km).
(a) Traffic information over A10 ring within 24 hours

(b) Traffic information over A10 ring at time point 08:46am

**Figure 4.26:** An example of traffic information over A10 ring Amsterdam on the 3rd of March, in 2011

From this figure we can observe that the traffic congestion appears in afternoon of the road section between hectometers 15\((km)\) and 30\((km)\). There exists a short period congestion in the morning from hectometer 10\((km)\) to 17\((km)\). The yellow part might be the result of traffic speed control. A more convenient view of traffic situation is presented in figure 4.26(b), in which the circle represents the trace of A10 ring. This figure shows a overview at given time point 08:46am. The dark red part indicates a heavy congestion, while the light red represents a coming or a fading traffic congestion.

The idea of nodes selection sketched in section 4.5.2 is based on traffic flow
theory and the fact that perturbations in traffic propagates with a certain speed and along certain directions. Applying the same idea to each location on the A10 ring makes it possible to predict the traffic situation over the whole network. In this section we implement the prediction experiment on the basis of data set collected by MONICA system and filtered using Treiber and Helbing filter. Details about data acquisition and processing are discussed in section 4.3. The filtered data set with clockwise traffic direction contains 318 location points, starting from hectometer 1.4\((km)\) to 33.08\((km)\). We arrange the data set with columns representing data at each location, in which from left to right the hectometer increases with step size 0.1\((km)\). Each column in the filtered data set contains the speed data measured per minutes of 365 days in 2011. Thereby, the sample size of the filtered data set is 525,600 \(\times\) 318. In order to predict the speed for each location, we select the nodes based on the traffic theory as described in section 4.5.2.

Since BN Model I performs better than BN Model II based on the analysis in section 4.5.2, together with its simplicity we decided to use BN Model I as the prediction model for each location. Historical model using statistical measure median is employed as a comparison. The train data is generated by 51 days data randomly selected from 365 days in the whole year. For the testing days and prediction time horizon we select:

- the 18th of July, Monday, with prediction time horizon 20 minutes;
- the 4th of March, Friday, with prediction time horizon 30 minutes;
- the 10th of November, Tuesday, with prediction time horizon 30 minutes;
- the 30th of September, Friday, with prediction time horizon 40 minutes.

Figure 4.27 presents the prediction results and error measures’ performance on the testing day - 18th of July, Monday in 2011 - of A10 ring network. From figure 4.27 it becomes clear that both models BN I mean and BN I median are able to capture the congestion period. The two main congestion parts in the real data and BN prediction are shown approximately at the same time and location coordinates with similar size. The errors of BN model occurs at the edge of the congestion area and the heavy congestion horizon (with color dark red). It indicates that the BN model tends to
underestimate the congestion. The H median fails to predict the short congestion period around hectometer 16(km)-21(km). Figures 4.27(b)-4.27(d) present the error measure performance of each model. It may be observed that H mean performs the worst, while the other four models turn out to have close performance except that H median outperforms at locations between hectometers 24(km) and 28(km). The prediction models have more stochastic fluctuations than the Naive model for all error measures.

For the sake of completeness, the network prediction results and error performance on day 4th of March, Friday, with prediction time horizon 30 minutes are presented in Figure 4.28. From the speed map one may see that the three models are able to predict the main parts of congestions, though the heavy traffic congestion situation (dark red) is underestimated. For H model there exists a gap in up-right red square, resulting high MAPE and sMAPE as shown in Figure 4.28(b) and 4.28(c). Not surprisingly, we observe that the blue lines, representing the error measure of the Naive model, are generally above the other lines, which indicates that the Naive model performs worse. At first glance, it seems hard to tell the pros and cons between BN model and H model as their error lines cross with each other. However, because of its intense fluctuation of H model shown in MAPE and sMAPE performance, we may say BN model is slightly better than H model. The false positive and false negative performance of each model turns out to be complicated. Generally, BN model performs better than the Naive model from FP and FN point of view.

