Delft University of Technology

Mixed Discrete-Continuous Railway Disruption-Length Models with Copulas

Zilko, Aurelius

DOI
10.4233/uuid:a551b9a2-b5da-4a51-8a3b-3d7f410d67cc

Publication date
2017

Document Version
Final published version

Citation (APA)
Zilko, A. (2017). *Mixed Discrete-Continuous Railway Disruption-Length Models with Copulas*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:a551b9a2-b5da-4a51-8a3b-3d7f410d67cc

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright
Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy
Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

This work is downloaded from Delft University of Technology.
For technical reasons the number of authors shown on this cover page is limited to a maximum of 10.

MIXED DISCRETE-CONTINUOUS RAILWAY DISRUPTION-LENGTH MODELS WITH COPULAS

AURELIUS ARMANDO ZILKO

COLOPHON

Aurelius Armando Zilko
Mixed Discrete-Continuous Railway Disruption-Length Models with Copulas, Delft, 2016
ISBN: 978-94-6186-776-6

Copyright © 2016 by A. A. Zilko
All rights reserved. No part of the material protected by this copyright notice may be reproduced or utilized in any form or by any means, electronic or mechanical, including photocopying, recording or by any information storage and retrieval system, without the prior permission of the author.

Cover designed by the author.
Printed in the Netherlands.

MIXED DISCRETE-CONTINUOUS RAILWAY DISRUPTION-LENGTH MODELS WITH COPULAS

Proefschrift

ter verkrijging van de graad van doctor
aan de Technische Universiteit Delft,
op gezag van de Rector Magnificus Prof. ir. K.C.A.M. Luyben,
voorzitter van het College voor Promoties,
in het openbaar te verdedigen op maandag 30 januari 2017 om
12:30 uur

door

Aurelius Armando ZILKO

Master of Science in Applied Mathematics
Technische Universiteit Delft, Nederland
geboren te Yogyakarta, Indonesië.

This dissertation has been approved by the

promotor: Prof. dr. F.H.J. Redig
copromotor: Dr. D. Kurowicka

Composition of the doctoral committee:

Rector Magnificus	chairperson
Prof. dr. F.H.J. Redig	Delft University of Technology
Dr. D. Kurowicka	Delft University of Technology

Independent members:

Prof. dr. G. Jongbloed	Delft University of Technology
Prof. dr. R.P.B.J. Dollevoet	Delft University of Technology
Prof. dr. D. Huisman	Erasmus University Rotterdam
Dr. A.K. Nikoloulopoulos	University of East Anglia, Norwich, England

Other member:

Dr. R.M.P. Goverde	Delft University of Technology

This research was funded by ExploRail, a collaboration between the Netherlands Organisation for Scientific Research (NWO), through Technology Foundation STW, and ProRail, project no. 12257: “Smart information and decision support forrailway operation control centres (SmartOCCR)”.

Contents

1 Introduction **1**

1.1 The Dutch Railway Network and Train Traffic Disruption 1

1.1.1 Types of Disruption 3

1.1.1.1 The Track Circuit (TC) 4

1.1.1.2 The Switch 7

1.1.2 Disruption Response Process 11

1.1.3 The SAP Database 12

1.2 Modelling Disruption Length 14

1.2.1 Modelling the Joint Distribution and the Copula 15

1.2.2 Vines 19

1.2.3 Discrete Variables 23

1.2.4 Parameter Estimation and Model Simplification 25

1.2.4.1 Parameter Estimation 25

1.2.4.2 Model Simplification 26

1.2.5 Literature Study 27

1.2.5.1 Copulas and BN in Transport Research 27

1.2.5.2 Copula Parameter Estimation with Discrete and Continuous Variables 28

1.2.6 Model Comparison 29

1.2.6.1 The Kullback-Leibler (KL) Divergence Test 30

1.2.6.2 The Probability Integral Transform (PIT) Test . . 30

1.2.6.3 The Vectorial Independence Test 31

1.2.6.4 The Akaike Information Criterion 32

1.2.6.5 The Bayesian Information Criterion 32

1.2.6.6 The Likelihood Ratio Test 32

1.2.6.7 The Vuong Test 33

1.2.6.8 Clarke's Distribution-Free Test 33

1.3 The Disruption-Length Models in Practice 34

1.4 Thesis Organization . . . 34

2 Copulas in a Mixed Discrete-Continuous Model 37
2.1 Multivariate Bernoulli Distribution with Copulas . . . 38
  2.1.1 Bivariate Bernoulli Distribution with Copulas . . . 39
  2.1.2 Trivariate Bernoulli Distribution with Copulas . . . 45
  2.1.3 The Non-Constant Conditional Copula-Vine Approach . . . 47
2.2 Mixed Discrete-Continuous Distributions with Copulas . . . 51
  2.2.1 Bivariate Case . . . 51
  2.2.2 Multivariate Case . . . 53
2.3 Modelling Mixed Data with the Copula-Vine Approach . . . 53
  2.3.1 An Algorithm to Obtain a Parsimonious Non-Constant Conditional Copula-Vine Model . . . 55
  2.3.2 Testing Algorithm 2.1 . . . 55
  2.3.3 Misspecification in Modelling . . . 61
2.4 Chapter Summary . . . 69

3 Construction of the Railway Disruption-Length Model 71
3.1 Data Analysis . . . 72
  3.1.1 Factors Influencing Latency Time . . . 72
    3.1.1.1 Time . . . 72
    3.1.1.2 Location . . . 73
    3.1.1.3 Weather and Overlapping Disruptions . . . 77
  3.1.2 Factors Influencing Repair Time . . . 78
    3.1.2.1 Contract Type . . . 78
  3.1.3 Causes . . . 79
3.2 Model Construction . . . 82
  3.2.1 The Multivariate Normal Copula Model . . . 83
  3.2.2 The Copula-Vine Model . . . 87
3.3 Model Comparison and Validation . . . 89
3.4 The Switch Disruption-Length Model . . . 91
  3.4.1 Influencing Factors . . . 91
    3.4.1.1 Cold Temperature . . . 91
    3.4.1.2 Causes of Switch Failures . . . 92
  3.4.2 The Model . . . 93
3.5 Chapter Summary . . . 96

4 The Model in Practice 97
4.1 Predicting Railway Disruption Lengths with the Models . . . 98
4.2 The Short-Turning and Passenger-Flow Models . . . 101
  4.2.1 The Short-Turning Model . . . 101
  4.2.2 The Passenger-Flow Model . . . 103
  4.2.3 Interaction Between the Three Models . . . 104
4.3 The Experiment . . . 105
  4.3.1 Experimental Setup . . . 105
  4.3.2 Case Studies . . . 109

        4.3.2.1 Case Study 1 109
        4.3.2.2 Case Study 1A 112
        4.3.2.3 Case Study 2 113
        4.3.2.4 Case Study 2A 115
    4.3.3 Discussion 117
  4.4 Chapter Summary 118

5 Summary, Recommendations, and Final Remarks **119**
  5.1 Thesis Summary 119
    5.1.1 Copulas in Multivariate Mixed Discrete-Continuous Problems 119
    5.1.2 The Railway Disruption-Length Models 120
  5.2 Recommendations 122
    5.2.1 Better Data Collection 122
    5.2.2 Models Expansion 124
  5.3 A Final Remark 125

References **127**

Appendicies

A Proof of Propositions **135**

B Experiments Results in Chapter 2 **139**

C Parameters of the Disruption Length Models **149**

Summary **155**

Samenvatting **157**

Ringkasan **161**

Acknowledgments **165**

Curriculum Vitae **167**

List of Publications **169**

CHAPTER 1

Introduction

1.1 The Dutch Railway Network and Train Traffic Disruption

The Dutch railway network is the busiest in Europe, with more than twenty-five active train operators as of 2015 ([ProRail, 2015a]). In 2014, the network operated 3.2 million train services and carried, on average, 1.1 million passengers each day totalling 145 million kilometres of passenger journeys for the year ([ProRail, 2015b]). Figure 1.1 shows the main-line train service map of the Netherlands, based on the 31 October 2012 timetable. This very dense network is managed and maintained by ProRail, the Dutch state-owned company whose main task is to assure that the railway system is reliable and safe for train operations. This includes producing and maintaining the railway timetable.

In practice, it is unavoidable that the operation of a railway system encounters unexpected incidents which disrupt the timetable. Depending on the length of the incident, different measures need to be taken to handle the situation during the downtime. Shorter incidents may require only timetable adjustment, while longer incidents may additionally require rolling stock and crew adjustment. These disruptions are the focus of this thesis.

Disruption length needs to be defined in a more precise manner. Adapting from highway traffic modelling (Highway Capacity Manual [2010]; Pereira et al. [2013]), we have divided it into four periods.

1. **Reporting time** is the interval between the actual occurrence of the incident and the moment when it is reported to the operator.
2. **Latency time** is the interval between the reporting time and the moment when the repair team arrives at the site.
3. **Repair time** is the time needed for the repair team to solve the problem.
4. **Recovery time** is the time needed for the train service to return to normal.

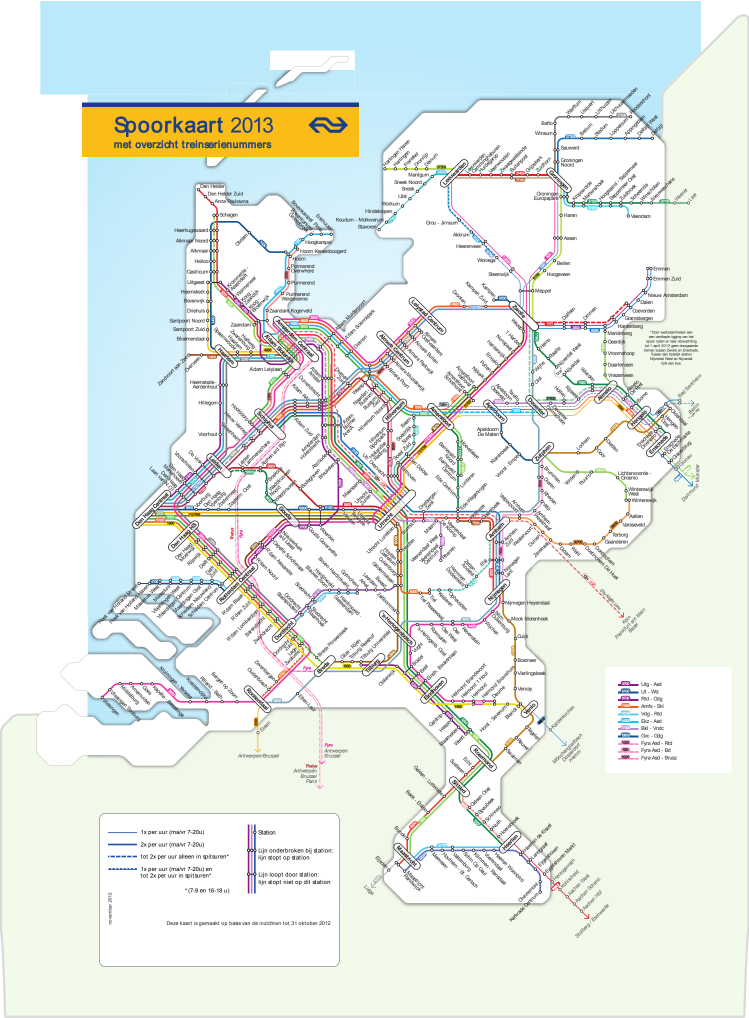


Figure 1.1: *Overview of Dutch train services in 2013.. Source: ProRail*

The first three periods represent the time when a section of railway is unavailable due to the unexpected incident. During the recovery time, the affected section has been reopened for train operation but traffic is still disrupted due to the earlier closure. The goal of this thesis is to construct a prediction model for the moment an affected section is ready to be reopened for use. Assuming that the reporting time is instantaneous, our focus is on the latency and repair time.

1.1.1 Types of Disruption

Many different unexpected incidents can disrupt operational train traffic. In general, these can be grouped into two types: technical and non-technical.

A complex railway network like the one in the Netherlands comprises a lot of different technical components. For example, as of 2012 the Dutch network consisted of 7,033 km of railway tracks, 2,731 level crossings, 7,195 switches (sets of points) and 11,683 signals ProRail [2015b]. These components fail over time for many different reasons, in a stochastic manner, and it is important to have the situation under control. The best way to do this is to prevent failures from happening in the first place, which is done through *maintenance* work. However, it is not uncommon to encounter situations where unexpected failures occur outside the maintenance schedule. Disruptions caused by such failures are said to be of a *technical* nature.

The various types of technical disruption are listed below, along with the number of registered urgent incidents on the Dutch railway network between 1 January 2011 and 30 June 2013 which we observe in our database (to be discussed shortly).

1. Switch (points) failures (2,974).
2. Track Circuit (TC) failures (2,113).
3. Signal failures (706).
4. Rail bar problems (592).
5. Rail foundation problems (285).
6. Interlocking problems (132).
7. Information and communication technology (ICT) problems (69).
8. Problems with rolling stock's electrical power (319).
9. Problems with power (not related to rolling stocks) (133).
10. Physical problems at stations (6).

As this historical data shows, the two predominant types of technical incident are switch and TC failures. In this thesis, two disruption-length models for incidents caused by these two types of failure are constructed. Models for other technical incidents can be constructed in a similar fashion, by following the model construction procedure we present here.

Other events can also disrupt train traffic. For instance, suicides and train collisions are not caused by the failure of any technical components but they still hinder traffic. Incidents of this type are said to be *non-technical* in nature.

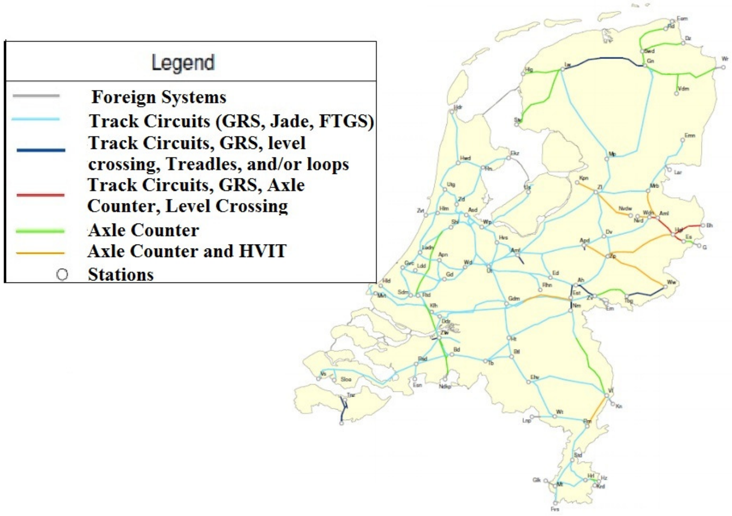
Non-technical incidents caused by suicide are observed particularly frequently in the Netherlands, with 425 recorded in our database for the years $2011-2012$. After such an incident, the site is declared a crime scene and the police first conduct an investigation. Once they have released the site, all technical components are checked by a repair team and the scene is cleared of any human remains before the section is reopened for operation.

A disruption-length model for incidents caused by suicide has not been constructed because ProRail experts believe that the disruption length in such cases is more or less certain: 120 minutes[1].

1.1.1.1 The Track Circuit (TC)

The TC is part of the train safety system, its task being to detect whether a block or section of track is occupied or not. It is the most commonly used train detection system in the Dutch railway network, as shown in Figure 1.2.

Figure 1.2: *Map of the train detection systems installed across the Dutch railway network. Source: ProRail*


Legend
Foreign Systems
Track Circuits (GRS, Jade, FTGS)
Track Circuits, GRS, level crossing, Treadles, and/or loops
Track Circuits, GRS, Axle Counter, Level Crossing
Axle Counter
Axle Counter and HVIT
Stations

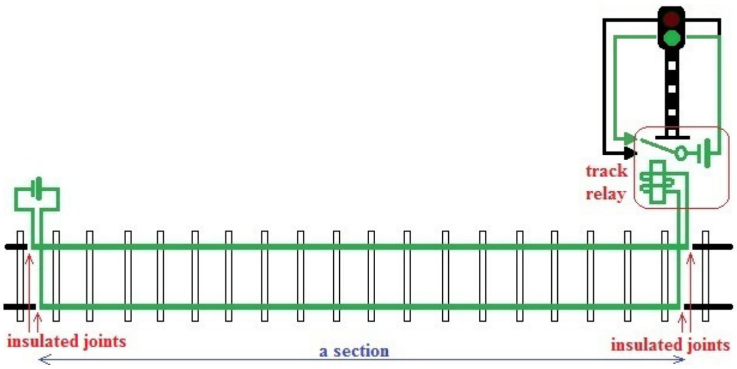
A TC has an electric current source at one end of the section and a detection device at the other end. The current flows along the section, through the rails. Sections are separated by joint insulators which keep the current flowing within their corresponding section. A track relay acts as the detection device. The track relay is in an up position when the section is clear and drops when the section is occupied. Specifically, the axles of the train produce a short circuit between the two rails so that the relay does not receive any current and hence the section is detected as occupied (Pachl [2004]). This is illustrated in Figure 1.3.

A TC system consists of many different electrical components. Failure of any of these leads to the corresponding section being erroneously detected as occupied and so disrupts train traffic. Sometimes, however, an external cause

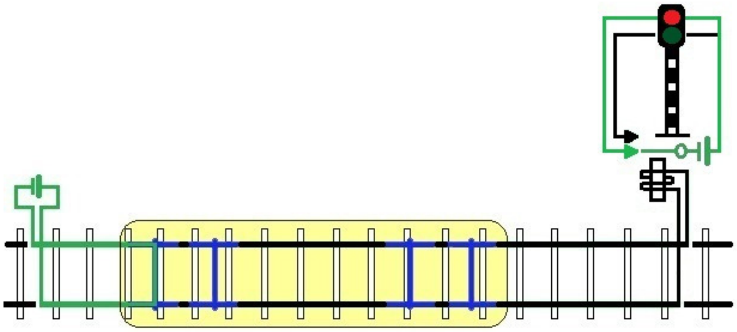
---

[1] Our database (to be described shortly) cannot be used to confirm this belief because it is based solely on the work of repair teams. This means that it does not contain full information about disruptions caused by suicide.

Figure 1.3: *How the track circuit system works.*



(a) An unoccupied block or section.



(b) An occupied block or section.

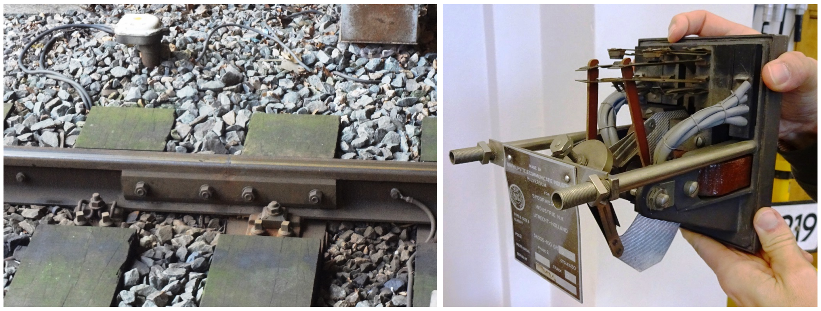
also leads to the same problem. The following are the main components of a TC and some external factors which can cause problems (Visser and Steenkamp [1981]).

1. **Joint insulator**. A joint insulator, depicted in Figure 1.4(a), separates two consecutive sections of railway with an insulator made of nylon plates and linings, so that the 75-hertz detection alternating current in one section does not flow to the adjoining one. If the joint insulator fails, the detection current in one section can flow into the next. Two common causes of this situation are:

   (a) Coins. A frequent problem with the insulation function is when someone deliberately puts conductive material on the joint insulator, most usually coins, thus allowing electric current from one section to flow into the next. To solve the problem, the repair team only needs to remove the conductive material from the joint insulator.

(b) Splinters/grinding chips and insulator problem. Other conductive foreign objects may also fall accidentally onto the joint insulator. Most frequently, the material that cause this is a splinter, a tiny piece of metal from the magnetic track brake or a train, or a grinding chip deposited from the overhead line or from interaction between the wheels and the rails. To solve this problem, the repair team needs to clear the splinter or chip from the joint insulator. It may also need to replace the nylon plates and linings to renew the insulating function of the insulator.

Figure 1.4: *Two components of the TC.*



(a) An insulated joint at Utrecht Centraal Station.

(b) A B2 Vane Relay.

2. **Relay cabinet**. A relay cabinet contains a few electrical components, such as a B2 vane relay (Figure 1.4(b)), a CR/VTB relay, a capacitor, a transformer, a fuse and wiring. Failure of any of these requires the repair team to replace the defective component with a new one.

3. **Arrester**. An arrester protects the TC from high voltage and creates a safe current path should the catenary[2] be disrupted and fall onto the track. A faulty arrester is replaced with a new one.

4. **Cable**. Cable problems arise if the cable itself is broken, its insulation is damaged or it is stolen for its valuable copper. Whatever the case, the faulty cable has to be replaced.

5. **Track-side electrical junction box**. The trackside electrical junction box connects the heavier trackside cable from the track circuit with the long, thinner cable to the relay cabinet. If it fails, it has to be replaced by a new box.

6. **Impedance bond**. In an electrically powered railway system, the rail is also used to carry the return traction direct current (DC) of 1500 volts. The

---

[2] An overhead wire used to transmit electrical power to trains.

impedance bond permits this DC traction current to pass between sections while blocking the track circuits alternating current (AC) of 75 hertz to stay within its respective section. When there is problem with an impedance bond, the repair team needs to replace it with a new one.

7. **External Reasons**. Some external factors which quite commonly cause TC problems.
    (a) *Human Error*. Sometimes, a worker makes an unintended mistake that leads to an erroneous section occupation detection. For example, they accidentally change the setting of a component during maintenance work. There is no one specific action needed to solve such a problem: that depends on what mistake has been made.
    (b) *Adjustment problem*. Sometimes, none of the components fails but the TC setting is not correct – most of the time due to high ambient temperature or too low ballast resistance. In this case, the engineering staff need to correct the setting by readjusting it. Hot days are therefore likely to produce more TC problems than average. On the afternoon of 28 June 2011, for instance, the temperature in the Netherlands exceeded 30^{o}C. As a result, on that day 19 urgent TC problems were detected on the Dutch railway network – well above the daily average of 2.32. Some remarks in the database indicated that high temperature played a role in these incidents.
    (c) *Short Circuit*. This problem occurs when an object causes a short circuit between the rail and the soil. The type of action needed depends on the degree of damage caused.

1.1.1.2 The Switch

A switch, set of points or a turnout[3], is an assembly of rails, movable points, and a frog, which creates the tangential branching of tracks and allows trains to switch from one track to another (Pachl [2004]). A mechanism to move the points from one position to another is also present by the switch. In the past the mechanism was in the form of a lever, which needed to be moved manually by a human operator. Nowadays, it is a remotely controlled electric motor.

The moving mechanism of a switch is divided into three different processes: steering, switching and controlling. Figure 1.6 represents the flow of these processes, starting from the top left of the diagram. When a traffic controller chooses a switch position (left or right), this command is forwarded to the point motor. This is the *steering* process. The motor then commands the point machine to move the switch blade to the desired position. This is the *switching* process. Once the blade is correctly adjusted, the point machine receives this information and relays it back to the traffic controller, who is notified that the switch is now in the desired position. This is the *controlling* process.

---

[3] The term “turnout” is commonly used in civil engineering when referring to this mechanism (Pachl [2004]). In this thesis, however, the term “switch” is used.

Figure 1.5: *Switch number 279B and 283B at Rotterdam Centraal Station.*

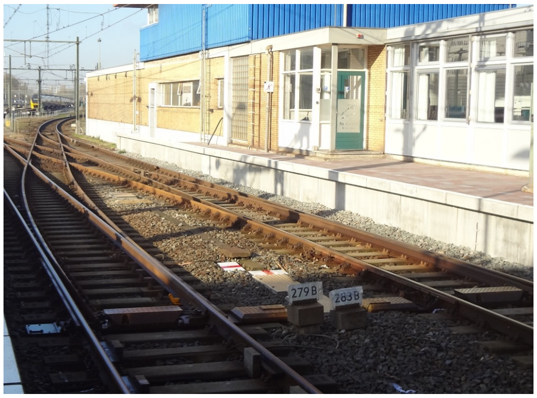
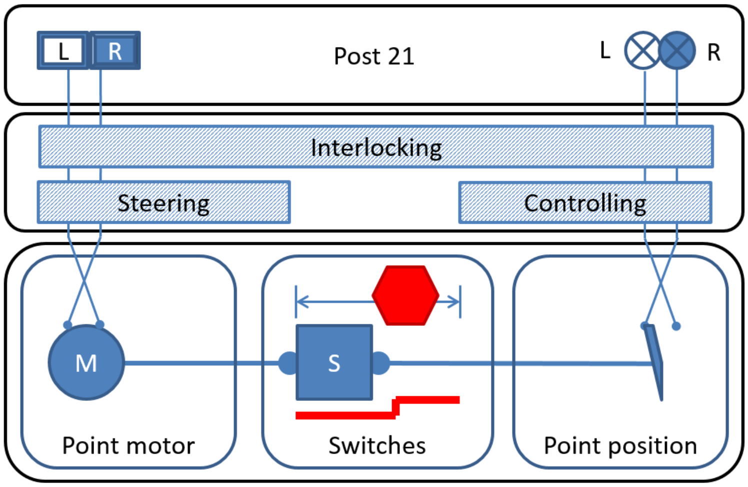


Figure 1.6: *The moving mechanism process of a switch. Source: ProRail.*



The most common problem with switches, which often leads to disrupted train traffic, is failure of their moving mechanisms caused by failure of any of the underlying processes. At ProRail, a switch is said to be “not in control” (NIC) when this situation occurs. Other, less frequent, switch-related problems include fractures of the rails, movable points or frogs and displacement of the substructure. ProRail distinguishes eight different types of NIC situation, as follows (with their ProRail’s codes).

1. **Train safety (steering)** (NIC-TB-S). A steering problem with the train safety system takes the form of an erroneous switch occupation detection where, when it happens, the switch must be adjusted to the correct position. The system also contains several small electrical components, such as fuses, relays and wiring, which can break over time. If they do, they need to be

replaced.

2. **Steering circuit** (NIC-1). A failure of the steering circuit can be caused by loose wiring or polluted devices. In this case the wiring needs to be fastened or the devices cleaned. As well as the electrical components listed above, the system also includes small parts such as sensors, hand cranks and motors, which need to be replaced when they are broken.

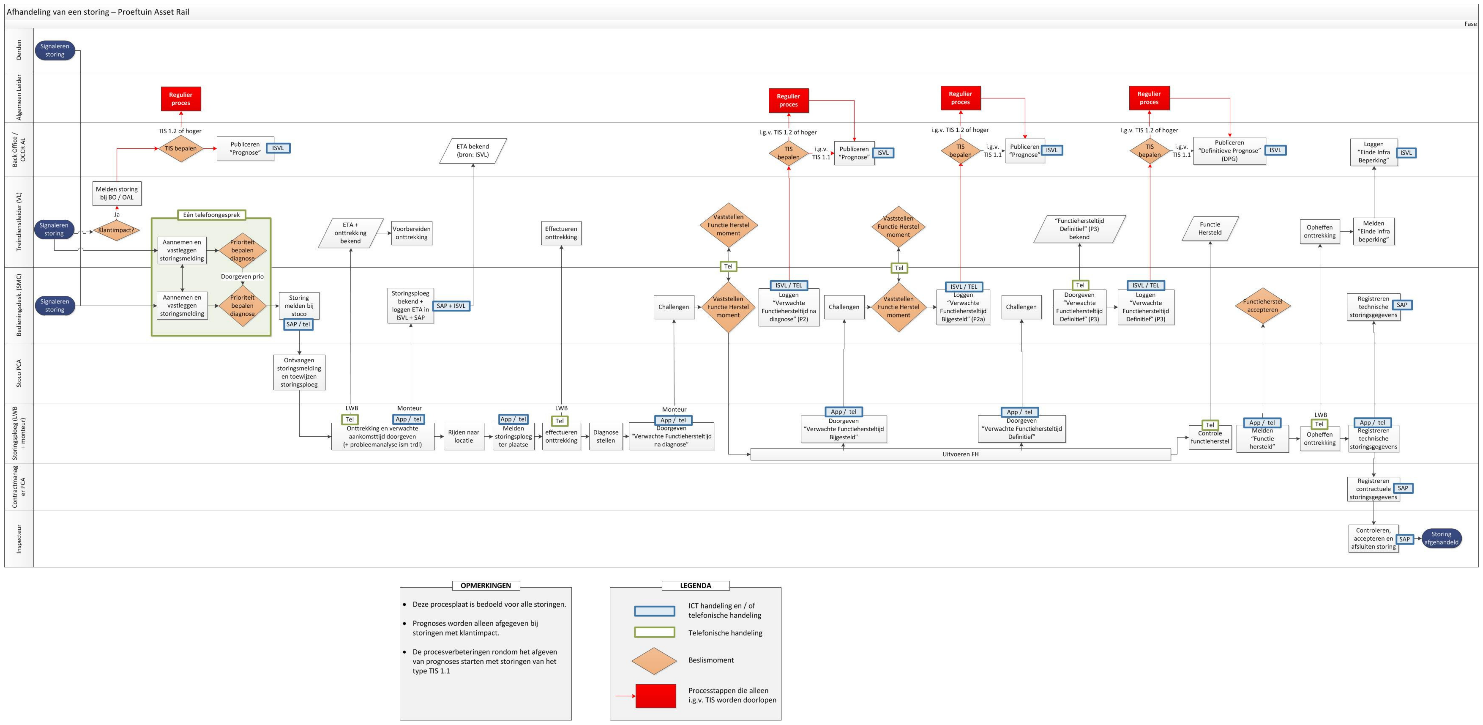
3. **Point machine (steering)** (NIC-2). This problem occurs during the command transmission from the point motor to the switch blade. It is usually caused by friction wear of the point machine's gears. In this case the entire point machine must be replaced.

4. **Blockage** (NIC-3A and NIC-3B). The blade's movement can be hindered by a foreign object blocking the switch. The object can be external, e.g. an ice block or a stone (NIC-3A), or internal, e.g. a bolt or a screw (NIC-3B). When this occurs, the blockage needs to be removed from the switch. Once that has been done, the repair team needs to check the switch to make sure it is now working properly.

5. **Point machine (controlling)** (NIC-4). This problem occurs during detection of the moved blade. It can be caused by an expired control rod in the point machine, which needs to be replaced.

6. **Circuit control** (NIC-5). A defective circuit controller causes the switch's status to not be updated, even though it is already in the desired position. In this case the circuit controller needs to be replaced.

7. **Train safety (controlling)** (NIC-TB-C). A control problem with the train safety system occurs during confirmation of the moved switch position. This is usually caused by a defective component in the circuit control of the train safety system. In this case the defective component needs to be replaced.


Afhandeling van een storing – Proeftuin Asset Rail
Fase
Derden
Algemeen Leider
Back Office / OCCR AL
Treindienstleider (VL)
Bedieningsdesk. (SMC)
Stoco PCA
Storingsploeg (LWB + monteur)
Contractmanager PCA
Inspecteur
Signaleren storing
Signaleren storing
Signaleren storing
Klantimpact?
Ja
Melden storing bij BO / OAL
TIS bepalen
TIS 1.2 of hoger
Regulier proces
Publiceren "Prognose"
ISVL
Eén telefoongesprek
Aannemen en vastleggen storingsmelding
Prioriteit bepalen diagnose
Doorgeven prio
Aannemen en vastleggen storingsmelding
Prioriteit bepalen diagnose
Storing melden bij stoco
SAP / tel
Ontvangen storingsmelding en toewijzen storingsploeg
LWB
Tel
Monteur
App / tel
Onttrekking en verwachte aankomsttijd doorgeven (+ probleemanalyse ism trdl)
ETA + onttrekking bekend
Voorbereiden onttrekking
Storingsploeg bekend + loggen ETA in ISVL + SAP
SAP + ISVL
ETA bekend (bron: ISVL)
Rijden naar locatie
App / tel
Melden storingsploeg ter plaatse
LWB
Tel
effectueren onttrekking
Effectueren onttrekking
Diagnose stellen
Monteur
App / tel
Doorgeven "Verwachte Functiehersteltijd na diagnose"
Challengen
Vaststellen Functie Herstel moment
Tel
Vaststellen Functie Herstel moment
ISVL / TEL
Loggen "Verwachte Functiehersteltijd na diagnose" (P2)
i.g.v. TIS 1.2 of hoger
Regulier proces
TIS bepalen
i.g.v. TIS 1.1
Publiceren "Prognose"
ISVL
Uitvoeren FH
App / tel
Doorgeven "Verwachte Functiehersteltijd Bijgesteld"
Challengen
Vaststellen Functie Herstel moment
Tel
Vaststellen Functie Herstel moment
ISVL / TEL
Loggen "Verwachte Functiehersteltijd Bijgesteld" (P2a)
i.g.v. TIS 1.2 of hoger
Regulier proces
TIS bepalen
i.g.v. TIS 1.1
Publiceren "Prognose"
ISVL
App / tel
Doorgeven "Verwachte Functiehersteltijd Definitief"
Challengen
Tel
Doorgeven "Verwachte Functiehersteltijd Definitief" (P3)
"Functiehersteltijd Definitief" (P3) bekend
ISVL / TEL
Loggen "Verwachte Functiehersteltijd Definitief" (P3)
i.g.v. TIS 1.2 of hoger
Regulier proces
TIS bepalen
i.g.v. TIS 1.1
Publiceren "Definitieve Prognose" (DPG)
ISVL
Tel
Controle functieherstel
Functie Hersteld
App / tel
Melden "Functie hersteld"
Functieherstel accepteren
LWB
Tel
Opheffen onttrekking
Opheffen onttrekking
Melden "Einde infra beperking"
Loggen "Einde Infra Beperking"
ISVL
App / tel
Registreren technische storingsgegevens
Registreren technische storingsgegevens
SAP
Registreren contractuele storingsgegevens
SAP
Controleren, accepteren en afsluiten storing
SAP
Storing afgehandeld
LEGENDA
ICT handeling en / of telefonische handeling
Telefonische handeling
Beslismoment
Processtappen die alleen i.g.v. TIS worden doorlopen

OPMERKINGEN

- Deze procesplaat is bedoeld voor alle storingen.
- Prognoses worden alleen afgegeven bij storingen met klantimpact.
- De procesverbeteringen rondom het afgeven van prognoses starten met storingen van het type TIS 1.1

Figure 1.7: *The disruption response process (in Dutch). Source: ProRail.*

1.1.2 Disruption Response Process

When a disruption occurs on the Dutch railway network, the disruption response process involves three main actors.

1. The **Train Traffic Controller** (***Treindienstleider (VL)***). When a disruption occurs, it is their responsibility to provide a safe working environment for the repair team by keeping the line clear during the repair.
2. The **Disruption Registration Center** (***Storingsmeldcentrum (SMC)***). Its task is to register the disruption and to keep relevant parties informed of progress in dealing with it.
3. The **contractors** (***aannemers***). In the Netherlands, several companies are contracted to perform repair work following incidents on different parts of the railway network. The four biggest are Strukton, ASSET Rail, BAM and Volker Rail.

Two other parties are also involved in the process, but somewhat less directly than the three mentioned above.

1. The **Back Office**. This receives information about the disruption and its predicted length from the SMC. Its task is to build a scenario on how to manage traffic during the period of disruption. For example, by rerouting, diverting[4] or cancelling trains. It then communicates this scenario to the train traffic controller, to assist them in managing train operations.
2. The **General Leader** (***De Algemeen Leider***). If the disruption is considered as having a major impact, the Back Office asks the *algemeen leider* to attend in person to supervise the ongoing situation on site, e.g. overseeing repairs or assisting passengers.

Representatives of all these parties are stationed at the Operational Control Centre Rail (OCCR) in Utrecht, where they work together to manage traffic during disruptions when the regular timetable is no longer being followed. Their goal is is to return traffic to normal as soon as possible.

Figure 1.7 presents a flowchart of the disruption response process for technical disruptions in the Netherlands. When a disruption occurs, the VL and SMC receive information about it. When it is serious, i.e. it impacts clients (the train operators, and hence passengers), a rough prediction of its length is made. This is called the “P1” prediction and is based on the average length of the same disruptions in the past.

In the meantime, the repair team from the relevant contractor is informed by the SMC about the disruption and is tasked to carry out repairs. From now on the OCCR is in close communication with the repair team, which informs the VL about its estimated time of arrival at the disruption site (the latency time). Once there, the team has 15 minutes to diagnose the problem, after which it

---

[4] When a train is *diverted*, it is on time but not running on its original route. When a train is *rerouted*, it is not on time and not running on its original route.

is required to make a prediction, using its own judgement, regarding the time needed for the repair. This is called the “P2” prediction. It is received by the SMC and is forwarded to the Back Office. The repair team is allowed to update its prediction later, in which case it is referred to as the “P2a” prediction. Each time an update is made, the information is forwarded to the Back Office.

Once a final prediction can be made, the repair team is required to update the OCCR again. This is the so-called “P3” prediction. Upon completion of its work, the repair team informs the OCCR that the problem is solved and the disrupted train traffic can be resumed. It is then required to register information about the disruption on an administrative form to be stored in a SAP database. Unfortunately, this procedure does not include recording the repair workers’ own predictions.

In current practice, therefore, the uncertainty surrounding the length of a disruption is dealt with by means of a series of predictions based on the repair workers own expertise and judgement. The goal of this thesis is to construct a prediction model which assists the OCCR by updating the disruption-length forecast every time new information about the situation is available, in a manner similar to the way the “P1”, “P2”, and “P3” predictions help them now. This model is based on historical data, the source of which is described in the next section.

1.1.3 The SAP Database

The data used in this thesis comes from the SAP database. This is an Excel-based database in which each column represents a specific piece of information and each row represents a recorded incident. All disruptions that involve the contractors are recorded in this database.

The following are the information items from the SAP database which are of importance in this thesis.

1. **System**. This column specifies the type of disruption.
2. **Contractor** and **Trace**. These columns specify the responsible contractor and its subdivision for each disruption.
3. **Priority**. This column specifies the priority of the incident, using integer values from 1 (major disaster) to 9 (least urgent).
4. **Contractor informed time**. This column shows the time the contractor was informed about the disruption.
5. **Repair team arrival time**. This column indicates the time the repair team arrived at the disruption site. The interval between the contractor informed time and the repair team arrival time is the latency time.
6. **Function recovery time**. This column indicates the time the failure was repaired and train traffic over the blocked section could be resumed. The interval between the repair team arrival time and the function recovery time is the repair time.

7. **Operational points from/to**.In the database, the disruption site is recorded as between two ”operational points”[5] at which GPS information is available. The coordinates of the disruption site are estimated to be the average of the two operational points' coordinates.

8. **Contract type**. In the Netherlands, there are currently two types of contract between ProRail and its contractors. The older “output-based contract” (*Output-procescontracten*, OPC) is based on the amount of work the contractor performs, whilst the newer “performance-based maintenance contract” (*Prestatiegericht Onderhoud*, PGO) introduces a penalty if the work takes too long. The contract type is indicated in this column.

9. **Remark**. This column contains free-text information about the incident, provided by the repair team when filling in the SAP form.

The training set used to construct our model is incident data collected between 1 January 2011 and 30 June 2013 . To validate the model, a second set of data has been used. The test set for disruptions more likely to occur in the summer contains incidents between 1 May 2014 and 31 October 2014, while that for disruptions more likely in the winter contains incidents between 1 October 2014 and 31 March 2015.

The SAP database also contains information about the cause of the incidents. Unfortunately, however, this is not recorded properly and is often presented as unstructured text that is hard to interpret. For instance, the cause of 67% of the urgent incidents caused by TC failures, one of the incident types that will be of interest later on in this thesis, is recorded either as “unknown” or “other”. This is a huge setback for our research, because this information is a crucial factor influencing repair time.

Fortunately to overcome this problem, information can be extracted from the non-standardized “Remark” column to some degree. By manually reading and looking for the right key words in each incident's entry, a diagnosis of the problem can sometimes be made. In performing this labour-intensive work, we were supervised by a track circuit expert and a switch expert from ProRail, who assisted us in defining the principal causes of TC problems (see subsection 1.1.1.1) and switch problems (see in Subsection 1.1.1.2) respectively, and in identifying the common key words for them. Of course, because the comments in the “Remark” column are written as free text, there is no standardization in the richness of the information they contain. At one extreme, some incidents are recorded with large amounts of explanatory detail. At the other, none at all is provided. Nevertheless, in this way we were able to reduce the proportion of “unknown” TC incidents to 30%. How this work was undertaken and how we made our decisions concerning that remaining 30% of the data will be described in more detail later in this thesis.

Material from other databases has been also used to complement that available in the SAP system. Information about the location of level crossings is

---

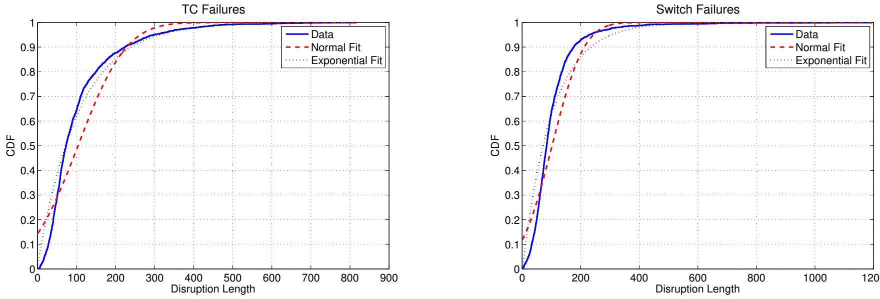
[5] The Dutch railway network is marked with operational points across the system: railway stations, junctions, bridges, etc.

used to approximate the distance between the estimated disruption site and the nearest level crossing. Information about the hourly frequency of trains passing each operational point is used to determine the density of traffic at disruption sites. Since weather is one factor of interest in the thesis, we have also consulted the hour-by-hour national weather data published by the Royal Netherlands Meteorological Institute (*Koninklijk Nederlands Meteorologisch Instituut*, KNMI) and available at: http://www.knmi.nl/nederland-nu/klimatologie/uurgegevens.

1.2 Modelling Disruption Length

A vast number of different mathematical algorithms and models proposed for recovery from a disrupted situation are available in literature. Cacchiani et al. [2014] provide an overview in which some of the cited works mention the uncertainty of disruption length. Given information about disruption length, the algorithms seek an optimal solution to recover from the disrupted situation in the form of timetable, rolling stock or crew rescheduling. This means that information about disruption length is crucial input for any of these algorithms and models to work.

However, disruption length is very uncertain: it is difficult to tell exactly how long a disruption will last. To tackle this, disruption length is going to be represented here as a probability distribution. This allows us to generate random samples of disruption length. This approach is relatively new in railway operation, but has been used in several earlier studies on highway traffic engineering. For instance, Golob et al. [1986], Giuliano [1989] and Sullivan [1997] use the log-normal distribution and Nam and Mannering [2000] use the Weibull distribution. In railway operation, Meng and Zhou [2011] model disruption length on a single-track rail line in China with the normal distribution, while Schranil and Weidmann [2013] model railway disruption length in Switzerland with the exponential distribution.



(a) Distribution of TC Disruption Length.

(b) Distribution of Switch Disruption Length.

Figure 1.8: *Observed distributions of disruption length (in minutes) in the data.*

Figure 1.8 presents the distributions of disruption for TC failures (Figure

1.8(a)) and switch failures (Figure 1.8(b)) in the Netherlands[6]. The observed distributions in the data are represented by the blue lines in both figures. The normal and exponential distribution are fitted using the standard maximum-likelihood approach and the results are presented as the dashed red curves and the dotted black curves, respectively. Both plots indicate that the distributions do not represent the disruption length we observe in the data.

Moreover, several factors influencing disruption length are considered. Our goal is to construct a joint distribution between the disruption length and these influencing factors for each disruption type, based on historical data. A prediction of disruption length can be made by conditioning this joint distribution on the observed values of the influencing factors, resulting in the conditional distribution of disruption length. Having a conditional distribution as the model output enables the OCCR to choose different quantiles of the distribution as the predictions of disruption length so as to optimize train traffic during disruption, depending on the situation.

1.2.1 Modelling the Joint Distribution and the Copula

It is often difficult to build a joint distribution that fits a set of data. Consider a model that involves only continuous variables. Formally, let $\mathbf{X} = (X_1, \ldots, X_n)$ be a continuous random vector with realization $\mathbf{x} = (x_1, \ldots, x_n)$.

One possible joint distribution model is the multivariate normal distribution. In this case, the joint density, $f_{1,\ldots,n}(x_1, \ldots, x_n)$, is defined as

$$f_{1,\ldots,n}(x_1, \ldots, x_n) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^T \Sigma^{-1} (\mathbf{x}-\mu)\right) \tag{1.1}$$

where μ is the mean vector of $\mathbf{X}$, Σ denotes the covariance matrix, and $|\Sigma|$ denotes the determinant of Σ. Notice that there are constraints in this model. First of all, Σ must be positive-definite. Moreover, this model only considers pairwise interactions between the variables as represented in Σ. Thirdly, each margin X_i as well as each higher dimensional margin must be normally distributed.

Another possible joint model is the multivariate t-distribution (see, e.g. Nadarajah and Kotz [2005]). However, the margins of such a model are constrained to be t-distributed.