Another example to describe the predictions with prediction time horizon 30 minutes took the testing day on 10th of November, Tuesday in 2011. Figure 4.29 presents the prediction and error measure results. From the visualized speed prediction map shown in Figure 4.29(a), one may see that all three models capture the afternoon congestion occurred between approximately hectometers 25(km) and 30(km). However, the congestion part represented by the up-right red square in real speed map was not well predicted by these three models. H median overestimates the morning rush hour traffic, while BN model underestimates it. From MAPE and sMAPE performance shown in Figure 4.29(b) and (c) we can observe that BN model outperforms among all models in general. For false positive, lines representing BN
model lie at the bottom, indicating BN model tends to predict the free flow traffic better than the other models in this testing case. The peaks appearing in false negative performance shown in Figure 4.29(e) of BN model are caused by its failure of predicting three congestions as explained at the beginning of this paragraph.

We are interested to see the prediction with longer prediction time horizon, say 40 minutes for instance. For this purpose, an experiment for testing day 30th of September, Friday with prediction time horizon 40 minutes was implemented and, the results are shown in Figure 4.30. We observe that three models fail to predict the congestions except the regular afternoon rush hour happening between hectometer 25(km) and 30(km). This should not come as a surprise because both BN model and H model are historical data based model. To some extent, BN model is able to predict the unexpected congestions caused by accident, terrible weather or else as shown in Figure 4.27(a). However, as prediction time horizon increases, BN model loses the accuracy in prediction because of the longer delay in obtaining observations. For MAPE performance three models behave very close, while from sMAPE performance we can observe that BN model performs better the Naive model still. For FP and FN error measure performance, BN model slightly outperforms among all models.

<table>
<thead>
<tr>
<th>Models</th>
<th>MAPE</th>
<th>sMAPE</th>
<th>False positive</th>
<th>False negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naive</td>
<td>0.043</td>
<td>0.020</td>
<td>0.004</td>
<td>0.004</td>
</tr>
<tr>
<td>BN model I mean</td>
<td>0.044</td>
<td>0.020</td>
<td>0.003</td>
<td>0.005</td>
</tr>
<tr>
<td>BN model I median</td>
<td>0.039</td>
<td>0.017</td>
<td>0.003</td>
<td>0.004</td>
</tr>
<tr>
<td>Historical median</td>
<td>0.038</td>
<td>0.017</td>
<td>0.001</td>
<td>0.004</td>
</tr>
</tbody>
</table>

Table 4.5: Error measures’ performance for chosen models, experimented on the 18th of July, 2011, with prediction time horizon 20 minutes

<table>
<thead>
<tr>
<th>Models</th>
<th>MAPE</th>
<th>sMAPE</th>
<th>False positive</th>
<th>False negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naive</td>
<td>0.107</td>
<td>0.041</td>
<td>0.013</td>
<td>0.025</td>
</tr>
<tr>
<td>BN model I mean</td>
<td>0.094</td>
<td>0.033</td>
<td>0.004</td>
<td>0.019</td>
</tr>
<tr>
<td>BN model I median</td>
<td>0.087</td>
<td>0.031</td>
<td>0.006</td>
<td>0.017</td>
</tr>
<tr>
<td>Historical median</td>
<td>0.116</td>
<td>0.033</td>
<td>0.011</td>
<td>0.011</td>
</tr>
</tbody>
</table>

Table 4.6: Error measures’ performance for chosen models, experimented on the 4th of March, 2011, with prediction time horizon 30 minutes

For a more convenient and general view of the performance for the three
(a) Speed map over A10 ring network. Prediction time horizon: 20 minutes. The color bar represents the speed

(b) MAPE performance

(c) sMAPE performance

(d) False positive performance

(e) False negative performance

Figure 4.27: Predictions over the whole network of A10 ring, on testing day 18th of July 2011, Monday. Prediction time horizon: 20 minutes
4.5: IMPLEMENTATION WITH FILTERED DATA