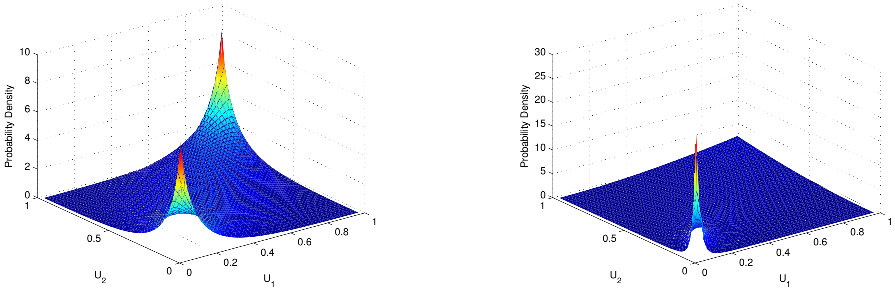
In practice, it might be the case that the marginal distributions of X_is are of different types. A popular way to solve this problem is by applying the *copula*. A copula is the n-dimensional joint distribution in the unit hypercube of n uniform random variables U_i. The theorem of Sklar [1959] serves as the basis of copula application. It states that any cumulative distribution function of $(X_1, \ldots, X_n)$, denoted as $F_{1,\ldots,n}$, can be rewritten in terms of the corresponding copula C as:

$$F_{1,\ldots,n}(x_1, \ldots, x_n) = C(F_1(x_1), \ldots, F_n(x_n)) \tag{1.2}$$

where $F_i(x_i)$ denotes the marginal distribution of the i-th variable. Consequently,

---

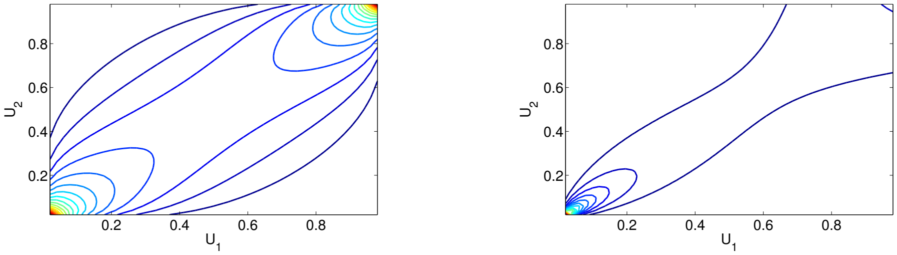
[6] In this thesis, disruption length is presented in minutes.



(a) Density of a normal copula with $\rho = 0.7167$ ($r = 0.7$).

(b) Density of a Clayton copula with $\theta = 2.1316$ ($r = 0.7$).

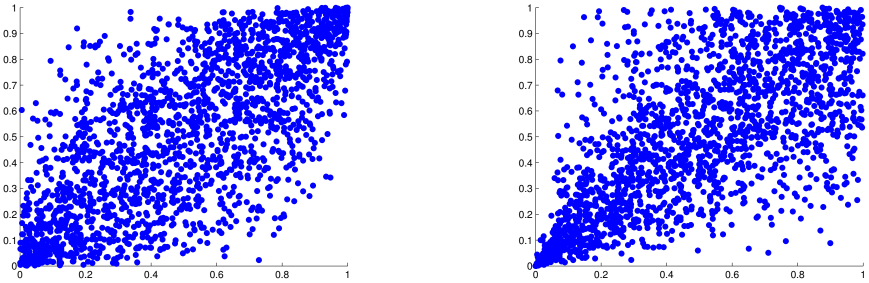
Figure 1.9: *Densities of two bivariate copulas.*



(a) Density contour plot of the normal copula in Figure 1.9(a).

(b) Density contour plot of the Clayton copula in Figure 1.9(b).

Figure 1.10: *Density contour plots of two bivariate copulas.*



(a) Scatter plot of data sampled from a normal copula in Figure 1.9(a).

(b) Scatter plot of data sampled from a Clayton copula in Figure 1.9(b).

Figure 1.11: *Scatter plots of data from the two bivariate copulas.*

the copula representation of the joint density $f_{1,\ldots,n}(x_1,\ldots,x_n)$ is:

$$f_{1,\ldots,n}(x_1,\ldots,x_n) = c(F_1(x_1),\ldots,F_n(x_n)) \cdot f_1(x_1) \cdot \ldots \cdot f_n(x_n) \tag{1.3}$$

where c is the copula density. Moreover, the copula satisfying equation (1.2) is unique if the variables are continuous. However, if at least one variable is discrete, the copula is not unique.

In equation (1.3), note that the marginal densities f_i are separated from the copula density c. Therefore, in this case the joint density $f_{1,\ldots,n}$ can be modelled by modelling the copula c without any constraints on the marginal distributions.

When $n = 2$, there are many copula families that are available and easy to use (see, e.g., Nelsen [2006] and Joe [2014]). We next present two of those often used in practice.

Example 1.2.1. ***(bivariate normal copula)***. The bivariate normal, or Gaussian, copula is defined as follows:

$$C_\rho(u_1,u_2) = \Phi_\rho(\Phi^{-1}(u_1),\Phi^{-1}(u_2)) \tag{1.4}$$

where Φ^{-1} denotes the inverse cumulative distribution of a univariate standard normal distribution and Φ_ρ denotes the joint cumulative distribution of a bivariate normal distribution with zero mean and correlation ρ. The copula density can be derived from equations (1.3) and (1.1) for $n = 2$, which result in:

$$c_\rho(u_1,u_2) = \frac{1}{\sqrt{1-\rho^2}} \exp\left(-\frac{\rho^2\Phi^{-1}(u_1)^2 - 2\rho\Phi^{-1}(u_1)\Phi^{-1}(u_2) + \rho^2\Phi^{-1}(u_2)^2}{2(1-\rho^2)}\right). \tag{1.5}$$

Example 1.2.2. ***(bivariate Archimedean copula)***. The bivariate Archimedean copula has the following representation:

$$C_\theta(u_1,u_2) = \psi_\theta^{-1}(\psi_\theta(u_1) + \psi_\theta(u_2)) \tag{1.6}$$

where ψ_θ denotes the *generator* function that is continuous, strictly decreasing, convex, and satisfies $\psi_\theta(1) = 0$.

When the generator function is $\psi_\theta(u) = \frac{1}{\theta}(u^{-\theta} - 1)$ where $\theta \in [0,\infty)$, (1.6) becomes:

$$C_\theta(u_1,u_2) = \left(u_1^{-\theta} + u_2^{-\theta} - 1\right)^{-1/\theta}. \tag{1.7}$$

This copula is known as the **Clayton copula** (Clayton [1978]) and is interesting for some applications because it captures the lower tail dependence between the variables.

Figure 1.9 presents the density of a bivariate normal copula (Figure 1.9(a)) and a Clayton copula (Figure 1.9(b)) when the Spearman's rank correlation between U_1 and U_2 is $r = 0.7$[7]. Figure 1.10 presents the contour plots of the two copula densities. It can be seen that the two copulas behave very differently, even when the rank correlations are the same.

---

[7] This corresponds to $\rho = 0.7167$ for the normal copula and $\theta = 2.1316$ for the Clayton copula.

Consequently, the samples of (U_1, U_2) generated from the two copulas would also look very different. Figure 1.11 presents the scatter plots of 2000 samples from the two copulas depicted in Figure 1.9.

When $n \geq 3$, a few copula families are available.

Example 1.2.3. ***(multivariate normal copula)***. The multivariate normal copula is an extension of the bivariate normal copula and is defined as follows:

$$C_R(u_1, \ldots, u_n) = \Phi_R(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_n)) \tag{1.8}$$

where Φ^{-1} is as in the bivariate case and Φ_R denotes the joint cumulative distribution of a multivariate normal distribution with zero mean and correlation matrix R, which is positive-definite. Moreover, all k-margins of the copula, with $k < n$, are constrained to be k-variate normal copula as well.

Example 1.2.4. ***(multivariate Archimedean copula)***. The bivariate Archimedean copula can be extended to the multivariate case as follows:

$$C_\theta(u_1, \ldots, u_n) = \psi_\theta^{-1}(\psi_\theta(u_1) + \ldots + \psi_\theta(u_n)). \tag{1.9}$$

McNeil and Nešlehová [2009] show that ψ_θ^{-1} needs to be n-monotone[8], i.e. differentiable up to the order $n-2$, to satisfy:

$$\frac{d^k}{dt^k}(-1)^k \psi_\theta^{-1}(t) \geq 0, \quad k = 0, 1, \ldots, n-2$$

and $\frac{d^{n-2}}{dt^{n-2}}(-1)^{n-2}\psi_\theta^{-1}(t)$ is non-negative, non-increasing and convex on $(0, \infty)$. If ψ_θ^{-1} is n-monotone for any dimension n, the generator is said to be *completely monotone* (Kimberling [1974]).

Note that all k-margins of a multivariate Archimedean copula are identical, i.e. $C(u_1, \ldots, u_k, \mathbf{1}) = \psi_\theta^{-1}(\sum_{i=1}^k \psi_\theta(u_i))$ with the same dependence parameter θ. This constraint makes this copula less attractive from an application point of view.

Example 1.2.5. ***(nested Archimedean copula)***. Joe [1997] presents the construction of an n-dimensional copula C_n by nesting $n-1$ bivariate Archimedean copulas. This copula is defined iteratively for $n > 2$ by:

$$\begin{aligned} C_n(u_1, \ldots, u_n; \psi_1, \ldots, \psi_{n-1}) &= \psi_1^{-1}(\psi_1(u_1) + \psi_1(C_{n-1}(u_2, \ldots, u_n; \psi_2, \ldots, \psi_{n-1})) \\ &= \psi_1^{-1}(\psi_1(u_1) + \psi_1 \circ \psi_2^{-1}(\psi_2(u_2) + \psi_2(C_{n-2}(u_3, \ldots, u_n; \psi_3, \ldots, \psi_{n-1}))) \\ &\vdots \\ &= \psi_1^{-1}(\psi_1(u_1) + \psi_1 \circ \psi_2^{-1}(\psi_2(u_2) + \ldots \psi_{n-2} \circ \psi_{n-1}^{-1}(\psi_{n-1}(u_{n-1}) + \psi_{n-1}(u_n))\ldots)). \end{aligned} \tag{1.10}$$

Note that $\psi_{n-1}^{-1}(\psi_{n-1}(u_{n-1}) + \psi_{n-1}(u_n))$ corresponds to (1.6).

---

[8] McNeil and Nešlehová [2009] use the term "d-monotone" which corresponds to the d-dimensional joint distribution. In this thesis, we adjust the term to "n-monotone" because our joint distribution is n-dimensional.

With this approach, asymmetries in the joint distribution can be modelled by choosing a different bivariate Archimedean copula for each ψ_k. However, McNeil [2008] shows that $\psi_{k-1} \circ \psi_k^{-1}$ has to be completely monotonic for $k = 2, \ldots, n-1$ which means that ψ_k cannot be chosen independently.

Therefore, while a copula helps in separating the marginal densities f_i from the dependence structure (represented by the copula density c), the construction of c when $n \geq 3$ is still difficult. Depending on the chosen technique, marginal constraint or functional constraint needs to be satisfied in the model construction. In the next subsection, we present a modelling strategy which avoids these problems as it involves a set of algebraically independent bivariate copulas. This strategy uses a graphical structure called "vines" and is shown to be very useful in the model construction when $n \geq 3$.

1.2.2 Vines

Note that with the standard decomposition and the condition of positive (conditional) densities, the joint density can be rewritten as:

$$f_{1,\ldots,n}(x_1, \ldots, x_n) = f_1(x_1) \cdot f_{2|1}(x_2|x_1) \cdot \ldots \cdot f_{n|1,\ldots,n-1}(x_n|x_1, \ldots, x_{n-1}) \tag{1.11}$$

where $f_{n|1\ldots,n-1}$ denotes the conditional density of X_n given $X_1, \ldots, X_{n-1}$. There are many different possibilities to decompose the joint density.

Let $\mathbf{V}$ denote the conditioning set of the j-th term on the right-hand side of (1.11) and $\mathbf{V}_{\setminus i}$ denotes $\mathbf{V}$ without X_i. Each term $f_{X_j|\mathbf{V}}$ for all j such that $i < j$ on the right-hand side of (1.11) can be rewritten with a bivariate copula as follows

$$f_{X_j|\mathbf{V}} = \frac{f_{X_j,X_i|\mathbf{V}_{\setminus i}}}{f_{X_i|\mathbf{V}_{\setminus i}}} = c_{ji|\kappa}(F_{X_j|\mathbf{V}_{\setminus i}}, F_{X_i|\mathbf{V}_{\setminus i}}; \mathbf{V}_{\setminus i}) f_{X_j|\mathbf{V}_{\setminus i}} \tag{1.12}$$

where κ denotes the index of all variables in the set $\mathbf{V}_{\setminus i}$ and $C_{ji|\kappa}$ denotes the bivariate copula between X_j and X_i conditioned on the variables in $\mathbf{V}_{\setminus \mathbf{i}}$ with copula density $c_{ji|\kappa}$. The second equality in (1.12) comes from the copula representation of the joint density as in (1.3).

With this approach the full joint density can be modelled with a set of (conditional) bivariate densities that are represented with a set of bivariate copulas. Note that the (conditional) bivariate densities are algebraically independent where the bivariate copulas can be chosen freely and do not depend on each other. This is called the "copula-vine" approach (Kurowicka and Cooke [2006]).

The joint density decomposition can be represented graphically using a structure called "vines", which was introduced in Cooke [1997] and developed further in Bedford and Cooke [2002]. A vine is a nested set of $n-1$ trees consisting of nodes and (undirected) edges where the edges of the i-th tree are the nodes of the $(i+1)$th tree. In this thesis, we consider only a special form of vine; that is, the *regular* vine in which two edges in tree i are joined by an edge in tree $i+1$ only if they share a common node in tree i.

The following example illustrates the copula-vine approach in three dimensions.

Example 1.2.6. Consider three continuous random variables X_1, X_2, X_3 whose joint density is decomposed and represented with three bivariate copulas C_{12}, C_{23} and $C_{13|2}$ as:

$$\begin{aligned} f_{1,2,3}(x_1,x_2,x_3) &= f_3(x_2)f_{3|2}(x_3|x_2)f_{1|2,3}(x_1|x_2,x_3) \\ &= f_{1,3|2}(x_1,x_3|x_2)f_2(x_2) \\ &= c_{13|2}(F_{1|2}(x_1|x_2),F_{3|2}(x_3|x_2);x_2)f_{1|2}(x_1|x_2)f_{3|2}(x_3|x_2)f_2(x_2) \\ &= c_{13|2}(F_{1|2}(x_1|x_2),F_{3|2}(x_3|x_2);x_2)c_{12}(F_1(x_1),F_2(x_2))c_{23}(F_2(x_2),F_3(x_3)) \\ &\quad f_1(x_1)f_2(x_2)f_3(x_3) \end{aligned} \tag{1.13}$$

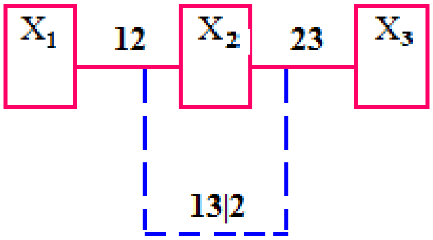


Figure 1.12: *A regular vine on three variables.*

Figure 1.12 illustrates a vine structure that represents the decomposition in (1.13)[9]. The first tree is coloured red. In this tree, the nodes correspond to the variables and the edges correspond to the unconditional bivariate copula C_{12} and C_{23}. The red edges become the nodes of the second tree whose edge, coloured blue, corresponds to the conditional bivariate copula $C_{13|2}$.

To sample the joint distribution of $(X_1, \ldots, X_n)$, we next present a procedure on the vine structure in Example 1.2.6. For more details, interested readers are referred to Kurowicka and Cooke [2006].

In the procedure, we assume the marginal distributions F_i to be continuous and invertible so that one can use them to transform each variable into uniform $(0,1)$. Therefore, without loss of generality X_1, X_2, X_3 are uniform on $(0,1)$.

First, three independent uniform $(0,1)$ variables U_1, U_2, U_3 are sampled. Then:

$$\begin{aligned} x_1 &= u_1, \\ x_2 &= C^{-1}_{2|1:x_1}(u_2), \\ x_3 &= C^{-1}_{3|2:x_2}(C^{-1}_{3|2|C_{1|2:x_2}(x_1)}(u_3)), \end{aligned}$$

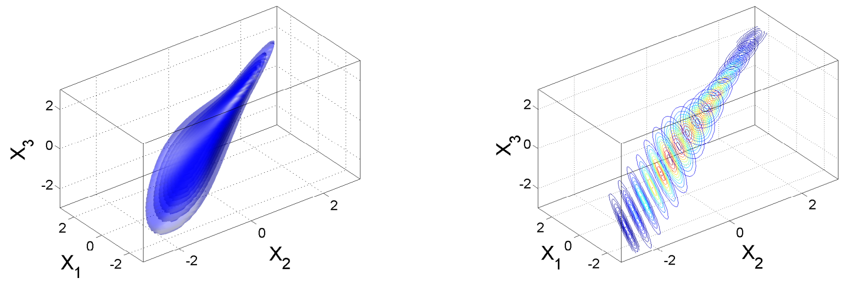
where $C_{j|i:x_i}$ denotes the cumulative distribution function of X_j given $X_i = x_i$ under the copula C_{ij} and $C^{-1}_{j|i:x_i}$ denotes its inverse.

---

[9] To save space, the bivariate copula $C_{ij|k}$ is presented as $ij|k$ in all vine figures in this thesis.

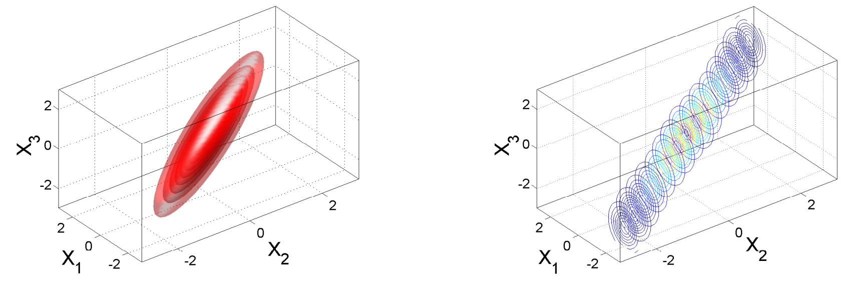
In general the copula $C_{ji|\kappa}$ in (1.12) depends on the conditioning variables $\mathbf{V}_{\backslash i}$. When the copula is assumed not to depend on $\mathbf{V}_{\backslash i}$, the copula is said to be "constant" with respect to the conditioning variables (otherwise it is "non-constant"). The choice of whether the conditional copula is constant or non-constant affects the constructed joint density, as illustrated in the following example.

Example 1.2.7. Consider the three dimensional vine structure in Example 1.2.6. Let X_i be standard normally distributed, C_{12} and C_{23} be the normal copulas with parameters $\rho_{12} = \rho_{23} = 0.8$, and $C_{13|2}$ be normal copula with parameter that depends on X_2 as $\rho_{13|2}(x_2) = 0.9\cos(\pi \cdot x_2)$. The joint density is computed with equation (1.13).



(a) Isosurface of the joint distribution. (b) Contour slice of the joint distribution.

Figure 1.13: *The generated joint distribution of* (X_1, X_2, X_3) *with non-constant* $C_{13|2}$.



(a) Isosurface of the joint distribution. (b) Contour slice of the joint distribution.

Figure 1.14: *The generated joint distribution of* (X_1, X_2, X_3) *with constant* $C_{13|2}$.

Figure 1.13(a) shows several layers of isosurfaces, which represent the points (x_1, x_2, x_3) in the $3D$ space with constant density $f_{1,2,3}(x_1, x_2, x_3)$. Figure 1.13(b) presents the contour of the joint density when sliced at various values of X_2. Note that the slices are all ellipses, but with different eccentricities. Figure

1.13(b) shows how the correlation between $X_1|X_2$ and $X_3|X_2$ changes from strongly positive when X_2 is small to strongly negative when X_2 is large. This is a direct implication of the functionality of $\rho_{13|2}$ with respect to X_2.

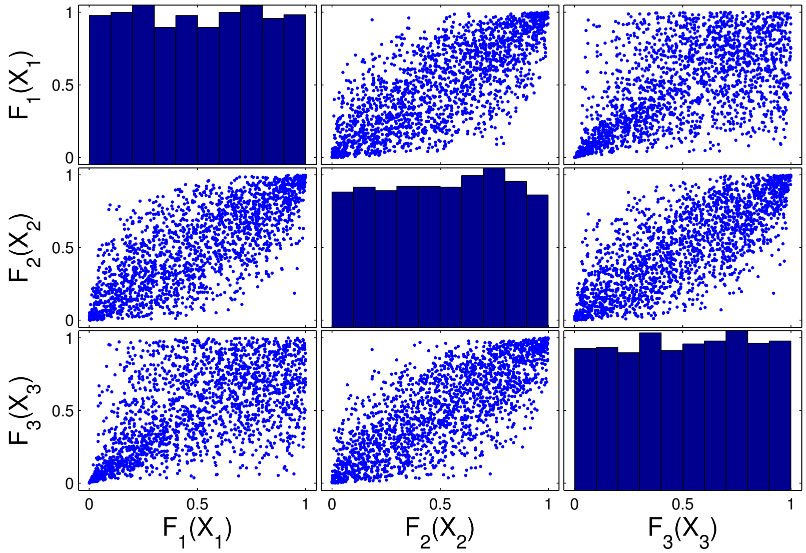
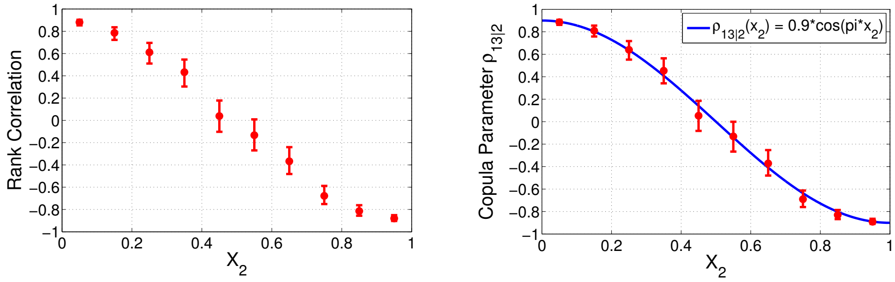


Figure 1.15: *Bivariate copulas of the samples of* (X_1, X_2, X_3) *with non-constant conditional copula.*



(a) Rank correlations.

(b) Copula parameters and the function of $\rho_{13|2}(x_2)$.

Figure 1.16: *Rank correlations and copula parameters between* $X_1|X_2$ *and* $X_3|X_2$ *for the ten different groups of* X_2.

If $C_{13|2}$ is constant, the resulting joint distribution will be different. For instance, consider a second joint distribution where $C_{13|2}$ is normal with constant

$\rho_{13|2} = 0$ (that is the "average" correlation of the first example) and everything else is as before. Spanhel and Kurz [2015] show that this is the closest constant copula in terms of the Kullback-Leibler distance to the non-constant copula in the first example. Again, the joint density is computed with equation (1.13). This construction results in the trivariate normal distribution.

Figure 1.14(a) shows several layers of isosurfaces, which represent the points (x_1, x_2, x_3) in the $3D$ space with constant density $f(x_1, x_2, x_3)$. In this case, the isosurfaces are all elliptical. Figure 1.14(b) presents the contour of the joint density when sliced at various values of X_2 and the constant behaviour of $\rho_{13|2}$ is observed. Note that all slices are ellipses with constant eccentricities.

Therefore, the choice of constant or non-constant conditional copulas in the model construction might lead to very different joint distributions.

Given a set of data, however, it is not easy to distinguish whether the underlying joint distribution is constructed using constant or non-constant conditional copulas. Figure 1.15 presents the scatter plots of the bivariate copulas of the 2000 samples generated from the joint distribution depicted in Figure 1.13. It can be seen that the dependence between (X_1, X_3) is not represented by the bivariate normal copula. This is because this pair is modelled through the conditional copula which is normal but non-constant.

Using the vine structure in Example 1.2.6 to model the samples' joint distribution, one needs to decide whether to model the conditional copula $C_{13|2}$ with constant or non-constant copulas. One way to determine this is by dividing the data into, for instance, ten groups by discretizing X_2 with equal length. For each group, the rank correlation or the copula parameter between $X_1|X_2$ and $X_3|X_2$ is computed along with the confidence bound. By comparing these rank correlations or the copula parameters, it is possible to conclude whether the conditional copula can be modelled as a constant copula. Figure 1.16 shows the results. Kurz [2013] presents several tests which can be used to identify whether the constant conditional copula assumption can be used in a set of data.

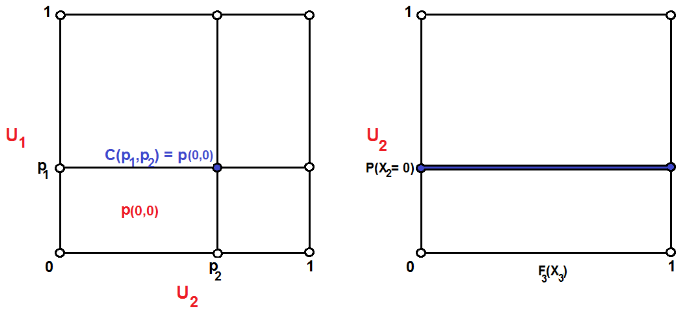
It can be seen that the conditional copula $C_{13|2}$ cannot be modelled as a constant copula. The function $\rho_{13|2}(x_2) = 0.9\cos(\pi * x_2)$ is plotted as the blue line in Figure 1.16(b) and is captured by the confidence bounds of the copula parameters.

Thus far, we have only considered the joint distribution construction of continuous variables. In the next subsection, we briefly discuss model construction when some of the variables are discrete, resulting in a mixed discrete-continuous joint distribution. This is the main topic of this thesis and is discussed in more detail in Chapter 2.

1.2.3 Discrete Variables

Consider three random variables X_1, X_2 and X_3, where X_1 and X_2 are Bernoulli with $\mathbb{P}(X_i = 0) = p_i$ for $i = \{1, 2\}$ and X_3 is continuous with marginal distribution F_3.

Example 1.2.8. Consider the bivariate joint distributions of (X_1, X_2) and (X_2, X_3). To represent them with copulas, we introduce two *latent* variables U_1 and U_2,



(a) Unit square corresponding to (U_1, U_2).

(b) Unit square corresponding to (U_2, F_3).

Figure 1.17: *Unit square corresponding to the latent variable* (U_1, U_2) *and* (U_2, F_3).

which are uniform on $(0,1)$ such that $\mathbb{P}(X_i = 0) = F(U_i \le p_i) = p_i$. In this case the Sklar's equation (1.2) becomes:

$$\mathbb{P}(X_1 \le 0, X_2 \le 0) = F_{U_1,U_2}(p_1, p_2) = C_{12}(p_1, p_2) \tag{1.14}$$

for (X_1, X_2) and:

$$\mathbb{P}(X_2 \le 0, X_3 \le x_3) = F_{U_2,X_3}(p_2, x_3) = C_{23}(p_2, F_3(x_3)) \tag{1.15}$$

for (X_2, X_3). The copulas C_{12} and C_{23} are not unique and are only constrained at that part of the unit square indicated by the blue point and line in Figure 1.17.

Furthermore, the joint distribution of (X_1, X_2, X_3) can also be represented with the copula-vine approach.

Example 1.2.9. Using the vine structure in Figure 1.12 to represent the joint distribution, the joint probability can be rewritten as:

$$\begin{aligned}\mathbb{P}(X_1 = x_1, X_2 = x_2, X_3 = x_3) &= f_3(x_3)\mathbb{P}(X_2 = x_2|X_3 = x_3)\mathbb{P}(X_1 = x_1|X_2 = x_2, X_3 = x_3)\\ &= f_3(x_3)\left(\sum_{s_j=0}^{x_2}(-1)^{s_j}C_{2|3:x_3}(\mathbb{P}(X_2 \le x_2 - s_j))\right)\\ &\quad\left(\sum_{s_j=0}^{x_1}(-1)^{s_j}C_{1|2:x_2|F(X_3\le x_3|X_2=x_2)}(\mathbb{P}(X_1 \le x_1 - s_j|X_2 = x_2))\right)\end{aligned} \tag{1.16}$$

where $C_{j|i:x_i}$ denotes the conditional copula of variable X_j given $X_i = x_i$ which can be computed from the copula C_{ij}. The expressions $F(X_3 \le x_3|X_2 = x_2)$ and $\mathbb{P}(X_1 \le x_1 - s_j|X_2 = x_2)$ can be computed as:

$$F(X_3 \le x_3|X_2 = x_2) = \frac{\sum_{s_j=0}^{x_2}(-1)^{s_j}C_{23}(\mathbb{P}(X_2 \le x_2 - s_j), F_3(x_3))}{\mathbb{P}(X_2 = x_2)}$$

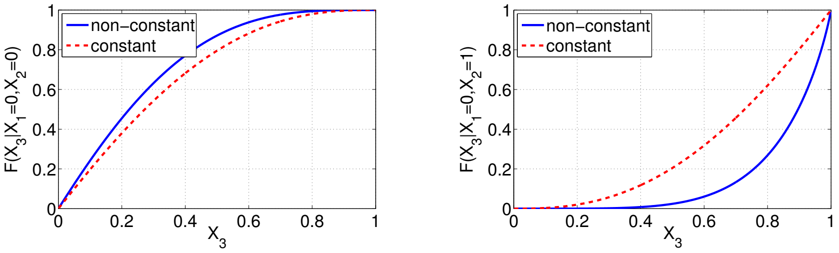
and:

$$\mathbb{P}(X_1 \le x_1 - s_j | X_2 = x_2) = \frac{\sum_{s_i=0}^{x_2} (-1)^{s_i} C_{12}(\mathbb{P}(X_1 \le x_1 - s_j), \mathbb{P}(X_2 \le x_2 - s_i))}{\mathbb{P}(X_2 = x_2)},$$

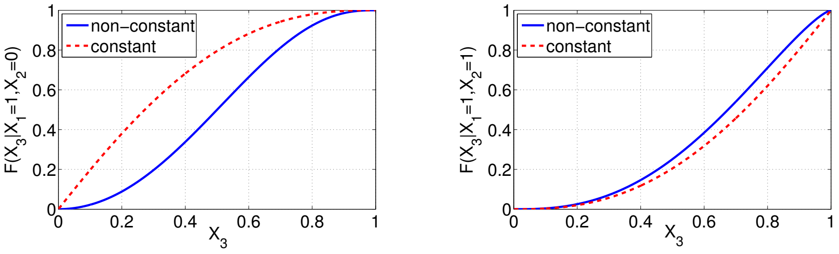
respectively.

Note that the conditional copula $C_{13|2}$ used to compute the third term in (1.16) can be different for $X_2 = 0$ and $X_2 = 1$, which corresponds to the non-constant copula-vine approach in the fully continuous case.

As in the fully continuous case, choosing constant or non-constant conditional copula $C_{13|2}$ affects the constructed joint distribution. The following example illustrates this.



(a) $F(X_3 \le x_3 | X_1 = 0, X_2 = 0)$ (b) $F(X_3 \le x_3 | X_1 = 0, X_2 = 1)$



(c) $F(X_3 \le x_3 | X_1 = 1, X_2 = 0)$ (d) $F(X_3 \le x_3 | X_1 = 1, X_2 = 1)$

Figure 1.18: *The conditional distribution* $F(X_3 \le x_3 | X_1 = x_1, X_2 = x_2)$ *given different realizations of* X_1 *and* X_2.

Example 1.2.10. Consider $\mathbb{P}(X_1 = 0) = \mathbb{P}(X_2 = 0) = 0.5$ and X_3 uniform on $(0,1)$ where the joint distribution of (X_1, X_2, X_3) is constructed with the vine structure in Figure 1.12, i.e. the joint probability is in the form as in Example 1.2.9. Moreover, let C_{12} and C_{23} be the bivariate normal copula with parameters $\rho_{12} = \rho_{23} = 0.8$.

We construct two joint models. The first applies non-constant conditional copula $C_{13|2}$ where the copula is normal with parameters $\rho_{13|2=0} = 0.6$ and $\rho_{13|2=1} = -0.6$ which correspond to the conditioning variable $X_2 = 0$ and $X_2 = 1$, respectively. In the second model, $C_{13|2}$ is taken to be constant where the copula is normal with parameter $\rho_{13|2=0} = \rho_{13|2=1} = 0$.

Figure 1.18 shows the conditional distribution $F(X_3 \leq x_3 | X_1 = x_1, X_2 = x_2)$ of the two joint models. The plots show that the choice of constant or non-constant conditional copula affects the conditional distribution.

We have seen that the copula-vine approach can be used to construct very complicated joint distribution. Different bivariate copulas from different families can be used in the construction, and non-constant conditional copulas can be considered to add a further layer of complexity. Furthermore, the approach can be used when the variables are continuous, discrete or mixed.

1.2.4 Parameter Estimation and Model Simplification

1.2.4.1 Parameter Estimation

Given a set of data, the copula parameters of the copula-vine model can be estimated. The estimation is sequential, starting from the first tree, which consists of pairs of unconditional copula. When both discrete and continuous variables are present in the model, there are three types of pair: continuous, discrete and mixed discrete-continuous. The copula parameter is computed using equation (1.2), (1.14) or (1.15).

After all parameters in the first tree are estimated, we estimate those in the second tree. This is done by computing the *pseudo*-samples of the margins using the already estimated copulas in the first tree[10]. For instance, the parameter of $C_{ij|k}$ is computed as in the first tree, using pseudo-samples of $X_i|X_k$ and $X_j|X_k$ computed using the copulas C_{ik} and C_{jk} with parameters estimated in the previous step.

The estimation goes up to the higher trees in the same fashion until all parameters in the copula-vine model are computed.

1.2.4.2 Model Simplification

In the presence of data, we are interested in the copula-vine model which "best" represents that data. The flexibility of the copula-vine approach in modelling highly complicated joint distribution is especially useful when a highly complicated dependence structure is present in the data. Sometimes, however, the model does not need to be this complicated in order to represent the data effectively. In other words, a "parsimonious" model is desirable.

There are two possible approaches to obtain a parsimonious model. In the "forward" approach, information regarding insignificant parameters is already known and so only the significant ones are estimated. Such information may be available in the form of (conditional) independence statements between the variables and might come from, for instance, the model setup, experts' knowledge or by testing (conditional) independence from the data.

A popular way to illustrate the (conditional) independence statements between a set of variables is by using a directed acyclic graphical structure called the

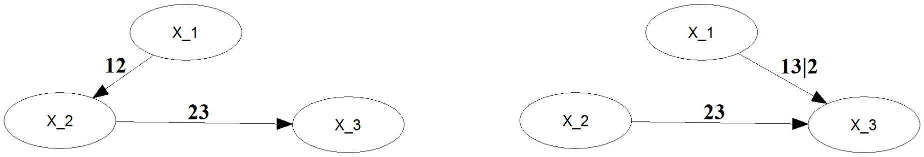
---

[10] The term "*pseudo*-samples" is used because these "samples" are not directly observed in the data but can only be computed after certain other parameters have first been estimated.

Bayesian network (BN). The graphical structure of a BN contains the (conditional) independence statements between the variables where the absence of an arc between two nodes corresponds to (conditional) independence between the two variables the two nodes represent. Underlying the graphical structure are the conditional probabilities between the variables, which specify the variables' relationships.

When all the variables are continuous, the conditional probabilities are in the form of conditional densities from which a joint density between the variables in the BN (with the conditional independence statements) is constructed. Kurowicka and Cooke [2005], extended in Hanea et al. [2006] and Hanea et al. [2010], show that the copula-vine approach can be used for this purpose. This type of BN model is called the "copula Bayesian network".

Example 1.2.11. Let X_1, X_2, X_3 be continuous variables as in Example 1.2.6. Consider two BN structures representing these three variables as in Figure 1.19[11]. The vine structure in Figure 1.12 can be used to model these two BNs.



(a) Structure 1.

(b) Structure 2.

Figure 1.19: *Two BN structures of three continuous variables* X_1, X_2, X_2.

The absence of an arc between X_1 and X_3 in Figure 1.19(a) implies conditional independence between X_1 and X_3 given X_2. This means that the copula $C_{13|2}$ in the decomposition in equation (1.13) is constant with respect to the variable X_2 and is the independence copula where $c_{13|2}(F_{1|2}(x_1|x_2), F_{3|2}(x_3|x_2)) = 1$.

In Figure 1.19(b), the absence of an arc between X_1 and X_2 implies independence between the two variables. This means that the copula C_{12} is the independence copula so $c_{12}(F_1(x_1), F_2(x_2)) = 1$ in the decomposition in equation (1.13).

Sometimes the (conditional) independence information is unknown. In this case, the "backward" approach can be used. Under this, all parameters are considered initially and then the insignificant ones are removed from the model. To determine the significance of a parameter, parametric bootstrapping can be performed. The parameter's confidence bound can then be computed. With this, the decision whether or not to remove the parameter from the model can be taken.

Several copula-vine software packages are available for use in practice. When the conditional copulas can be assumed to be constant, the `R` package `VineCopula` can be used to fit a regular copula-vine model to a set of data (Schepsmeier et al.

---

[11] The indices above the arcs represent the bivariate copulas used to represent the arcs, just as in Figure 1.12.

[2015]). When the bivariate copulas can be assumed to be the normal copulas, the `UNINET` [12] software developed at Delft University of Technology can be used[13]. `UNINET` has already been successfully applied in several other fields involving continuous variables, such as aviation safety modelling (Morales Napoles [2009]), fire safety (Hanea and Ale [2009]), and oil reservoir simulation (Hanea et al. [2013]; Zilko [2012]).

In this thesis we are going to construct a railway disruption-length model with the help of copulas. Moreover, some of the influencing factors (presented in Chapter 3) turn out to be discrete variables resulting in a mixed discrete-continuous model that needs to be dealt with. The next subsection presents several studies concerning the use of copulas in other transport research fields, as well as examples found in the literature of copula modelling in a mixed discrete-continuous setting.

1.2.5 Literature Study

1.2.5.1 Copulas and BN in Transport Research

The use of copulas in railway research is still fairly new. However, it is not a foreign concept in transport research generally. Srinivas et al. [2006] use several different copula families to model the dependence between vehicle axle weights. Wan and Kornhauser [1997] construct a copula-based model to predict travel time for use in a routing decision-making problem. Ng and Lo [2013] model the air-quality conformity in transportation networks with copulas. These three studies show the benefits to be gained by modelling the dependence between variables with copulas.

One approach we consider in this thesis is the copula Bayesian network. Under this, the disruption-length model is represented graphically as a BN. To quantify the conditional probabilities represented by the BN structure, we use the multivariate normal copula. The copula parameter R is estimated using the maximum-likelihood approach.

In transport research, BN modelling has been used in several studies but in a different setting where all the variables were discrete. For instance, Gregoriades and Mouskos [2013] quantify accident risk in road traffic in Cyprus and Chen et al. [2015] construct a dependence model of travellers' preference for toll road utilization in Texas with BNs. In the railway field, Oukhellou et al. [2008] use the technique to perform broken-rail diagnosis. In the case of continuous variables in Gregoriades and Mouskos [2013], the variables are discretized to obtain a fully discrete model. Because all the variables are discrete, these studies quantify the conditional probabilities with the conditional probability table (CPT).

Our second approach is to construct the disruption-length model with the copula-vine by considering several non-constant conditional copulas. This approach is new in railway research.

---

[12] `UNINET` is available at `www.lighttwist.net/wp/uninet`.

[13] All BN figures in the thesis are created using `UNINET`.

1.2.5.2 Copula Parameter Estimation with Discrete and Continuous Variables

A popular way to estimate the copula parameter is by finding an empirical dependence measure, for instance Kendall's tau, τ, or the Spearman's rho, ρ, and equating it to the copula parameter. This approach is based on the assumption that the empirical dependence measure is a one-to-one mapping with the copula parameter. Chapter 5 of Nelsen [2006] provides some of these relationships for several copula families.

This approach works well when the variables are continuous. When some of the variables are discrete, however, Genest and Nešlehová [2007] have shown that it is highly biased. Nevertheless, the maximum likelihood technique can still be used, even though it is much more computationally expensive.

Maximum-likelihood estimation of copula parameters for the n-dimensional discrete model requires an approximation of multidimensional integral or evaluating 2^n finite differences of the copula to find the value of the probability mass function. Due to computational costs, many copula applications on discrete models have only worked on lower dimensional problems. Nikoloulopoulos and Karlis [2008] construct a four-dimensional Bernoulli distribution with the help of several different copula families with three parameters, while Song et al. [2009] build a trivariate discrete distribution with the normal copula. In both cases, the copula models work well and the authors highlight that the dependence structure between the variables come not only from the copula but also from the margins.

Nikoloulopoulos [2013] proposes the computation of rectangle probabilities using the simulated maximum-likelihood approach method. This new approach has been shown in Nikoloulopoulos [2015] to be applicable in dimensions of up to 225 even though computational burden becomes heavy as dimension and sample size increase. Another alternative technique to estimate the parameters uses the Bayesian methods, as proposed by Smith and Khaled [2012]. However, this technique is also computationally intensive.

The copula-vine approach with constant conditional copula is used to construct the multivariate discrete distribution in Panagiotelis et al. [2012]. It is shown here that this approach reduces the computation cost of calculating the probability mass function to $2n(n-1)$ evaluations only, which makes this model applicable even for very high-dimensional problems. With this approach, unlike in the fully continuous setting, it is "easier" to detect from the data whether the constant conditional copula assumption holds or not. This is because more samples are observed for each state of the conditioning variable(s) from which a copula needs to be fitted. Zilko and Kurowicka [2016] explore the model construction with non-constant conditional copulas and some of their results are presented in Chapter 2 of this thesis.

As in the purely discrete models, problems are also encountered when the models are of mixed discrete and continuous type. Most applications of copulas available in the literature are to low-dimensional problems. Song et al. [2009] model bivariate mixed binary discrete (disposition) and continuous (severity of burn injury) variables with a normal copula. De Leon and Wu [2010] propose two strategies to compute the maximum likelihood for a bivariate mixed discrete and continuous distributions, with a simulation study and an application to the same dataset as in Song et al. [2009]. He et al. [2012] use the normal copula to construct two and three-dimensional mixed discrete and continuous models, each with one discrete variable to study the relationship between the genotype (discrete) and a few continuous phenotypes, such as cholesterol density and protein concentration. Stöber et al. [2015] construct a six-dimensional mixed discrete and continuous model with five binary variables and one continuous variable representing six chronic diseases by following the copula-vine approach with constant conditional copulas as described in Panagiotelis et al. [2012].

1.2.6 Model Comparison

In this thesis we need to test whether the distribution of a continuous random variable can be modelled with a distribution from a parametric family. For instance, early in this section we fit the distributions of disruption length in the data to the normal and exponential distribution. Plotting the fitted and empirical distributions, as in Figure 1.8, indicates that neither parametric distribution represents the disruption length we observe in the data. Goodness-of-fit tests can be performed for this purpose.

For a one-dimensional case, many goodness-of-fit tests are available. The two we will use most in this thesis are the Kolmogorov-Smirnov (KS) test and the Cramér-von Mises (CvM) test. Let F denote the hypothesized distribution of a random variable X whose empirical distribution is $\hat{F}$. The KS test measures:

$$\sup_x |\hat{F}(x) - F(x)|$$

while the CvM test measures:

$$\int_{-\infty}^{\infty} [\hat{F}(x) - F(x)]^2 dF(x).$$

From each test, the p-value is computed to determine whether there is evidence in the data that F is not close enough to $\hat{F}$. The p-value measures how extreme the observed difference is. When the p-value is lower than a significance level α, the difference between F and $\hat{F}$ is too high to be explained by noise in the data and so the test concludes that F cannot be used to represent $\hat{F}$. In this thesis, we choose $\alpha = 0.05$ as the significance level.

Performing the KS and CvM tests to the fitted normal and exponential distribution of the disruption length yields p-values in the order of 10^{-3} or lower. This confirms our observation that neither distribution is appropriate to represent the disruption-length data.

The two-sample KS and CvM tests are also available. Instead of testing the fit of a parametric distribution to a set of data, these measure whether the distributions of two sets of data are close to each other or not. The two-sample KS and CvM tests are used in Chapter 3 of this thesis.

The joint models constructed in this thesis are copula-based. Once a copula is fitted, it is of interest to perform goodness-of-fit test to see whether the copula can indeed be used to represent the data. The following goodness-of-fit tests are used throughout this thesis.

1.2.6.1 The Kullback-Leibler (KL) Divergence Test

Let the joint probability of n discrete variables $\mathbf{P} = \mathbb{P}(X_1 = x_1, \dots, X_n = x_n)$ be modelled with the copula C_R as:

$$\mathbf{Q} = \mathbb{P}(X_1 = x_1, \dots, X_n = x_n) = \sum_{s_1=0}^{1} \dots \sum_{s_n=0}^{1} (-1)^{\sum_i s_i} C_R(\mathbb{P}(X_1 \leq x_1 - s_1), \dots, \mathbb{P}(X_n \leq x_n - s_n)).$$

The KL test measures the difference between $\mathbf{P}$ and $\mathbf{Q}$ as:

$$D_{KL} = \sum_{x_1} \dots \sum_{x_n} \mathbf{P} \log\left(\frac{\mathbf{P}}{\mathbf{Q}}\right) \tag{1.17}$$

with the null hypothesis $\mathbf{P} = \mathbf{Q}$ tested against the alternative $\mathbf{P} \neq \mathbf{Q}$. Under the null-hypothesis for large enough number of samples N, $2 \cdot N \cdot D_{KL}$ is chi-squared distributed from where the p-value can be computed (see, e.g., Cover and Thomas [2006]).