(a) Speed map over A10 ring network

(b) MAPE performance

(c) sMAPE performance

(d) False positive performance

(e) False negative performance

Figure 4.28: Predictions over the whole network of A10 ring, on testing day 4th of March 2011, Friday. Prediction time horizon: 30 minutes
CHAPTER IV: TRAFFIC INFORMATION MODELLING

(a) Speed map over A10 ring network

(b) MAPE performance

(c) sMAPE performance

(d) False positive performance

(e) False negative performance

**Figure 4.29:** Predictions over the whole network of A10 ring, on testing day 10th of November 2011, Tuesday. Prediction time horizon: 30 minutes
4.5 : IMPLEMENTATION WITH FILTERED DATA

(a) Speed map over A10 ring network

(b) MAPE performance

(c) sMAPE performance

(d) False positive performance

(e) False negative performance

Figure 4.30: Predictions over the whole network of A10 ring, on testing day 30th of September 2011, Friday. Prediction time horizon: 40 minutes
models, we average the error measure of all locations. The results are presented in Tables from 4.5 to 4.8. Generally, BN I model is better in predicting unexpected congestions than H model, and in drawing the comprehensive traffic situation than the Naive model. This superiority is constrained by the length of prediction time horizon. As prediction time horizon grows, BN model tends to lose the accuracy. Within prediction time horizon interval $[20 \text{ minutes}, 40 \text{ minutes}]$, BN I model would be a good choice to predict the whole network traffic information.

### 4.6 Implementation with Raw Data

The filtered data presents a complete graph of traffic on A10 ring, and further provides an opportunity for us to predict the traffic information over the whole network. However, the filtration itself performs as a predictor, thus brings extra variation to the data. BN model shows a promising prediction on the basis of the filtered data, as discussed in the previous section. How BN model would perform on the basis of the raw data becomes a following topic engineers concern about.

The raw data collected by MONICA system has 15% – 20% missing data approximately. The missing data are recorded by zeros. Further, the locations collecting the traffic data are not as dense as the filtered data. The raw data of 2011 on A10 ring consists by hand consists of $525,600 \times 76$ data, where the columns represent

<table>
<thead>
<tr>
<th>Models</th>
<th>MAPE</th>
<th>sMAPE</th>
<th>False positive</th>
<th>False negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naive</td>
<td>0.093</td>
<td>0.037</td>
<td>0.014</td>
<td>0.014</td>
</tr>
<tr>
<td>BN model I mean</td>
<td>0.094</td>
<td>0.032</td>
<td>0.002</td>
<td>0.019</td>
</tr>
<tr>
<td>BN model I median</td>
<td>0.091</td>
<td>0.029</td>
<td>0.002</td>
<td>0.019</td>
</tr>
<tr>
<td>Historical median</td>
<td>0.098</td>
<td>0.035</td>
<td>0.013</td>
<td>0.017</td>
</tr>
</tbody>
</table>

**Table 4.7:** Error measures’ performance for chosen models, experimented on the 10th of November, 2011, with prediction time horizon 30 minutes

<table>
<thead>
<tr>
<th>Models</th>
<th>MAPE</th>
<th>sMAPE</th>
<th>False positive</th>
<th>False negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naive</td>
<td>0.112</td>
<td>0.038</td>
<td>0.016</td>
<td>0.016</td>
</tr>
<tr>
<td>BN model I mean</td>
<td>0.098</td>
<td>0.029</td>
<td>0.002</td>
<td>0.016</td>
</tr>
<tr>
<td>BN model I median</td>
<td>0.095</td>
<td>0.026</td>
<td>0.002</td>
<td>0.015</td>
</tr>
<tr>
<td>Historical median</td>
<td>0.097</td>
<td>0.026</td>
<td>0.002</td>
<td>0.016</td>
</tr>
</tbody>
</table>

**Table 4.8:** Error measures’ performance for chosen models, experimented on the 30th of September, 2011, with prediction time horizon 40 minutes
4.6: IMPLEMENTATION WITH RAW DATA

(a) Hectometers of raw data
(b) Speed map of raw data

Figure 4.31: Hectometers and speed map on 4th of March, of raw data

hectometers and rows represent time. Note that the raw data is much sparser than the filtered, as the column number reduces from 318 to 76. The locations distribute continuously as shown in Figure 4.31(a), and Figure 4.31(b) presents a speed map example on day 4th of March, in which the white part represents the missing data.