1.2.6.2 The Probability Integral Transform (PIT) Test

We follow the presentation of the test in Breymann et al. [2003]. The test is based on the Rosenblatt probability integral transform (Rosenblatt [1952]). The null hypothesis of the PIT test is that C_R models the dependence of the joint distribution of n continuous variables against the alternative hypothesis that this is not the case.

Let $\mathbf{U} = (\tilde{U}_1, \dots, \tilde{U}_n)$ be random vector with uniform margins. Then, define $U_1, \dots, U_n$ as:

$$U_1 = \tilde{U}_1, \quad U_2 = C_{R_{[1,2]}}(\tilde{U}_2 | \tilde{U}_1), \dots, \quad U_n = C_R(\tilde{U}_n | \tilde{U}_1, \dots, \tilde{U}_{n-1}) \tag{1.18}$$

where C_R denotes a copula with parameter R and $R_{[1,2]}$ is an element of R corresponding to the dependence between the first and second variables. If the copula represents the data well, U_i are uniform on $(0,1)$ for each i and are independent. This means that the test-statistic:

$$V = \sum_{i=1}^{n} \left(\Phi^{-1}(U_i)\right)^2, \tag{1.19}$$

is chi-square distributed with n degrees of freedom. The vector $\mathbf{U}$ is obtained from data by transforming the margins to uniform using their marginal distributions.

Berg [2009] shows that the PIT test performs well in testing the normal copula hypothesis. Because this is the copula we are interested in, the PIT test is used in this thesis. Several other copula goodness-of-fit tests, along with their power comparisons, are available in Berg [2009].

However, when the marginal distributions are estimated with the empirical distributions, Dobrić and Schmid [2007] report a reduced power of the test to reject a wrong null hypothesis.

One common assumption in the copula-vine approach is the constant conditional copula assumption. The following test can be used to check whether this is validated by the data.

1.2.6.3 The Vectorial Independence Test

In the case of continuous variables, a test developed by Quessey [2010] can be used to test the assumption of constant conditional copula. The null hypothesis of this test is the independence between $(U,V) = (F_{X_j|\mathbf{V}_{\setminus\mathbf{i}}}, F_{X_i|\mathbf{V}_{\setminus\mathbf{i}}})$ and $\mathbf{Z} = \mathbf{V}_{\setminus\mathbf{i}}$ and is tested against the alternative hypothesis where independence does not hold.

Therefore, under the null hypothesis, for $(u,v,\mathbf{z}) \in [0,1]^n$,

$$C_{U,V,\mathbf{Z}}(u,v,\mathbf{z}) = C_{U,V,\mathbf{Z}}(u,v,\mathbf{1})C_{U,V,\mathbf{Z}}(1,1,\mathbf{z}). \tag{1.20}$$

Kojadinovic and Holmes [2009] and Quessey [2010] measure the difference between the left-hand and the right-hand sides of (1.20) using the CvM test-statistic. Under the null hypothesis, the test statistic should be zero. To indicate whether the observed test statistic in the data is close enough to zero or not, Kurz [2013] estimates the p-value based on the work of Quessey [2010]. For further details, readers are referred to Kurz [2013].

In this thesis, several possible disruption-length models are constructed and comparisons made between them. Two models can be compared against each other by checking their likelihoods. In principle, a model with higher likelihood represents the data better. However, the number of model parameters should be considered as well. A fair model comparison takes into the account both the likelihood and the number of parameters.

The following are five different ways to compare competing models' likelihoods and their respective number of parameters. They can be used as means for model selection.

1.2.6.4 The Akaike Information Criterion

The Akaike information criterion (AIC) is introduced in Akaike [1974]. For each model, its AIC score is computed as:

$$AIC = 2k - 2\ln(L) \tag{1.21}$$

where k denotes the number of parameters involved in the model and L denotes the model's likelihood. Different models are compared against each other using their respective AIC scores. In principle, the model with the higher likelihood, and hence the lower AIC score, is the better model. Note that the AIC "penalizes" a model with a higher number of parameters.

1.2.6.5 The Bayesian Information Criterion

Another popular criterion for model comparison is the Bayesian information criterion (BIC) or Schwarz criterion. This is introduced in Schwarz [1978] and is defined as:

$$BIC = k\ln(N) - 2\ln(L) \tag{1.22}$$

where N denotes the number of samples and k and L are as in (1.21). As with the AIC, the preferred model is the one with the lower BIC score.

While the AIC and BIC can be used to determine which of two competing models is the better one to represent the data concerned, they do not indicate "how much better" one is than the other. In other words, the significance of the observed difference between the two models cannot be concluded from either the AIC or the BIC.

For this reason, a number of statistical tests have been developed. Three of these are presented next.

1.2.6.6 The Likelihood Ratio Test

The likelihood ratio test compares the ratio of the likelihoods of the two competing models. To use it, the two competing models have to be *nested*. Two models are said to be *nested* if one can be transformed into the other by imposing constraints on the first model's parameters. If not, the two models are *non-nested*.

The test statistic is:

$$LRT = -2\ln\left(\frac{L_S}{L_G}\right) \tag{1.23}$$

where L_S and L_G denote the maximum likelihood of the simpler and the more general model, respectively. The maximum likelihood of the more general model will always be greater than or equal to that of the simpler model. The significance of the difference in the two likelihoods is then tested. Wilkes [1938] shows that the test statistic LRT is asymptotically χ^2-distributed with degrees of freedom equal to the number of parameters difference between the two models.

While the likelihood ratio test is useful when two competing models are nested, in practice they often are not. Two tests which do allow comparison of competing non-nested models are presented below.

1.2.6.7 The Vuong Test

The Vuong test, introduced by Vuong [1989], also uses the likelihood ratio to measure the difference between the two competing models.

Consider two competing models F and G with parameters α and β, respectively. It tests whether the mean of the likelihood ratios between the two models is close to one, indicating that both are equally close to the data. Vuong [1989] shows that, under general conditions, the following holds in the null hypothesis:

$$Z = \frac{L\tilde{R}_N(\alpha,\beta)}{\sqrt{N}\omega_n} = \frac{LR_N(\alpha,\beta) - K_N(F,G)}{\sqrt{n}\omega_n} \xrightarrow{D} N(0,1) \tag{1.24}$$

where $LR_N(\alpha,\beta) = \ln(L_F) - \ln(L_G)$ is the difference between the log-likelihoods of the two models, $K_N(F,G)$ is a correction factor accounting for the difference in the number of parameters between the two models, N is the number of samples, and ω_n is the standard deviation of the individual log-likelihood ratio. The null hypothesis of the test is that Z is standard normally distributed while the alternative hypothesis states that it is not. Vuong [1989] suggests the use of a correction factor corresponding to the AIC, $K_N(F,G) = a - b$, or to the BIC, $K_N(a,b) = \frac{a-b}{2}\ln(N)$, where a and b are the number of parameters of model F and G, respectively.

1.2.6.8 Clarke's Distribution-Free Test

As an alternative to the Vuong Test, Clarke [2003] proposes a paired sign test to measure the difference between the individual log-likelihoods from two non-nested models. Under the null hypothesis, half of the individual likelihoods of one model should be larger than the other model's. The test statistic is:

$$B = \sum_{i}^{N} I\left(\log\left(\frac{L_{F_i}}{L_{G_i}}\right) > 0\right) \tag{1.25}$$

where N is the number of samples, L_{F_i} and L_{G_i} are the individual likelihoods of sample i, respectively, and I is the indicator function. The null hypothesis is that B is distributed binomially with parameters N and $p = 0.5$. This is tested against the alternative hypothesis where B is not binomially distributed with parameters N and $p = 0.5$. To account for the different number of parameters between the two models, Clarke [2003] proposes applying the *average* correction factor to the individual log-likelihood ratios with either the AIC (a/N and b/N to model F and G, respectively) or the BIC ($\frac{a}{2N}\ln(N)$ and $\frac{b}{2N}\ln(N)$ to model F and G, respectively).

1.3 The Disruption-Length Models in Practice

The constructed joint distribution is conditioned on the observed values of the influencing factors. The result is in the form of the conditional distribution of disruption lengths. Therefore, the model output is in the form of a probability distribution.

In practice it is difficult to implement a train traffic optimization algorithm with a stochastic disruption-length input, especially with a highly complex railway network like the one in the Netherlands. Consequently, the optimization algorithm is often simplified and takes a deterministic value of disruption length as its input. In our case, this means that one value needs to be chosen from the conditional distribution of disruption length to be taken as the disruption-length prediction.

Statistically, choosing the mean of this distribution minimizes the mean-squared error (MSE) of the prediction while choosing the median minimizes the average absolute deviation. However, the chosen prediction affects the train traffic and the passengers, who are the end customers of ProRail. A prediction that is optimistic (too short) might have different impact on the customers than a prediction that is pessimistic (too long). Therefore, it is of model application importance to study the impact of different choices of prediction on the train traffic and the passengers.

To investigate this issue, a study has been performed in collaboration with the Department of Transport and Planning at Delft University of Technology. The chosen disruption-length model was applied, together with a train short-turning model and a passenger-flow model, in a disrupted train traffic in the region of Houten, the Netherlands. This disrupted area is part of the important A2 railway corridor, connecting Amsterdam and Eindhoven via Utrecht and 's-Hertogenbosch (Den Bosch)[14].

In the study, the disruption completely blocked a section of line between Utrecht-Lunetten and Houten stations, so that no train service could run between them until the problem was solved. Naturally, this would impact train traffic and the passengers travelling in the region. The short-turning model computed the optimal train timetable during the disruption, based on the chosen prediction of its length. The passenger-flow model then used this timetable to compute passenger movements around the area.

The results are presented in Chapter 4 of this thesis.

1.4 Thesis Organization

The rest of the thesis is organized as follows:

In Chapter 2 we study the use of copulas in fully discrete and mixed discrete-continuous settings. The copula-vine approach is adapted to construct such a mixed model, and an algorithm to fit the model parameters is proposed. The algorithm is also tested in several simulation studies to observe its behaviour.

The application part of this thesis starts in Chapter 3, the main topic of which is the construction of the TC disruption-length model. We begin the chapter with data analysis to determine the important factors influencing latency and repair times. With these identified, a small number of candidate disruption-length models for TC failures are considered and compared to determine which is most appropriate to predict the length of a disruption to the railway. The construction of the switch disruption-length model follows the same pattern and is also presented briefly in this chapter.

In Chapter 4 we show how the disruption-length model we develop in this thesis can be used in the real world. The practical challenges are also presented.

---

[14] The full corridor extends to Alkmaar in the north-west of the Netherlands and Maastricht in the south-east.

The disruption-length model is used in a collaboration with the short-turning and passengers-flow models, that are developed by two PhD candidates in the Department of Transport and Planning of Delft University of Technology. The three models are applied in a disrupted train traffic in the Houten area. A number of case studies are considered from which some preliminary conclusions are drawn.

Finally, this thesis closes with a summary and recommendations, which we present in Chapter 5.

CHAPTER 2

Copula In A Mixed Discrete-Continuous Model[1]

In this chapter we concentrate on theoretical issues concerning the use of copulas for purely discrete and mixed discrete-continuous models. We are interested in the normal copula in particular, because this is the one intended for use in the railway disruption-length models later on. The goal is to construct a model that allows fast and accurate prediction of our variable of interest for different combinations of values of other variables in the model.

For the purely discrete models, we start by considering the multivariate Bernoulli distribution. We study the possibility of representing the dependence of such distribution with the multivariate normal copula, the one which corresponds to the copula Bayesian network model. The challenges and problems of such an approach are investigated. After that, the copula-vine approach popular in fully continuous dependence modelling is adapted to the multivariate Bernoulli case. This is a more flexible approach, which allows more parameters to be involved and so means that, in theory, a more accurate model can be constructed.

Next, we extend our study by considering some variables to have more than two states. Eventually, we end up with a mixed discrete-continuous problem. Again, the copula-vine approach can be adapted to the mixed discrete-continuous setting.

To estimate the parameter values of the mixed discrete-continuous copula-vine model, a sequential algorithm is proposed. The performance of this algorithm, including the consideration of some possible copula and vine structure misspecifications, is tested in low-dimensional problems.

---

[1] The first part of this chapter is based on Zilko and Kurowicka [2016].

2.1 Multivariate Bernoulli Distribution with Copulas

Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a random vector taking values in $\{0,1\}^n$ and $\mathbf{x} = (x_1, \ldots, x_n)$ be a realization of $\mathbf{X}$. The joint probability is

$$\begin{aligned}\mathbb{P}(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) &= p(x_1, x_2, \ldots, x_n)\\ &= p(0,0,\ldots,0)^{\prod_{j=1}^{n}(1-x_j)} p(1,0,\ldots,0)^{x_1 \prod_{j=2}^{n}(1-x_j)} \ldots p(1,1,\ldots,1)^{\prod_{j=1}^{n} x_j}\end{aligned} \tag{2.1}$$

where all the p's must add up to 1. The marginal distribution of X_i is

$$\mathbb{P}(X_i = 0) = p_i = \sum_{x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_n \in \{0,1\}} p(x_1, \ldots, x_{i-1}, 0, x_{i+1}, \ldots, x_n).$$

Another popular representation of a multivariate Bernoulli distribution is the log-linear expansion. Taking the logarithm of the probability in (2.1) and collecting the appropriate terms leads to:

$$\log p(x_1, x_2, \ldots, x_n) = \log p(0,0,\ldots,0) + \sum_i u_i x_i + \sum_{i,j} u_{ij} x_i x_j + \sum_{ijk} u_{ijk} x_i x_j x_k + \ldots + u_{12\ldots n} x_1 x_2 \ldots x_n. \tag{2.2}$$

The u-terms in (2.2) represent the two, three, …, n-way interactions between the variables (see, e.g., Whittaker [1990]) and can be obtained from the probabilities as follows:

$$\begin{aligned}u_1 &= \log \frac{p(1,0,0,\ldots,0)}{p(0,0,0,\ldots,0)},\\ u_{12} &= \log \frac{p(1,1,0,\ldots,0)p(0,0,0,\ldots,0)}{p(1,0,0,\ldots,0)p(0,1,0,\ldots,0)},\\ u_{123} &= \log \frac{p(1,1,1,0,\ldots,0)p(1,0,0,0,\ldots,0)p(0,1,0,0,\ldots,0)p(0,0,1,0,\ldots,0)}{p(1,1,0,0,\ldots,0)p(1,0,1,0,\ldots,0)p(0,1,1,0,\ldots,0)p(0,0,0,0,\ldots,0)}.\end{aligned} \tag{2.3}$$

The interactions between the variables contain information about dependence. The term u_{12} is also known as the log cross-product ratio (cpr) between variables X_1 and X_2. Note that the cross-product ratio $cpr(X_1, X_2)$ can be rewritten in terms of conditional probabilities of variables X_1 and X_2 given that all remaining variables $X_3, \ldots, X_n$ equal zero. Moreover, u_{123} is the logarithm of the ratio of the cross product ratio of variables X_1, X_3 given $X_2 = 1$, and the cross product ratio of X_1, X_3 given $X_2 = 0$.

The symbol u_{ij} represents the two-way dependence between the variables X_i and X_j. However the dependence between X_i and X_j is also affected by all higher-order interactions containing these variables.

Example 2.1.1. (the trivariate Bernoulli Distribution). The trivariate Bernoulli distribution of (X_1, X_2, X_3) is

$$\mathbb{P}(X_1 = x_1, X_2 = x_2, X_3 = x_3) = p(0,0,0)^{(1-x_1)(1-x_2)(1-x_3)} \ldots p(1,1,1)^{x_1 x_2 x_3} \tag{2.4}$$

for $x_1, x_2, x_3 \in \{0, 1\}$. Its log-linear expansion is:

$$\log p(x_1, x_2, x_3) = u_\emptyset + u_1 x_1 + u_2 x_2 + u_3 x_3 + u_{12} x_1 x_2 + u_{13} x_1 x_3 + u_{23} x_2 x_3 + u_{123} x_1 x_2 x_3. \tag{2.5}$$

where $u_\emptyset = \log p(0, 0, 0)$ and the u-terms are as presented in (2.3).

The conditional distribution of X_1 and X_3 given $X_2 = x_2$ is a bivariate Bernoulli distribution. Let the log cross-product ratio of this conditional distribution be denoted as $u_{13|2=x_2}$. When $u_{13|2=x_2} = 0$, the variables $X_1|X_2 = x_2$ and $X_3|X_2 = x_2$ are independent. Moreover, when $u_{13|2=x_2} = 0$ for both realizations of $x_2 = 0$ and $x_2 = 1$, the variables X_1 and X_3 are conditionally independent given variable X_2. Note that $u_{13} = u_{13|2=0}$ and $u_{123} = \log\left(\frac{cpr(X_1, X_3|X_2=1)}{cpr(X_1, X_3|X_2=0)}\right)$. Hence X_1 and X_3 are conditionally independent given variable X_2 if and only if $u_{123} = 0$ and $u_{13} = 0$.

The above relationships can be generalized for higher-order interactions and allow independencies and conditional independencies to be read from the log-linear expansion by examining the u-terms. Moreover, if the random vector $(X_1, \ldots, X_n)$ has the Bernoulli distribution, then it is easy to see that the conditional distributions are also Bernoulli.

The dependencies between variables with Bernoulli distributions are contained in the u-terms of their log-linear expansions. In this thesis, we consider modelling these dependencies by means of copula. Our study starts with the simplest case possible, the bivariate Bernoulli distribution.

2.1.1 Bivariate Bernoulli Distribution with Copulas

To illustrate how copulas are used to model discrete distributions and to present the graphical interpretation of equation (1.2), a bivariate Bernoulli random vector (X_1, X_2) with margins p_1, p_2 is considered. Moreover, let U_1 and U_2 be uniform random variables with copula C. The probability mass function of (X_1, X_2) can be represented in terms of the *latent* variables U_1 and U_2 with copula C as follows:

$$\mathbb{P}(X_1 = x_1, X_2 = x_2) = \begin{cases} p(0,0), & U_1 \leq p_1, U_2 \leq p_2; \\ p(0,1), & U_1 \leq p_1, U_2 > p_2; \\ p(1,0), & U_1 > p_1, U_2 \leq p_2; \\ p(1,1), & U_1 > p_1, U_2 > p_2. \end{cases} \tag{2.6}$$

Figure 2.1 shows the above construction graphically. The two axes in Figure 2.1 correspond to the latent vector (U_1, U_2). The range $U_1 \in (0, p_1]$ in the vertical axes corresponds to the realization $X_1 = 0$; and $U_2 \in (0, p_2]$ on the horizontal axes to $X_2 = 0$. The mass in the bottom left rectangle is $\mathbb{P}(X_1 = 0, X_2 = 0) = p(0, 0)$.

In this case, equation (1.2) takes the form:

$$\mathbb{P}(X_1 \leq x_1, X_2 \leq x_2) = C(\mathbb{P}(X_1 \leq x_1), \mathbb{P}(X_2 \leq x_2)). \tag{2.7}$$

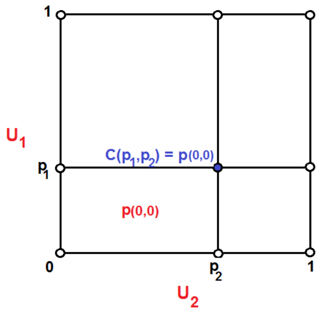


Figure 2.1: *The unit square corresponding to the latent variables* (U_1, U_2)

Since $\mathbb{P}(X_1 \leq 1, X_2 \leq x_2) = \mathbb{P}(X_2 \leq x_2) = C(1, \mathbb{P}(X_2 \leq x_2))$ for $x_2 \in \{0,1\}$ and $\mathbb{P}(X_1 \leq x_1, X_2 \leq 1) = \mathbb{P}(X_1 \leq x_1) = C(\mathbb{P}(X_1 \leq x_1), 1)$ for $x_1 \in \{0,1\}$ hold for any copula, the only constraint on C to realize the distribution of (X_1, X_2) is:

$$p(0,0) = \mathbb{P}(X_1 \leq 0, X_2 \leq 0) = C(\mathbb{P}(X_1 \leq 0), \mathbb{P}(X_2 \leq 0)) = C(p_1, p_2). \tag{2.8}$$

For continuous random vectors, there exists a unique copula that models the dependence of the joint distribution. However, this copula is constrained to satisfy the Sklar's theorem in every point of the unit hypercube. In the bivariate Bernoulli case, where the constraint is at only one point in the unit square, any copula satisfying (2.8) will be appropriate to model the dependence of (X_1, X_2).

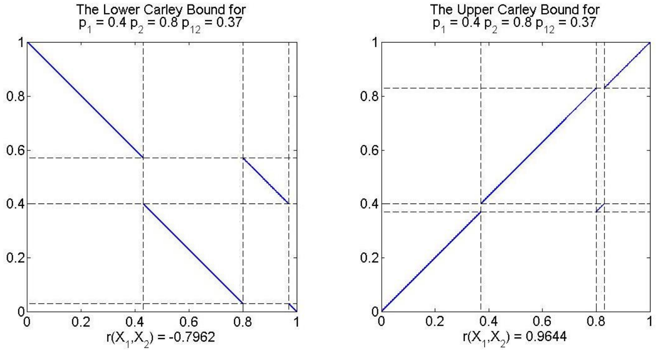
The bounds on copulas satisfying (2.8) have been presented in Carley [2002]. The upper (lower) Carley bound belongs to the family of copulas constructed as a shuffle of the upper (M) and lower (W) Fréchet bounds, where $M = \min(u, v)$ and $W(u,v) = \max(u + v - 1, 0)$ for $(u, v) \in (0,1)^2$ (Nelsen [2006]). The mass in M and W is concentrated uniformly on the diagonal and anti-diagonal of the unit square, respectively. For (U_1, U_2) with copula M (W), the Spearman's correlation is $\rho(U_1, U_2) = 1(-1)$. The following example illustrates this.

Example 2.1.2. Consider a bivariate Bernoulli random variable with $p_1 = 0.4$, $p_2 = 0.8$, and $p(0,0) = 0.37$. The upper and lower Carley bounds of all copulas satisfying (1.2) for this case are shown in Figure 2.2. The mass of the lower and upper Carley bound copulas is distributed uniformly along the anti-diagonal and diagonal lines in the Figure 2.2, respectively. Both copulas satisfy (2.8) because the mass in the bottom left-hand rectangle $(0, p_1] \times (0, p_2] = (0, 0.4] \times (0, 0.8]$ is 0.37.

The correlation of the lower Carley bound is -0.7962, while the correlation of the upper Carley bound is 0.9644. Therefore, when $p_1 = 0.4, p_2 = 0.8$ and $p(0,0) = 0.37$, the correlation of any copula satisfying (2.8) varies from as low as -0.7962 to as high as 0.9644.

If a one-parametric copula family has already been chosen to work with, the non-uniqueness problem of the copulas satisfying (2.8) is avoided. However, one might then face the problem of non-existence of a copula in the chosen class that

Figure 2.2: *The lower and upper Carley bounds with their corresponding correlation.*



satisfies equation (2.8). The theorem below gives the conditions a copula has to satisfy in order to be able to recover a bivariate Bernoulli distribution.

Theorem 2.1.1. *Let (X_1, X_2) be a Bernoulli distributed random vector. Let C_θ be a one-parametric copula that is continuous with respect to θ and satisfies:*

$$\lim_{\theta \to \theta_L} C_\theta(u,v) = W(u,v) \text{ and } \lim_{\theta \to \theta_U} C_\theta(u,v) = M(u,v) \text{ for } (u,v) \in (0,1)^2$$

for some θ_L and θ_U, where $W(u,v)$ and $M(u,v)$ are the lower and upper Fréchet bounds, respectively. Then, there exists a θ which satisfies:

$$C_\theta(\mathbb{P}(X_1 \leq 0), \mathbb{P}(X_2 \leq 0)) = C_\theta(p_1, p_2) = p(0,0) = \mathbb{P}(X_1 \leq 0, X_2 \leq 0). \tag{2.9}$$

Proof. Since any copula at point (p_1, p_2) has to lie between the lower and upper Fréchet bounds at (p_1, p_2), $p(0,0)$, it has to satisfy:

$$W(p_1, p_2) \leq p(0,0) \leq M(p_1, p_2). \tag{2.10}$$

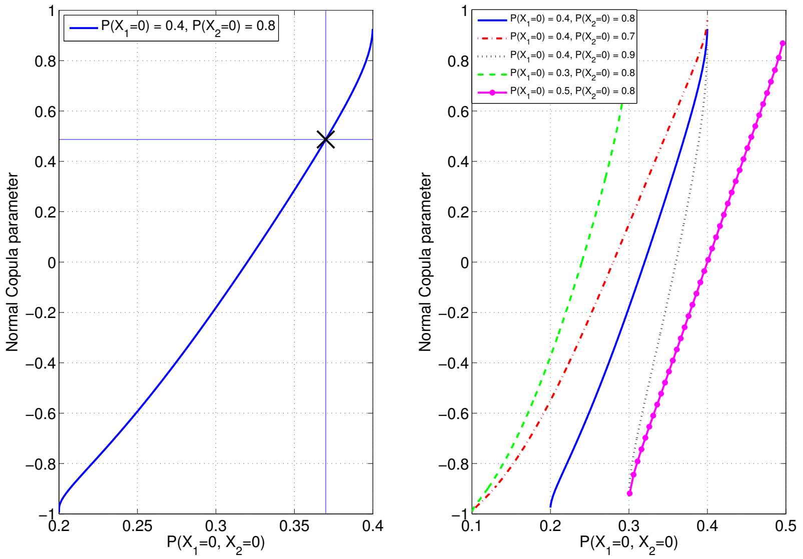
Since $\lim_{\theta \to \theta_L} C_\theta(p_1, p_2) = W(p_1, p_2)$, $\lim_{\theta \to \theta_U} C_\theta(p_1, p_2) = M(p_1, p_2)$, and C_θ is continuous with respect to θ, inequality (2.10) together with the Intermediate Value Theorem guarantee the existence of $\theta \in [\theta_L, \theta_U]$, such that equation (2.9) is satisfied. ■

Corollary 2.1.2. *The bivariate normal copula as in equation (1.4) is continuous with respect to the parameter ρ and:*

$$\lim_{\rho \to -1} C_\rho(u,v) = W(u,v) \quad \text{and} \lim_{\rho \to 1} C_\rho(u,v) = M(u,v).$$

Therefore, according to Theorem 2.1.1, the solution to equation (2.9) exists for the normal copula.

Figure 2.3: *Plot of the parameter of the normal copula versus the joint probability of variables* X_1 *and* X_2.

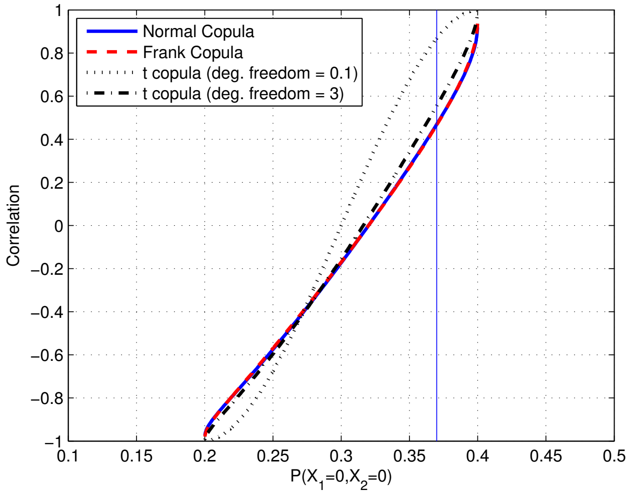


The corollary above states that one can always find a normal copula which corresponds to a bivariate Bernoulli random variable. We concentrates mainly on the normal copula as this is the one we intend to use in the disruption-length models later on.

Figure 2.3 (left) shows the relationship between the parameter of the normal copula and the probability $p(0,0) = \mathbb{P}(X_1 = 0, X_2 = 0)$ of Bernoulli distribution with margins $p_1 = 0.4$ and $p_2 = 0.8$. The joint probability $p(0,0)$ is bounded by $\max(p_1 + p_2 - 1, 0)$ and $\min(p_1, p_2)$. When $p(0,0) = 0.37$, as in Example 2.1.2, the parameter of the normal copula is $\rho = 0.4868$. Figure 2.3 (right) illustrates the relationship between the parameter of the normal copula and $p(0,0)$ in case of different univariate margins.

The normal copula is often applied in practice. However, the relationships between the margins, $p(0,0)$, and the parameters of other copulas can be found as well. These relationships are not available in straightforward analytic form, but in the bivariate case they can be calculated easily. Figure 2.4 illustrates the relationship between $p(0,0)$ and Spearman's correlations realized by normal, Frank's and Student-t copulas in the case of bivariate Bernoulli distribution, as in Example 2.1.2, showing that the relationships differ slightly for different copula families.

Figure 2.4: *Plot of the Spearman's correlation of the copula versus the joint probability between variables X_1 and X_2.*

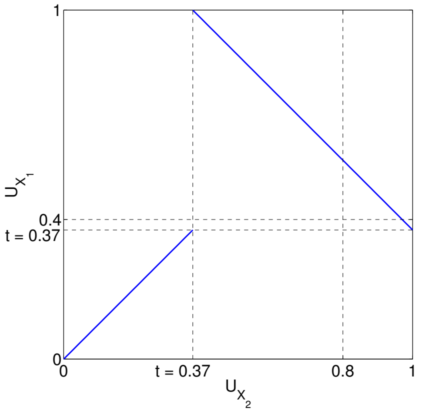


Example 2.1.3. Consider a copula C_t, where $t \in [0,1)$ and is defined as:

$$C_t(u,v) = \begin{cases} \max(u+v-1,t), & (u,v) \in [t,1]^2; \\ \min(u,v), & \text{otherwise} \end{cases} \tag{2.11}$$

as in Exercise 2.10 of Nelsen [2006].

Figure 2.5: *Copula C_t with $t = 0.37$ as in equation (2.11) representing the bivariate Bernoulli distribution in Example 2.1.2.*



Obviously, $\lim_{t\to 0} C_t(u,v) = W(u,v)$ and $\lim_{t\to 1} C_t(u,v) = M(u,v)$. Moreover, rewriting C_t as a function of t yields:

$$C_t(u,v) = \begin{cases} u+v-1, & \text{for } 0 \le t \le \max(0,u+v-1); \\ t, & \text{for } \max(0,u+v-1) \le t \le \min(u,v); \\ \min(u,v), & \text{for } t \ge \min(u,v). \end{cases}$$

This means that C_t is a continuous function with respect to t.

According to Theorem 2.1.1, therefore, the solution to equation (2.9) exists for this copula. With $t = 0.37$, this copula represents the bivariate Bernoulli distribution in Example 2.1.2 and is depicted in Figure 2.5.

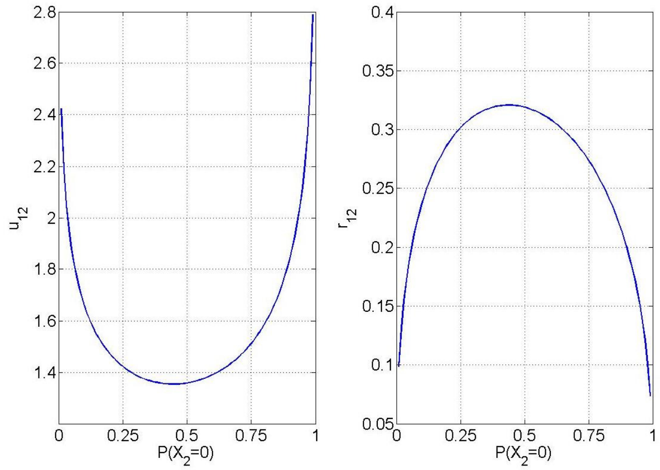
The next example shows that not every copula family can be used to recover a given bivariate Bernoulli distribution.

Example 2.1.4. The Morgenstern copula defined as:

$$C_\theta(u,v) = uv(1+\theta(1-u)(1-v)) \tag{2.12}$$

where $(u,v) \in (0,1)^2$ and $\theta \in [-1,1]$ is known to not span the entire lower and upper Fréchet bounds (Nelsen [2006]). Therefore, the Morgenstern copula might not satisfy equation (2.9) for some choices of p_1, p_2, and $p(0,0)$. It is easy to see this for the maximum value of $\theta = 1$, $C_{\theta=1}(0.4,0.8) = 0.3584$, which means that the bivariate Bernoulli distribution in Example 2.1.2 cannot be obtained with this copula.

Figure 2.6: *The log cross product ratio and rank correlation of bivariate Bernoulli with varying* $p_2 \in (0,1)$.



Next, we check whether the properties of the latent random vector (U_1,U_2) translate to equivalent properties of its corresponding Bernoulli distributed variables (X_1,X_2). The example below shows that, in contrast to continuous variables, two Bernoulli distributions constructed with the same copula have very different dependencies.

Example 2.1.5. Let a bivariate normal copula with parameter $\rho = 0.4868$ be a distribution of latent variables U_1, U_2. We fix the first margin as $p_1 = 0.4$ and the second margin can vary: $p_2 \in (0,1)$. For each p_2, the log cross-product ratio and rank correlation of the corresponding Bernoulli distribution are calculated.

As presented in Figure 2.6, it turns out that both values are different for different choices of p_2. Moreover, the minimum log cross-product ratio is obtained at $p_2 = 0.4459$ and the maximum rank correlation is obtained at $p_2 = 0.4375$.

Example 2.1.5 shows that dependencies of a Bernoulli distributed random vector depend not only on the copula, but also on the marginal distributions. This observation is in line with previous contributions, in Denuit and Lambert [2005], Mesfioui and Tajar [2005], and Nešlehová [2007].

2.1.2 Trivariate Bernoulli Distribution with Copulas

In the case of a three-dimensional Bernoulli distribution, the Sklar's equality (1.2) takes the following form:

$$\mathbb{P}(X_1 \le x_1, X_2 \le x_2, X_3 \le x_3) = C(\mathbb{P}(X_1 \le x_1), \mathbb{P}(X_2 \le x_2), \mathbb{P}(X_3 \le x_3)).$$

Given that the univariate margins of the Bernoulli random vector (X_1, X_2, X_3) are fixed and all probabilities have to add up to 1, a copula C needs to satisfy the following four equations:

$$\begin{cases} \mathbb{P}(X_1 \le 0, X_2 \le 0, X_3 \le 0) = C(\mathbb{P}(X_1 \le 0), \mathbb{P}(X_2 \le 0), \mathbb{P}(X_3 \le 0)), \\ \mathbb{P}(X_1 \le 0, X_2 \le 0, X_3 \le 1) = C(\mathbb{P}(X_1 \le 0), \mathbb{P}(X_2 \le 0), 1), \\ \mathbb{P}(X_1 \le 0, X_2 \le 1, X_3 \le 0) = C(\mathbb{P}(X_1 \le 0), 1, \mathbb{P}(X_3 \le 1)), \text{ and} \\ \mathbb{P}(X_1 \le 1, X_2 \le 0, X_3 \le 0) = C(1, \mathbb{P}(X_2 \le 0), \mathbb{P}(X_3 \le 1)), \end{cases} \tag{2.13}$$

to model the dependence of (X_1, X_2, X_3). The second, third, and fourth equations of (2.13) correspond to the three bivariate margins of the copula. The first equation completes the information needed to construct a trivariate Bernoulli distribution.

Consider a normal copula C_R with a symmetric and positive definite matrix R of bivariate correlations:

$$R = \begin{pmatrix} 1 & \rho_{12} & \rho_{13} \\ \rho_{12} & 1 & \rho_{23} \\ \rho_{13} & \rho_{23} & 1 \end{pmatrix}.$$

It is easy to see why there might be a problem with the existence of a normal copula that realizes a given trivariate Bernoulli distribution. The three correlations in the correlation matrix are determined by the second, third, and fourth equations in (2.13), and each can be computed as in Section 2.1.1. These three correlations have to (i) form a positive definite matrix R and, if this is the case, (ii) satisfy the first equation in (2.13). In the following example, these problems are highlighted.

Example 2.1.6. Consider two trivariate Bernoulli distributions, both with margins $\mathbb{P}(X_1 = 0) = 0.4$, $\mathbb{P}(X_2 = 0) = 0.8$, and $\mathbb{P}(X_3 = 0) = 0.2$. The probabilities of the second distribution are presented in brackets.

- $\mathbb{P}(X_1 = 0, X_2 = 0, X_3 = 0) = 0.01$ (0.0100)
- $\mathbb{P}(X_1 = 0, X_2 = 0, X_3 = 1) = 0.36$ (0.3789)
- $\mathbb{P}(X_1 = 1, X_2 = 0, X_3 = 0) = 0.16$ (0.1855)
- $\mathbb{P}(X_1 = 1, X_2 = 0, X_3 = 1) = 0.27$ (0.2256)
- $\mathbb{P}(X_1 = 0, X_2 = 1, X_3 = 0) = 0.02$ (0.0011)
- $\mathbb{P}(X_1 = 0, X_2 = 1, X_3 = 1) = 0.01$ (0.0100)
- $\mathbb{P}(X_1 = 1, X_2 = 1, X_3 = 0) = 0.01$ (0.0034)
- $\mathbb{P}(X_1 = 1, X_2 = 1, X_3 = 1) = 0.16$ (0.1855)

The bivariate margins of these distributions are:

- $\mathbb{P}(X_1 = 0, X_2 = 0) = 0.37$ (0.3889)
- $\mathbb{P}(X_2 = 0, X_3 = 0) = 0.17$ (0.1955)
- $\mathbb{P}(X_1 = 0, X_3 = 0) = 0.03$ (0.0111)

Note that the bivariate margin of (X_1, X_2) of the first trivariate Bernoulli has already been discussed in Section 2.1.1.

From the last three equations in (2.13), we find $\rho_{12} = 0.4868, \rho_{13} = -0.4868, \rho_{23} = 0.1340$ for the first distribution and $\rho_{12} = 0.7, \rho_{13} = -0.7, \rho_{23} = 0.6$ for the second. The correlations obtained for the second distribution do not form a positive-definite matrix. While a positive-definite matrix is formed with the correlations of the first distribution, with this parameter $C_R(0.4, 0.8, 0.2) = 0.0298 \neq 0.01 = \mathbb{P}(X_1 = 0, X_2 = 0, X_3 = 0)$. This means that neither distribution can be recovered with the trivariate normal copula.

A Bernoulli distribution inherits some properties from its corresponding copula. However, it is not easy to specify conditions under which a Bernoulli distribution can be constructed with a given copula. Conditions under which the construction is possible can be found only in some special cases.

Proposition 2.1.3. *If the trivariate Bernoulli distribution* (Y_1, Y_2, Y_3) *obtained from a trivariate normal copula* C_R *has margins* $\mathbb{P}(Y_i = 0) = 0.5$ *for all* $i \in \{1, 2, 3\}$, *then the three-way interaction* u_{123} *is zero.*

Proof. Since the normal copula realizes the trivariate Bernoulli distribution of (Y_1, Y_2, Y_3) and $\forall y_1, y_2, y_3 \in \{0, 1\}$ when $\mathbb{P}(Y_i = 0) = 0.5$, the radial symmetry of the trivariate normal distribution implies:

$$p(y_1, y_2, y_3) = p(1 - y_1, 1 - y_2, 1 - y_3)$$

Therefore, in this case we obtain:

$$\begin{aligned} u_{123} &= \log\left(\frac{p(1,1,1)p(1,0,0)p(0,1,0)p(0,0,1)}{p(1,1,0)p(1,0,1)p(0,1,1)p(0,0,0)}\right) \\ &= \log\left(\frac{p(0,0,0)p(0,1,1)p(1,0,1)p(1,1,0)}{p(1,1,0)p(1,0,1)p(0,1,1)p(0,0,0)}\right) = 0. \end{aligned}$$

■

By the construction of the Bernoulli distribution from a copula, we can immediately obtain the following result.

Proposition 2.1.4. *If the margins of the Bernoulli distribution are* $\mathbb{P}(X_i = 0) = 0.5$ *for all* $i \in \{1,2,3\}$ *and the distribution has a radial symmetry, i.e.* $p(x_1,x_2,x_3) = p(1-x_1,1-x_2,1-x_3)$ *for* $x_i \in \{0,1\}$*, then* (X_1,X_2,X_3) *can be realized with a trivariate normal copula.*

Proof. Since the margins are equal to 0.5, writing $p(1,1,1)$ in terms of $p(0,0,0)$ leads to:

$$p(1,1,1) = -0.5+p(0,0,0)+p(0,0,1)+p(0,0,0)+p(0,1,0)+p(0,0,0)+p(1,0,0)-p(0,0,0).$$

Because of the radial symmetry, the above equation becomes:

$$p(0,0,0) = \frac{\mathbb{P}(X_1 \leq 0, X_2 \leq 0, X_3 \leq 1) + \mathbb{P}(X_1 \leq 0, X_2 \leq 1, X_3 \leq 0) + \mathbb{P}(X_1 \leq 1, X_2 \leq 0, X_3 \leq 0) - 0.5}{2}.$$

The numerator of the right-hand side of the above equation is determined from the second, third, and fourth equations of (2.13). Therefore, the first equation of (2.13) is automatically satisfied by the parameters obtained from the other three equations. ■

Proposition 2.1.5. *If the three-way interaction of a trivariate Bernoulli distribution* (X_1,X_2,X_3) *is zero and* $\mathbb{P}(X_i = 0) = 0.5$ *for all* $i \in \{1,2,3\}$*, then a trivariate normal copula is able to realize* (X_1,X_2,X_3)*.*

Proof. The proof can be found in the Appendix. ■

Proposition 2.1.5 implies that a zero three-way interaction does not guarantee the existence of a normal copula that corresponds to the given Bernoulli distribution.

Let the latent vector (U_1,U_2,U_3) be joined by a normal copula C_R, such that U_1 and U_3 are independent conditionally on U_2. This happens when the correlations in R are such that $\rho_{13} = \rho_{12} \cdot \rho_{23}$ (Whittaker [1990]). The example below shows that the conditional independence of latent variables does not translate to the corresponding Bernoulli random vector (Y_1,Y_2,Y_3).

Example 2.1.7. Let (Y_1,Y_2,Y_3) be a trivariate Bernoulli distribution realized by a bivariate normal copula C with parameters $\rho_{12} = 0.5$, $\rho_{13} = -0.5$, and $\rho_{23} = \rho_{12} \cdot \rho_{13} = -0.25$ and let $\mathbb{P}(Y_1 = 0) = 0.4$, $\mathbb{P}(Y_2 = 0) = 0.8$, and $\mathbb{P}(Y_3 = 0) = 0.2$. We have that U_2, U_3 are conditionally independent given U_1.

The variables $Y_2|Y_1$ and $Y_3|Y_1$ where (Y_1,Y_2,Y_3) is constructed with copula C are not conditionally independent. In fact, the u-terms of the distribution of (Y_1,Y_2,Y_3) are: $u_{123} = -0.1408 \neq 0$ and $u_{23} = -0.7485 \neq 0$.

Even when $\mathbb{P}(Y_i = 0) = 0.5$ for all $i \in \{1,2,3\}$, the conditional independence of the latent variables does not translate to conditional independence of the Y's. With the same correlation matrix as above and all margins equal to 0.5, the u-terms are: $u_{23} = -0.6491 \neq 0$ and $u_{123} = 0$.

Corollary 2.1.6. *From Proposition 2.1.5, if* X_1 *and* X_3 *are conditionally independent given* X_2 *and the margins are* $\mathbb{P}(X_i = 0) = 0.5$ *for all* $i \in \{1,2,3\}$*, then* (X_1,X_2,X_3) *can be recovered with a trivariate normal copula.*

2.1.3 The Non-Constant Conditional Copula-Vine Approach

In the previous subsection we investigated the existence of a normal copula for a given Bernoulli distribution and showed that, in general, it is not easy to give conditions which assert the existence or non-existence of the normal copula for a specified Bernoulli distribution.

In this subsection, we adapt the use of the copula-vine approach, as presented in subsection 1.2.2, to model a multivariate Bernoulli distribution. Panagiotelis et al. [2012] have investigated this approach for a general multivariate discrete distribution. Analogous to the density decomposition in equation (1.11), the joint probability mass function of $(X_1, \dots, X_n)$ can be decomposed as, for instance:

$$\begin{aligned}\mathbb{P}(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) = \mathbb{P}(X_1 = x_1) \cdot \mathbb{P}(X_2 = x_2 | X_1 = x_1) \cdot \dots \cdot \\ \mathbb{P}(X_n = x_n | X_1 = x_1, \dots, X_{n-1} = x_{n-1}).\end{aligned} \tag{2.14}$$

Following the same notation as in subsection 1.2.2, each term $\mathbb{P}(X_j = x_j | \mathbf{V})$ on the right-hand side of (2.14) can be rewritten as:

$$\mathbb{P}(X_j = x_j | \mathbf{V} = \mathbf{v}) = \frac{\mathbb{P}(X_j = x_j, X_i = x_i | \mathbf{V}_{\setminus i} = \mathbf{v}_{\setminus i})}{\mathbb{P}(X_i = x_i | \mathbf{V}_{\setminus i} = \mathbf{v}_{\setminus i})}. \tag{2.15}$$

This is the discrete analogue to the density representation in equation(1.12). Again, just as in the fully continuous case, the numerator of the right-hand side of (2.15) can be rewritten and then expressed with the bivariate copula:

$$\mathbb{P}(X_j = x_j, X_i = x_i | \mathbf{V}_{\setminus i} = \mathbf{v}_{\setminus i}) = \sum_{s_j=0}^{x_j} \sum_{s_i=0}^{x_i} (-1)^{(s_j+s_i)} \mathbb{P}(X_j \le x_j - s_j, X_i \le x_i - s_i | \mathbf{V}_{\setminus i} = \mathbf{v}_{\setminus i}) \tag{2.16}$$

$$= \sum_{s_j=0}^{x_j} \sum_{s_i=0}^{x_i} (-1)^{(s_j+s_i)} C_{X_j, X_i | \mathbf{V}_{\setminus i}} \Big(\mathbb{P}(X_j \le x_j - s_j | \mathbf{V}_{\setminus i} = \mathbf{v}_{\setminus i}), \mathbb{P}(X_i \le x_i - s_i | \mathbf{V}_{\setminus i} = \mathbf{v}_{\setminus i}) \Big).$$

For $i \neq k < j$, the arguments of the copula on the right-hand side of the above expression can also be expressed with the copula:

$$\mathbb{P}(X_j \le x_j - s_j | \mathbf{V}_{\setminus i} = \mathbf{v}_{\setminus i}) = \frac{\sum_{s_k=0}^{x_k} (-1)^{s_k} C_{X_j, X_k | \mathbf{V}_{\setminus i,k}} \Big(\mathbb{P}(X_j \le x_j - s_j | \mathbf{V}_{\setminus i,k}), \mathbb{P}(X_k \le x_k - s_k | \mathbf{V}_{\setminus i,k}) \Big)}{\mathbb{P}(X_k = x_k | \mathbf{V}_{\setminus i,k})}. \tag{2.17}$$

Note that the decomposition above is performed sequentially until the conditioning set $\mathbf{V}$ is empty. Naturally, as in the fully continuous case, all (conditional) copulas appearing in this decomposition can be organized using a regular vine (Kurowicka and Cooke [2006]).