As an experiment and the first step investigation, we inherited the BN algorithm described in section 4.4 but simplified it with two rules:

- If at time point \( t \), there exists at least one zero in the train data at time \( t \), then the prediction is set as zero;

- If at time point \( t \), there exists at least one zero among observations at time \( t \), then the prediction is set as zero.

Since we are only aiming to test the qualification of BN model on raw data, Historical model and the congestion error measure FN and FP will not be excluded. BN I model is employed in experiment.

Figure 4.32 presents a prediction on basis of raw data, on day 4th of March, with prediction time horizon 30 minutes. From speed map we can observe that BN predictions have more missing values than real data, resulted by the above two rules. The main congestion parts are well predicted by BN I model. Figure 4.32(b) shows the error measure MAPE performance. Note that most blue stars locate above the red and green lines except predictions around hectometer 27(km). It indicates that generally
BN I model performs better than the Naive model. The sMAPE performance shown in Figure 4.32(c) again corroborates that BN I model predicts better in this case.

Figure 4.33 presents a prediction on basis of raw data, on day 18th of June, with prediction time horizon 30 minutes. From the error measure performance shown in Figure 4.33(b) and (c), we see that BN I median performs better than BN I mean, and the Naive model performs the worst.

As a short conclusion, BN model works well on raw data as well. Besides, BN model can be used as a filter tool. By using BNs, as long as there exists one observation the prediction can be obtained. We will not investigate further on this topic but leave it as the future work.
4.6 : IMPLEMENTATION WITH RAW DATA

(a) Speed map over A10 ring network

(b) MAPE performance

(c) sMAPE performance

Figure 4.32: Predictions over the whole network of A10 ring, on testing day 4th of March 2011, Friday. Prediction time horizon: 30 minutes
(a) Speed map over A10 ring network

(b) MAPE performance

(c) sMAPE performance

**Figure 4.33:** Predictions over the whole network of A10 ring, on testing day 18th of July 2011, Monday. Prediction time horizon: 30 minutes
CHAPTER V
CONCLUSIONS AND FUTURE WORK

Several conclusions have been presented throughout the research. We summarize the key findings and discuss the possible future relevant work in this chapter.

5.1 Conclusions
5.1.1 Theoretical part

The investigation of Bayesian Networks presents a way to model the uncertainties such as the traffic domain. There are different methods to train the data. Among them the Non-parametric Bayesian Network using copulas shows its advantages in efficiency. In a Continuous Non-parametric Bayesian Network (CNBN), the nodes are associated with continuous invertible distribution functions, and the arcs are assigned with (conditional) rank correlations, as the parameter of chosen copula.

In our research, the Gaussian copula was applied. The Gaussian copula does not always represent the data fairly, thereby Gaussian tests are needed to check the validation of Gaussian copula. Three tests: Bivariate Gaussian copula test, Blanket test and D-test are applied in this research.

5.1.2 Project 1: Traffic Loads in Bridges

Traffic loads plays an important role in determining the bending moment in bridges, and further in safety. The main two goals for this project were to gain the traffic configuration given the out of sample bending moment, and the dependent relationship between traffic configuration and bridge bending moment.

The data acquisition for this project were generated by Monte-Carlo simulation.
Then they were categorized into N sections to form the variables further used in modelling. The first goal of this project motivates us to conditionalize on out of sample values, thus parametric margin fitting is needed. Normal and mixed Normal margin fitting methods were applied.

We have presented CNBNs application in modeling the dependence structure of bridge bending moment and vehicle configuration and, information updating. The BNs were built by using data mining in the software UNINET. The inference results generated in UNINET are treated as benchmarks. The values of relative difference measures $RD_N$ and $RD_M$ show that the mixed Normal margin fitting models perform better than the Normal margin fitting model mostly, especially for the variables with multi clusters. When there appears negative values in statistical measures after information updating, re-conditionalization needs to be applied. For most variables, re-conditionalization does improve the accuracies.