In Panagiotelis et al. [2012], the assumption has been made that all conditional copulas do not depend on the conditioning variables. In general, however, the conditional copulas do not have to be constant so the following conditions have been proposed in Panagiotelis et al. [2012] to ensure the existence of constant conditional copulas for the multivariate Bernoulli case.

Proposition 2.1.7. *Let* $p_{j,(1)},\dots,p_{j,(\kappa_1)}$ *denote the ordered* κ_1 *distinct values of* $\mathbb{P}(X_j \leq 0|\boldsymbol{V}_{\setminus i} = \boldsymbol{v}_{\setminus i})$ *and* $p_{i,(1)},\dots,p_{i,(\kappa_2)}$ *denote the ordered* κ_2 *distinct values of* $\mathbb{P}(X_i \leq 0|\boldsymbol{V}_{\setminus i} = \boldsymbol{v}_{\setminus i})$*. A constant bivariate copula* C *exists over the conditioning set* $\boldsymbol{V}_{\setminus i} = \boldsymbol{v}_{\setminus i}$ *if it solves*

$$\mathbb{P}(X_j \leq 0, X_i \leq 0|\boldsymbol{V}_{\setminus i} = \boldsymbol{v}_{\setminus i}) = C\left(\mathbb{P}(X_j \leq 0|\boldsymbol{V}_{\setminus i} = \boldsymbol{v}_{\setminus i}), \mathbb{P}(X_i \leq 0|\boldsymbol{V}_{\setminus i} = \boldsymbol{v}_{\setminus i})\right)$$

for each member in the conditioning set. For this to happen, all of the $(\kappa_1+1)(\kappa_2+1)$ *values of* $\mathbb{P}(p_{j,(a)}, p_{i,(b)}) - \mathbb{P}(p_{j,(a-1)}, p_{i,(b)}) - \mathbb{P}(p_{j,(a)}, p_{i,(b-1)}) + \mathbb{P}(p_{j,(a-1)}, p_{i,(b-1)})$ *must be non-negative.*

Even if a constant conditional copula exists for the above construction, this does not have to be a normal copula. To the best of the authors' knowledge, there is no result ensuring when the constant normal copula exists.

For large models the assumption of constant conditional copulas is understandable. In the case of moderate-sized models with variables that do not contain many states, it might not be prohibitive to consider the non-constant conditional copula-vine model. In such cases, different copulas can be specified for each combination of conditioning variables in (2.16).

Theorem 2.1.8. *Any multivariate Bernoulli random variables can be represented with the bivariate normal copulas with the non-constant conditional copula-vine model.*

Proof. Since the conditional distribution of a multivariate Bernoulli distribution is also multivariate Bernoulli and the copulas are allowed to be different, the result follows immediately from Theorem 2.1.1. ■

To illustrate how the non-constant conditional copula-vine model works, a simple example on the trivariate Bernoulli variable is presented.

Example 2.1.8. Let (X_1, X_2, X_3) be a trivariate Bernoulli distribution and let the vine structure in Example 1.13 in Chapter 1 as shown in Figure 1.12 be chosen. This means that two bivariate unconditional copulas, C_{12} and C_{23}, are fixed. Next, conditional copulas need to be specified for both realizations of variable X_2. These are denoted as $C_{13|2=0}$ and $C_{13|2=1}$.

The parameters of four bivariate normal copulas have to be such that equations (2.13) are all satisfied. The parameters ρ_{12} and ρ_{23} of the normal copula C_{12} and C_{23} are found from the second and fourth equation of (2.13), respectively. We find the parameter of copula $C_{13|2=0}$ to satisfy the first equation and $C_{13|2=1}$ to the fourth equation as follows:

$$\begin{aligned}\mathbb{P}(X_1 \leq 0, X_2 \leq 0, X_3 \leq 0) =& \mathbb{P}(X_1 \leq 0, X_2 = 0, X_3 \leq 0)\\ =& \mathbb{P}(X_1 \leq 0, X_3 \leq 0|X_2 = 0)\mathbb{P}(X_2 = 0)\\ =& C_{13|2=0}(\mathbb{P}(X_1 \leq 0|X_2 = 0), \mathbb{P}(X_3 \leq 0|X_2 = 0))\mathbb{P}(X_2 = 0)\end{aligned}$$

and

$$\begin{aligned}\mathbb{P}(X_1 \le 0, X_2 \le 1, X_3 \le 0) =& \mathbb{P}(X_1 \le 0, X_2 = 1, X_3 \le 0) + \mathbb{P}(X_1 \le 0, X_2 = 0, X_3 \le 0)\\ =& C_{13|2=1}(\mathbb{P}(X_1 \le 0|X_2 = 1), \mathbb{P}(X_3 \le 0|X_2 = 1))\mathbb{P}(X_2 = 1) +\\ & \mathbb{P}(X_1 \le 0, X_2 = 0, X_3 \le 0)\end{aligned}$$

Since there is no constraint on copulas in the above representation, any copula can be chosen and all four equations in (2.13) are satisfied. These conditional copulas correspond to $\mathbb{P}(X_1 \le 0, X_3 \le 0|X_2 = 0)$ and $\mathbb{P}(X_1 \le 0, X_3 \le 0|X_2 = 1)$; both are bivariate Bernoulli distributions.

For the first Bernoulli distribution in Example 2.1.6, the parameters of the normal copulas can be calculated as above and are equal to: $\rho_{12} = 0.4868$, $\rho_{23} = 0.1340$, $\rho_{13|2=0} = -0.7612$, and $\rho_{13|2=1} = 0.8552$. We see that the normal copula parameters are very different for different realization of variable X_2.

Moreover, $(X_1, X_3|X_2)$ cannot be represented with any constant conditional copula. In this example, $\mathbb{P}(X_1 = 0|X_2 = 0) = 0.4625 > \mathbb{P}(X_1 = 0|X_2 = 1) = 0.1500$, $\mathbb{P}(X_3 = 0|X_2 = 0) = 0.2125 > \mathbb{P}(X_3 = 0|X_2 = 1) = 0.1500$, and $\mathbb{P}(X_1 = 0, X_3 = 0|X_2 = 0) = 0.0125 < \mathbb{P}(X_1 = 0, X_3 = 0|X_2 = 1) = 0.1000$. This results in the probability in the region $([0, \mathbb{P}(X_1 = 0|X_2 = 0)] \times [0, \mathbb{P}(X_3 = 0|X_2 = 0)]) \setminus ([0, \mathbb{P}(X_1 = 0|X_2 = 1)] \times [0, \mathbb{P}(X_3 = 0|X_2 = 1)])$ being negative, which is a violation of the condition in Proposition 2.1.7.

For a joint normal copula, each marginal and conditional copula is normal and the conditional copulas do not depend on the conditioning variables. This property does not translate to the trivariate Bernoulli distribution (Y_1, Y_2, Y_3) implied by the normal copula. It is not always the case that the conditional copulas $C_{13|2=0}$ and $C_{13|2=1}$ are equal.

Proposition 2.1.9. *Let the univariate margins of a trivariate Bernoulli distribution* (X_1, X_2, X_3) *be* $\mathbb{P}(X_i = 0) = 0.5$ *for all* $i \in \{1, 2, 3\}$. *The trivariate normal copula realizes* (X_1, X_2, X_3) *if and only if* $C_{ij|k=0} = C_{ij|k=1}$ *for any combination of* $i, j, k \in \{1, 2, 3\}$ *where* $C_{ij|k}$ *is a radially symmetric copula.*

Proof. The proof can be found in the Appendix. ■

According to Proposition 2.1.9, if $C_{13|2=0}$ and $C_{13|2=1}$ are the independent copulas and the margins are 0.5, then the distribution of (Y_1, Y_2, Y_3) can be represented by the normal copula and the variables Y_1 and Y_3 are conditionally independent given variable Y_2. However, the latent variables U_{Y_1} and U_{Y_3} are *not* conditionally independent given U_{Y_2}. This is illustrated in the following example.

Example 2.1.9. Let (Y_1, Y_2, Y_3) be a trivariate Bernoulli distribution with $\mathbb{P}(Y_i = 0) = 0.5$ for all $i \in \{1, 2, 3\}$ and both $C_{13|2=0}$ and $C_{13|2=1}$ be the independent copulas. Assume that the bivariate margins (Y_1, Y_2) and (Y_2, Y_3) are represented by the bivariate normal copula with parameters 0.5 and −0.5, respectively. Y_1 and Y_3 are, thus, conditionally independent given Y_2 and the trivariate normal copula with parameters $r_{12} = 0.5$, $r_{23} = -0.5$, and $r_{13} = -0.1736$ represents the trivariate Bernoulli distribution (Y_1, Y_2, Y_3). However, $r_{12} \cdot r_{23} = -0.25 \ne -0.1736 = r_{13}$, which means that the latent variables U_{Y_1} and U_{Y_3} are not conditionally independent given U_{Y_2}.

Any multivariate Bernoulli distribution can always be recovered using the non-constant conditional copula-vine approach. This means that it can be constructed using building blocks consisting of bivariate normal copulas. This is not the case, however, when the conditioning copulas in the copula-vine approach are assumed not to depend on the conditioning variables, nor in the case of multivariate normal copula.

2.2 Mixed Discrete-Continuous Distributions with Copulas

In this section, we discuss the extension of the copula modelling to discrete distributions with more than two states and mixed discrete-continuous models. In cases where the variables have more states, a copula used in the construction of such a discrete distribution must satisfy more constraints.

2.2.1 Bivariate Case

We start the exposition with the simplest possible case: a bivariate discrete distribution (X_1, X_2) with one margin taking values on $\{0,1\}$ and another taking values on $\{0,1,2\}$. In this case, a copula C that realizes (X_1, X_2) must satisfy the following conditions:

$$\begin{cases} \mathbb{P}(X_1 \le 0, X_2 \le 0) = C(\mathbb{P}(X_1 \le 0), \mathbb{P}(X_2 \le 0)), \\ \mathbb{P}(X_1 \le 0, X_2 \le 1) = C(\mathbb{P}(X_1 \le 0), \mathbb{P}(X_2 \le 1)). \end{cases} \tag{2.18}$$

The normal copula cannot always represent (X_1, X_2) anymore because of the over-determined system (2.18) that needs to be satisfied.

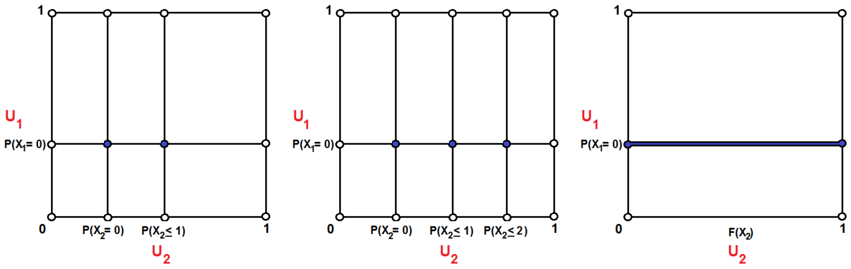
With more states of the variables, the problem deteriorates simply because the number of equations in (2.18) increases while the number of parameters remains the same. When X_2 is continuous, it has an infinite number of states and a copula C that is able to recover the distribution of (X_1, X_2) needs to satisfy the following constraint for all realizations of X_2:

$$\mathbb{P}(X_1 \le 0, X_2 \le x_2) = C(\mathbb{P}(X_1 \le 0), \mathbb{P}(X_2 \le x_2)). \tag{2.19}$$

Figure 2.7 illustrates the problems of finding a copula graphically for distributions with more than two states. Figure 2.7(a) is when X_2 has three states, Figure 2.7(b) is when X_2 has four states, and Figure 2.7(c) is when X_2 is continuous. The blue dots in Figure 2.7(a) and 2.7(b) show where in the unit square the copula is constrained. When X_2 is continuous, the copula is constrained at all points on the horizontal blue line in Figure 2.7(c).

Using (2.19), to see whether the copula C can model the mixed discrete-continuous bivariate variable (X_1, X_2), the conditional distribution of X_2 given

Figure 2.7: *The unit square corresponding to the latent variable* (U_1, U_2) *for distributions of* X_2 *with different numbers of states.*



(a) X_2 with three states. (b) X_2 with four states. (c) X_2 continuous.

X_1 should be compared with the conditional distribution of a copula:

$$\mathbb{P}(X_2 \leq x_2 | X_1 \leq 0) = \frac{C(\mathbb{P}(X_1 \leq 0), \mathbb{P}(X_2 \leq x_2))}{\mathbb{P}(X_1 \leq 0)}. \tag{2.20}$$

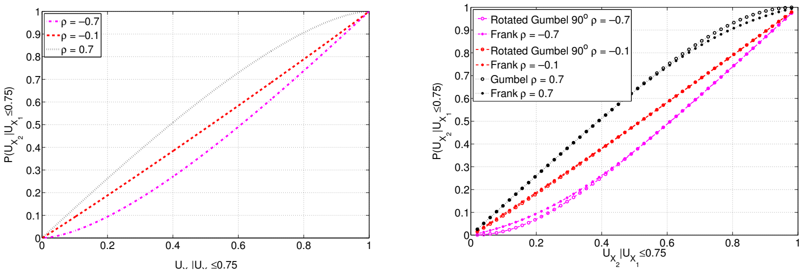
The following example illustrates this.

Example 2.2.1. Let (X_1, X_2) be a mixed discrete-continuous bivariate random variable with X_1 binary, $\mathbb{P}(X_1 = 0) = 0.75$, and X_2 continuous with marginal distribution $\mathbb{P}(X_2 \leq x_2)$. Figure 2.8(a) shows the conditional distribution of the latent variable U_2 given $U_1 \leq 0.75$ for normal copulas with different parameters. Figure 2.8(b) illustrates the conditional distributions as in (a) but with Frank and Gumbel copulas used to model dependence between U_2 and U_1.

When the variable X_1 is also taken to be continuous, the conditions on the copula become:

$$\mathbb{P}(X_1 \leq x_1, X_2 \leq x_2) = C(\mathbb{P}(X_1 \leq x_1), \mathbb{P}(X_2 \leq x_2)),$$

Figure 2.8: *The conditional distribution of the latent variable* U_2 *given variable* $U_1 \leq 0.75$ *with different copulas.*



(a) Three different $U_2|U_1 \leq 0.75$ from three normal copulas. (b) Six different $U_2|U_1 \leq 0.75$ from three Frank copulas and three Gumbel copulas.

which is none other than Sklar's equation (1.2). In this case, the copula must conform to all points in the unit square of the latent variable (U_1, U_2).

2.2.2 Multivariate Case

In the higher-dimensional case, it becomes even more difficult to find the copula that generates the mixed model. The non-constant conditional copula-vine approach as presented in Section 2.1.3 can be seen as the most flexible model in this case. The conditional probabilities in (2.16) containing only discrete variables in the conditioning set have to be replaced by the following:

$$\begin{aligned}
&\mathbb{P}(X_j = x_j, X_i \leq x_i | \mathbf{V}_{\setminus i} = \mathbf{v}_{\setminus i}) = && (2.21)\\
&= \sum_{s_j=0}^{x_j} (-1)^{s_j} \mathbb{P}(X_j \leq x_j - s_j, X_i \leq x_i | \mathbf{V}_{\setminus i} = \mathbf{v}_{\setminus i})\\
&= \sum_{s_j=0}^{x_j} (-1)^{s_j} C_{X_j, X_i | \mathbf{V}_{\setminus i}} \left(\mathbb{P}(X_j \leq x_j - s_j | \mathbf{V}_{\setminus i} = \mathbf{v}_{\setminus i}), \mathbb{P}(X_i \leq x_i | \mathbf{V}_{\setminus i} = \mathbf{v}_{\setminus i}) \right) && (2.22)
\end{aligned}$$

where X_j corresponds to the binary variable, X_i corresponds to the continuous variable, and $\mathbf{V}_{\setminus i}$ contains only discrete variables. In this case, the approach as illustrated in Example 2.2.1 can be used for each combination of conditioning variables in (2.22).

2.3 Modelling Mixed Data with the Copula-Vine Approach

In the previous section we have seen how the copula-vine approach can be used to represent the joint distribution of mixed discrete-continuous variables. In this section we discuss how that approach can be applied to construct a joint model of a dataset with mixed discrete and continuous variables.

To apply the copula-vine approach, a regular vine structure first needs to be chosen. With n variables, there are $\frac{n!}{2} 2^{\left(\frac{(n-2)(n-3)}{2}\right)}$ possible regular vine structures (Morales Napoles [2009]). This is due to the many possible decompositions of the joint distribution which means that, even with only a moderate number of variables in the model, the number of possible regular vines is already very large. For instance, when $n = 10$, there are $487,049,291,366,400$ possible regular vines.

To narrow down the set of possible regular vine structures, in this thesis we choose to keep the purely continuous and purely discrete parts of the model as sub-vines of the full vine. Let n_d denote the number of discrete variables and $n_c = n - n_d$ the number of continuous variables. Even after clustering the continuous and discrete variables (each with an already chosen regular vine structure), there are still $2^{n_d + n_c - 2} = 2^{n-2}$ different ways to merge these sub-vines together

(Cooke et al. [2015]). Unless otherwise stated, in this thesis we merge the two sub-vines by connecting the discrete and continuous nodes with the highest correlation.

Next, bivariate copula families corresponding to the edges in the vine need to be chosen. In principle, any such family can be fitted. In this thesis, we focus on the bivariate normal copula because this is the one we intend to use later on in the railway disruption-length models.

Once the vine structure and the bivariate copulas have been chosen, the model's parameters are estimated from data to obtain the full copula-vine model.

The likelihood of the constructed model is computed using the same decomposition as in the purely continuous (equation (1.11) - (1.12)) and purely discrete (equation (2.14) - (2.15)) cases, following the chosen vine structure and evaluating each term in the decomposition as a bivariate (conditional) copula. The likelihood, therefore, takes the form of a product of these bivariate (conditional) copulas evaluated at the data points.

There are three possible cases for these bivariate (conditional) copulas between the variables X_j and X_i:

1. When both X_j and X_i are discrete, we need to compute $\mathbb{P}(X_j = x_j, X_i = x_i|\mathbf{V}_{\setminus i} = \mathbf{v}_{\setminus i})$. This can be done as in equation (2.16).
2. When both X_j and X_i are continuous, $f_{X_j|\mathbf{V}_{\setminus i}, X_i|\mathbf{V}_{\setminus i}}(X_j = x_j, X_i = x_i|\mathbf{V}_{\setminus i} = \mathbf{v}_{\setminus i})$ can be computed as in equation (1.3).
3. When one of X_j or X_i is discrete and the other is continuous. If the discrete variable is X_j, then the likelihood can be computed by multiplying the probability $\mathbb{P}(X_j = x_j|X_i = x_i, \mathbf{V}_{\setminus i} = \mathbf{v}_{\setminus i})f(X_i = x_i|\mathbf{V}_{\setminus i} = \mathbf{v}_{\setminus i})$ of each sample. The first term can be computed with the bivariate copula as follows

$$= \sum_{s_j=0}^{x_j} (-1)^{s_j} C_{X_j|\mathbf{V}_{\setminus i}=\mathbf{v}_{\setminus i} \big| F(X_i \leq x_i|\mathbf{V}_{\setminus i}=\mathbf{v}_{\setminus i})}(\mathbb{P}(X_j \leq x_j - s_j|\mathbf{V}_{\setminus i} = \mathbf{v}_{\setminus i})) \tag{2.23}$$

   while the second term can be rewritten as another bivariate (conditional) copula following the sub-vine corresponding to the set $\{X_i, \mathbf{V}_{\setminus i}\}$.

The copula arguments, which are the conditional distribution of X_j given the variables in $\mathbf{V}_{\setminus i}$, are expressed with bivariate copulas following the corresponding sub-vine structure. Following the notation in equation (2.17), where X_k is an element of $\mathbf{V}_{\setminus i}$, there are four possible forms of the copula argument:

1. When both X_j and X_k are discrete, then $\mathbb{P}(X_j \leq x_j|\mathbf{V}_{\setminus i,k}, X_k)$ can be computed as in equation (2.17).
2. When X_j is discrete and X_k is continuous, then:

$$\mathbb{P}(X_j \leq x_j|\mathbf{V}_{\setminus i,k}, X_k = x_k) = C_{X_j|\mathbf{V}_{\setminus i,k} \big| F(X_k \leq x_k|\mathbf{V}_{\setminus i,k})}(\mathbb{P}(X_j \leq x_j|\mathbf{V}_{\setminus i,k})) \tag{2.24}$$

   as in equation (2.23).

3. When X_j is continuous and X_k is discrete, then:

$$F(X_j \leq x_j|\mathbf{V}_{\setminus i,k}, X_k = x_k) = \frac{\sum_{s_k=0}^{x_k}(-1)^{s_k} C_{X_j,X_k|\mathbf{V}_{\setminus i,k}}(F(X_j \leq x_j|\mathbf{V}_{\setminus i,k}), \mathbb{P}(X_k \leq x_k - s_k|\mathbf{V}_{\setminus i,k}))}{\mathbb{P}(X_k = x_k|\mathbf{V}_{\setminus i,k})} \tag{2.25}$$

4. When both X_j and X_i are continuous, then:

$$F(X_j \leq x_j|\mathbf{V}_{\setminus i,k}, X_k = x_k) = C_{X_j|\mathbf{V}_{\setminus i,k}\big|F(X_k \leq x_k|\mathbf{V}_{\setminus i,k})}(F(X_j \leq x_j|\mathbf{V}_{\setminus i,k})) \tag{2.26}$$

as in the fully continuous case.