Since the data sets can be simulated with different bridge length $L$ and auto correlation $r(S_t, S_{t-1})$ parameters. The analysis shows that the length of bridge $L$ has significantly influence in determining the bridge bending moment. Not only the maximal but also the mean bending moments of samples of 100 meters long bridge are lower than those of 50 meters long bridge. The bending moment is less sensitive to the auto correlation parameter than $L$. Generally, the higher the auto correlation is, the bigger the bending moment would be.

5.1.3 Project 2: Traffic Information Modelling

Traffic information is of importance for both road users and road managers. Among traffic situation measures, traffic speed is valuable as it can be used for calculating travelling time, which further provides the travellers the information in advance to arrange the schedule or route. Throughout this project, we were aiming to predict the traffic speed at A10 ring Amsterdam in the Netherlands, with certain prediction time horizon.

The data for this traffic were collected by MONICA system in 2011, supported by the Dutch Ministry of Transport, Public Works and Water Management in the Netherlands. There exists approximately 12% – 20% missing data in raw data. A
state-of-art algorithm, the Treiber and Helbing filter, was used for filling these missing values.

Traffic speed turns out to have different patterns, such as working days’ traffic pattern, weekend traffic pattern and so on. Three models: Naive model, CNBN model and Historical model were employed in this research. The prediction was experimented with both filtered data and the raw data.

For the prediction with filtered data, by analyzing the error measures we concluded that these three models perform differently. When the prediction time horizon is less than 20 minutes, the Naive model tends to outperform among them. When the prediction time horizon falls between 20 minutes and 40 minutes, BN model could be the best choice for prediction. It is able to predict the traffic situation with an accuracy of around 95% for a prediction horizon of 20 minutes, 91% for a prediction horizon of 30 minutes and 90% for a prediction horizon of 40 minutes, based on error measure MAPE. Both the Naive model and BN model would lose the accuracy when the prediction time horizon grows. When the prediction horizon is longer than 40 minutes, the Historical model tends to be more accurate than the other two models in predicting traffic speed.

The raw data has missing values, thus makes it difficult to model. BN model manages to predict the traffic speed but the accuracy is worse than prediction on the filtered data. With prediction time horizon 30 minutes, the BN model has approximately 88% accuracy.

5.2 Future Work

Because of the constraint of time, we have to stop our research here. In this section we discuss the possible future work that can be done.

In the "Traffic Loads in Bridges" project, we discussed that the Gaussian copula does not always represent the data fairly. Besides the three Gaussian copula testing methods we presented in this thesis, there exists more testing methods in literatures. One work could be done in the future is to apply other testing methods to check the suitability of Gaussian copula assumption. Or one can study the applications of the other parametric copulas, or mixed copula to measure dependence.
Later we fit Normal and mixed Normal distribution to be variables’ margins. The choice of other distributions could be a topic to be further studied. When processing the data, we categorized the data into $N$ sections, and cases $N = 1$ and $N = 3$ were investigated. In order to gain a more accurate traffic configuration, more categorizations or higher $N$ could be tried in the future.

In "Traffic Information Modelling" project, we propose to train our models on different locations on Dutch highways. We have presented a network prediction over A10 ring Amsterdam by using the Naive model, BN model and Historical model. It is interesting to propagate the prediction to the other road piece, and then generate a bigger network prediction.

The models in this research can be further improved. van den Haak [22] introduced a hybrid model, which combined the BN model with the Historical model. The improvement was promising. We propose to build a hybrid model by mixing the CNBN mode and the Historical model, as one of the future work. Last but not least, more environmental variables such as weather, incidents can be taken into consideration for modelling.

In the future, we hope that the investigation of traffic loads in bridges will make contributions to the bridge safety assessments. The predicted traffic speed can be used for calculating the travel time, which further provides traveller an accurate route recommendation.