The decomposition is performed sequentially until the conditioning set $\mathbf{V}$ is empty.

In a three dimensional case, the likelihood of the mixed discrete-continuous data presented in Example 1.2.9 can be computed by taking the products of the joint distribution representation in equation (1.16) computed for each sample.

2.3.1 An Algorithm to Obtain a Parsimonious Non-Constant Conditional Copula-Vine Model

As mentioned in subsection 1.2.4, a copula-vine model (especially when non-constant conditional copulas are considered) requires a high number of parameters resulting in a not parsimonious model. Some of these parameters might not be significant and should not be included in the model.

Algorithm 2.1 is proposed for this purpose. Given a regular vine structure and a set of bivariate copulas, this algorithm estimates the copula parameter values from data. For each parameter value, the confidence bound is computed through parametric bootstrapping and if the parameter is found to be insignificant, it is removed from the model.

Note that Algorithm 2.1 only considers non-constant conditional copulas on the pairs whose conditioning sets contain only discrete variables. In principle, non-constant conditional copulas could be assumed on any pair with a non-empty conditioning set.

2.3.2 Testing Algorithm 2.1

To test whether Algorithm 2.1 works, a small simulation study is performed. Consider three binary discrete variables, X_1, X_2, and X_3, with margins $\mathbb{P}(X_1 = 0) = 0.6, \mathbb{P}(X_2 = 0) = 0.4$, and $\mathbb{P}(X_3 = 0) = 0.7$, respectively, and three continuous variables X_4, X_5, and X_6 uniform on $(0,1)$. The six variables are joined together with a vine with the structure depicted in Figure 2.9. Such a structure is a special case of a regular vine called the *D-vine*.

The three discrete variables are shown as white rectangles and the three continuous variables as yellow rectangles. The edges in different trees are represented with different line styles. The red edges of the D-vine correspond to the

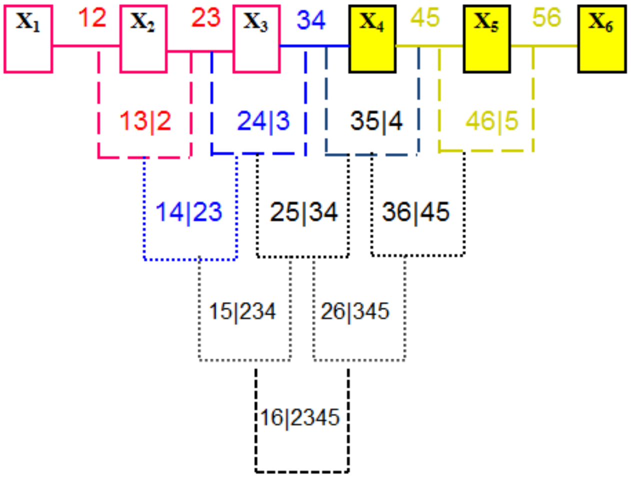
Algorithm 2.1 An algorithm to obtain a parsimonious non-constant copula-vine model

```
for tree = 1 to $n-1$ do
  if tree = 1 then
    for each pair of nodes do
      1. Estimate the parameter of the unconditional copula.
      2. Re-sample the pair $M$ times and find the parameter's 95% confidence bound.
      3. If the confidence bound contains 0, set the parameter to 0.
    end for
  else
    for each pair of nodes do
      if all conditioning variables are discrete then
        for each combination of conditioning variables values do
          1. Estimate the parameter of the unconditional copula.
          2. Re-sample the pair $M$ times and find the parameter's 95% confidence bound.
          3. If the confidence bound contains 0, set the parameter to 0.
        end for
        if all confidence bounds overlap then
          4. Fit a constant conditional copula to the pair.
          5. Re-sample the pair $M$ times and find the parameter's 95% confidence bound.
          6. If the confidence bound contains 0, set the constant parameter to 0.
        end if
      else
        1. Fit a constant conditional copula.
        2. Re-sample the pair $M$ times and find the parameter's 95% confidence bound.
        3. If the confidence bound contains 0, set the parameter to 0.
      end if
    end for
  end if
end for
```

purely discrete part of the model, the yellow lines correspond to the purely continuous part, and the rest represent the mixed pairs. The edges of the vine indicate the conditioned|conditioning variables and they correspond to the bivariate (conditional) copulas needed in the model. The red and blue edges in the second and third trees correspond to the pairs where non-constant conditional copulas can be applied. The pairs connected with the yellow and black edges are modelled with constant conditional copulas.

A set of bivariate normal copulas with certain parameter values is chosen and is used to generate datasets containing N samples from the D-vine structure. In

Figure 2.9: *A regular vine structure on six variables.*



this thesis, we consider $N = \{100, 500, 1000, 2000\}$. Algorithm 2.1 with $M = 1000$ is used to find the confidence bounds of the parameters of the bivariate normal copulas based on these samples.

To see how the copula-vine models recover the data, for each N we observe the following:

1. The log-likelihood.
2. The KL test for the eight-cell contingency table between the observed and predicted frequencies of the three binary variables.
3. The PIT test to test the trivariate normal copula for the three continuous variables.

The KL test and PIT test are goodness-of-fit tests of the discrete and continuous parts of the model, respectively. As for the mixed part of the model, to author's knowledge there is no usable goodness-of-fit test. We therefore test the model by its performance in predicting the outcome of some variables given observation of the others.

In practice, the variable of interest could be either the continuous variable, the discrete variable, or a set of variables in the model. In the experiment, in assessing the model's performance we consider only the prediction of either a continuous or a discrete variable. This choice is motivated by the application part of this thesis, where the railway disruption-length model is used to predict the disruption length.

Let the continuous variable X_6 be of interest. To test the fit of the conditional distribution of X_6, for each data point the quantile of the conditional distribution corresponding to the observed $X_6 = x_6$ is computed. The model represents the data well if these quantiles are distributed uniformly on $(0,1)$. This can be measured by, e.g., the Kolmogorov-Smirnov (KS) test.

A natural choice for the prediction would be the mean of the conditional distribution of X_6, denoted as $\hat{x}_6$. The root mean square error (RMSE) of the prediction is defined as:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{x}_{6i} - x_{6i})^2} \tag{2.27}$$

where x_{6i} denotes the i-th realization of X_6, and $\hat{x}_{6i}$ denotes the mean of the i-th conditional distribution of X_6. The closer the RMSE is to zero, the more accurate the prediction is.

The coefficient of determination R^2 is defined as

$$R^2 = 1 - \frac{RMSE^2}{\mathrm{Var}(X_6)}. \tag{2.28}$$

R^2 indicates the proportion of the variance of X_6 that is explained by the variables $X_1, \ldots, X_5$ through the copula-vine model.

If the discrete variable X_1 is of interest, the conditional distribution of X_1, $\mathbb{P}(X_1 = 0|X_2, \ldots, X_6)$, is computed from the copula-vine model. To make a prediction of X_1, a *threshold* needs to be chosen for $\mathbb{P}(X_1 = 0|X_2, \ldots, X_6)$. If the computed $\mathbb{P}(X_1 = 0|X_2, \ldots, X_6)$ is lower than the threshold, i.e. the probability of observing zero is "small", then the model's prediction of X_1 is $\hat{X}_1 = 1$. Otherwise, the prediction is $\hat{X}_1 = 0$. The threshold is chosen such that it maximizes, for instance, the proportion of correct prediction, i.e.

$$\mathbb{P}(\text{Correct Prediction}) = \mathbb{P}(X_1 = 0, \hat{X}_1 = 0) + \mathbb{P}(X_1 = 1, \hat{X}_1 = 1) \tag{2.29}$$

for the given threshold.

Because X_1 is binary, a popular way to make a prediction of X_1 is by using the generalized linear model (GLM) with the logit link (see, e.g., Nelder and Wedderburn [1972]). We shall also compare the performance of the predictions of X_1 made with the copula-vine model and the GLM. The built-in function `fitglm` in `MATLAB` is used to obtain the GLM. We consider up to two-way interactions, and parameters found to be insignificant are removed from the GLM.

Therefore, additionally to statistics 1, 2, and 3 above, we look at the following:

4. The fit of the conditional distribution of X_6 by means of the KS test.
5. The *RMSE* and R^2 values of the prediction of X_6.
6. The threshold which yields the maximum proportion of correct predictions of X_1 and the proportion itself.
7. A comparison with a GLM with logit link to predict X_1.

Table 2.1: *A summary of the copula-vine models' performance for different numbers of samples N.*

N		100	500	1000	2000
1. Log-likelihood		−116.88	−517.47	−1005.15	−2015.80
2. Discrete fit		0.2025	0.6604	0.6667	0.7500
3. Continuous fit		0.3893	0.4427	0.3716	0.2376
Continuous prediction	4. Cond. dist.	0.5877	0.2713	0.8900	0.4085
	5a. $RMSE$	0.1517	0.1284	0.1286	0.1250
	5b. R^2	0.7575	0.8186	0.8139	0.8112
Discrete prediction	6a. Threshold	0.5	0.56	0.5	0.53
	6b. Proportion	77%	77.20%	75.30%	74.35%
	7. GLM	77.00%	78.80%	77.10%	77.20%

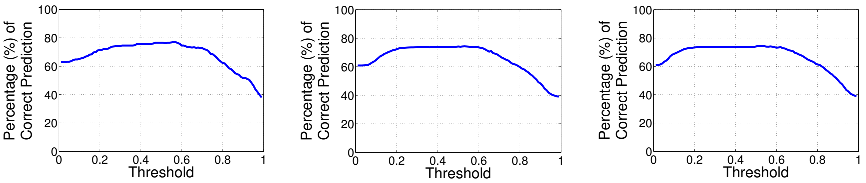
Table 2.1 summarizes the results.

The fitted copula parameters of the four models can be found in Table B.1 in Appendix B. It is shown there that the true parameter values are captured by the confidence bounds of all models. As would be expected, the confidence bounds are the widest when $N = 100$ and narrow down as N grows. Consequently, the model in which $N = 100$ has the most parameters set to zero.

Table 2.1 shows that the copula-vine models with parameters estimated using Algorithm 2.1 perform well in representing the data with different numbers of samples. The discrete part of the model is recovered well, as indicated by the p-values of the KL test. The PIT test indicates that the continuous part is also well-recovered[2].

To predict the continuous variable X_6, the model appears to benefit from a higher number of samples, as seen from the generally decreasing behaviour of the $RMSE$.

Figure 2.10: *Proportion of correct prediction as a function of threshold value.*



(a) $N = 500$, sequential. (b) $N = 2000$, sequential. (c) $N = 2000$, full-optimized.

Figure 2.10(a) and Figure 2.10(b) show the proportion of correct predictions (in %) of the variable X_1 as a function of threshold value for $N = 500$ and

---

[2] The PIT Test rejects the null-hypothesis that the copula recovers the dependence between the three continuous variables if any of the p-values evaluated from equation (1.18) or (1.19) is below the significance level of 5%. It is therefore sufficient to show the smallest evaluated p-value to draw a conclusion from the test, as presented in the table.

$N = 2000$, respectively. The results in Table 2.1 for discrete prediction are obtained from these figures by choosing the threshold that yields the maximum proportion for each N. In predicting the discrete variable X_1, however, the model does not appear to benefit significantly from a higher number of samples. Moreover, in terms of predicting X_1, the performances of the copula-vine and the GLM are similar for all N.

Algorithm 2.1 estimates the parameters sequentially. The model can be further "improved" by performing a full optimization of all parameters in the model together by maximizing the model's likelihood. Only parameters that are not zero after the sequential fitting are considered in the full optimization.

To test whether the improvement is significant, a copula-vine model with $N = 2000$ whose parameters are "fully-optimized" is constructed and compared with the copula-vine model with sequentially-estimated parameters. The parameters of the fully-optimized model can be found in Table B.2 in Appendix B. Table 2.2 summarizes the comparison between the two models.

Table 2.2: *Comparing the copula-vine models with sequential parameter estimation and fully-optimized parameter estimation for $N = 2000$.*

Method		Sequential	Fully Optim.
1. Log-likelihood		−2015.80	−2002.78
2. Discrete fit		0.7500	0.6412
3. Continuous fit		0.2376	0.2357
Continuous prediction	4. Cond. dist.	0.4085	0.5502
	5a. $RMSE$	0.1250	0.1267
	5b. R^2	0.8112	0.8058
Discrete prediction	6a. Threshold	0.53	0.52
	6b. Proportion	74.35%	74.10%

While the log-likelihood of the fully-optimized model is larger, the difference in log-likelihood is not statistically significant, as we observe with the Vuong and Clarke's distribution-free tests (see subsection 1.2.6). The test-statistic Z of the Vuong test is 1.6166 that is well within the 95% confidence bound of $(-1.96, 1.96)$. The test statistic B of the Clarke's distribution-free test is 1025, well within the 95% confidence bound of $(956, 1044)$. Moreover, the fully-optimized model does not outperform the sequential model in the prediction of continuous or discrete variables.

Estimating the "fully-optimized" parameters is certainly computationally expensive. On top of approximately three hours of computation time needed to sequentially estimate the parameters of the $N = 2000$ copula-vine model with Algorithm 2.1, seven hours of computation are required to obtain the "fully-optimized" parameters. In this thesis, all computation is performed in a computer with an Intel(R) Core i5-3470 3.2 GHz processor and 8 GB RAM.

For these reasons, from hereon we consider only the "sequential" copula-vine model.

In the experiment thus far, the GLM has only been used to predict the binary variable X_1 because the continuous variables are uniform on $(0,1)$. A limitation of the GLM is that the conditional distribution of the dependent variable is assumed to come from a distribution in the exponential family. However, the uniform variables can be transformed into, say, the standard normal distributions by applying the inverse cumulative distribution function. In this case, we use the GLM with the identity link to predict the transformed continuous variable X_6. Table 2.3 presents the performance of the GLM on the four transformed datasets.

Table 2.3: *Summary of the GLM performance in predicting the transformed variable X_6.*

N		100	500	1000	2000
GLM's	4. Cond. dist.	10^{-4}	10^{-17}	10^{-34}	10^{-72}
continuous	5a. $RMSE$	0.1257	0.1248	0.1256	0.1235
prediction	5b. R^2	0.8295	0.8286	0.8225	0.8156

Table 2.3 shows that the GLM produces good predicted mean as indicated by the low $RMSE^3$, even for small number of samples. By contrast with the copula-vine models, however, the GLM does not recover the conditional distribution of X_6.

2.3.3 Misspecification in Modelling

In practice, wrong assumptions might be made during the construction of a copula-vine model. For instance, the assumption of constant conditional copulas might be imposed while, in reality, are non-constant. Moreover, both the underlying bivariate copula families and the vine structure are generally unknown. The best copula family can be chosen by comparing a few different families and choosing the one closest to the data, while the vine structure can be chosen through a number of heuristic procedures (see Kurowicka and Joe [2011]). In this subsection, we look at those situations in which a wrong constant conditional copula assumption is made and in which either the copula family or the vine structure is misspecified.

Example 2.3.1. ***(wrong constant copula assumption)***. Consider the previous datasets constructed in the example in Subsection 2.3.2 and generated from a non-constant Copula-Vine. From these, for each N a constant copula-vine model with the same vine-structure is constructed with a set of bivariate normal copulas representing all arcs of the vine. To do this, Algorithm 2.1 is run with $M = 1000$, but every pair of nodes in the second tree and above is modelled with a constant conditional copula. Table B.3 in Appendix B presents the fitted parameters.

Table 2.4 summarizes the constant models' fit and performance in prediction for different values of N.

---

[3] The RMSEs are computed in the original uniform scale to enable comparison with the copula-vine model's performance as presented in Table 2.1.

Table 2.4: *Summary of the constant copula-vine models' performance for different numbers of samples N when the truth is non-constant.*

N		100	500	1000	2000
1. Log-likelihood		−117.9361	−518.60	−1011.12	−2030.96
2. Discrete fit		0.0017	$< 10^{-17}$	$< 10^{-17}$	$< 10^{-17}$
3. Continuous fit		0.3893	0.4427	0.3716	0.2376
Continuous prediction	4. Cond. dist.	0.3724	0.1223	0.8609	0.3945
	5a. $RMSE$	0.1415	0.1294	0.1290	0.1257
	5b. R^2	0.7839	0.8157	0.8128	0.8088
Discrete prediction	6a. Threshold	0.50	0.48	0.47	0.47
	6b. Proportion	64%	74.20%	71.20%	71.55%
	7. GLM	76%	78.80%	77.40%	77.20%

It is unsurprising that the discrete part of the data is not recovered by any of the models. The continuous part is modelled exactly as in the non-constant copula-vine models in subsection 2.3.2 because the parameters are estimated sequentially.

The models predict the continuous variable X_6 well for all N where the quantiles of the conditional distribution of X_6 are uniform $(0,1)$ for any N. The $RMSE$ tends to decrease with higher N. However for $N = \{500, 1000, 2000\}$, the $RMSE$s are slightly higher than in the non-constant copula-vine models. Moreover, the proportions of correct predictions of the binary variable X_1 are lower than in the non-constant copula-vine models.

Because the non-constant and constant copula-vine models are constructed from the same datasets, the models' likelihoods can be compared to see which are the better ones for the data. The results are summarized in Table 2.5.

Table 2.5: *Comparing the likelihoods of the non-constant and constant copula-vine models.*

N		100	500	1000	2000
AIC	non-constant	255.76	1062.94	2042.30	4065.61
	constant	251.87	1061.19	2046.24	4087.93
BIC	non-constant	284.42	1121.94	2120.82	4160.82
	constant	272.71	1111.77	2105.13	4160.74
LRT (p-value)		0.5514	0.3237	0.0357	10^{-5}
Vuong test's Z (Conf. bound)		−1.6735 $(-1.96, 1.96)$	−0.5104 $(-1.96, 1.96)$	0.8483 $(-1.96, 1.96)$	2.2205 $(-1.96, 1.96)$
Clarke's test's B (Conf. bound)		26 $(40, 60)$	227 $(228, 272)$	467 $(469, 531)$	1118 $(956, 1044)$

For low numbers of samples ($N = \{100, 500\}$), all tests favour the constant copula-vine models with the exception of the Vuong test which cannot distinguish which models are better. For $N = 1000$, from the AIC and likelihood ratio test (LRT) we conclude that the non-constant copula-vine model is better. However, BIC and Clarke's test indicate otherwise. The Vuong test, again, cannot identify which model is better. For $N = 2000$, all tests but the BIC state that the non-constant copula-vine model is the better one.

In this example, the BIC always favours the constant copula-vine models even though we know that these "underfit" the truth. This observation is in line with findings in, e.g. Johnson and Omland [2004], Vrieze [2012], and Aho et al. [2014], where the BIC tends to choose the simpler model.

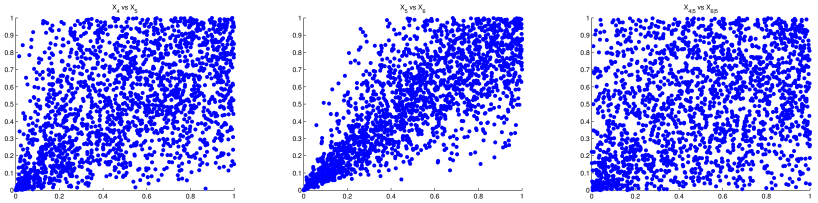
Example 2.3.1 shows that the choice of constant or non-constant conditional copula assumption affects the constructed copula-vine model. The models in Example 2.3.1 underestimate the complexity of the data, which results in lower likelihoods and in misfits in the discrete part of the models. In the example, however, this does not severely affect the models' predictive performance.

In the next example, we consider the situation when the copula family is misspecified.

Example 2.3.2. ***(copula misspecification)***. Consider three binary variables X_1, X_2, and X_3, with margins $\mathbb{P}(X_1 = 0) = 0.6, \mathbb{P}(X_2 = 0) = 0.4$, and $\mathbb{P}(X_3 = 0) = 0.7$, respectively, and three continuous variables X_4, X_5, and X_6 uniform on $(0,1)$. The six variables are joined together with a D-vine as in Figure 2.9.

The vine is sampled to obtain datasets containing N data points with a set of bivariate Clayton copulas with given parameters values. As before, we consider $N = \{100, 500, 1000, 2000\}$. From the datasets, for each N a non-constant copula-vine model with the same vine structure is constructed with a set of bivariate normal copulas representing all arcs of the vine. To do this, Algorithm 2.1 is run with $M = 1000$.

Figure 2.11: *Scatter plots of the continuous variables in the generated data.*



(a) X_4 vs X_5.

(b) X_5 vs X_6.

(c) $X_4|X_5$ vs $X_6|X_5$.

Figure 2.11 presents the scatter plots of three continuous bivariate margins for the data set with $N = 2000$. It can be seen that none of the copula is bivariate normal.

Table B.4 in Appendix B presents the fitted parameters. It is shown there that the fitted normal copula-vine models capture the true correlation values of the parameters. Table 2.6 summarizes the models' fit and performance in prediction for different values of N.

Table 2.6 shows that the discrete part of the data is recovered well for all N. This is an implication of the non-uniqueness of copulas in the fully discrete

Table 2.6: *Summary of the copula-vine models' performance for different numbers of samples N when the copula family is misspecified.*

N		100	500	1000	2000
1. Log-likelihood		−103.37	−621.18	−1183.33	−2376.99
2. Discrete fit		0.3464	0.5830	0.6885	0.5573
3. Continuous fit		0.0219	0.0037	0.0087	$7.7275e-4$
Continuous prediction	4. Cond. dist.	0.0084	0.0106	0.0013	$1.6285e-10$
	5a. $RMSE$	0.1844	0.1737	0.1686	0.1676
	5b. R^2	0.5485	0.6120	0.6426	0.6496
Discrete prediction	6a. Threshold	0.58	0.41	0.46	0.46
	6b. Proportion	86%	90.60%	88.60%	87.30%
	7. GLM	66%	91.00%	89.60%	88.75%

models and because all the discrete variables are binary. Even though the data is generated with bivariate Clayton copulas, copula-vine models with bivariate normal copulas can still be used to represent the discrete data.

As we would expect, the continuous part is not recovered well as indicated by the low p-values of the PIT test. Even when $N = 100$, this test already indicates the misfit in the continuous part of the model.

Moreover, the models do not predict the continuous variable X_6 well for all N. We see that the quantiles of the conditional distribution of X_6 are not uniform $(0,1)$ for any N. As before, the $RMSE$ tends to decrease with higher N.

Prediction of the discrete variable X_1 does not appear to benefit significantly from the higher number of samples in terms of the proportion of correct predictions. The copula-vine's performance in predicting X_1 is similar that of the GLM.

Table 2.7: *Comparison of the first and second copula-vine model for* $N = 2000$.

Model		First model	Second model
1. Log-likelihood		−2376.99	−1901.12
2. Discrete fit		0.5573	0.5573
3. Continuous fit		$7.7275e-4$	0.8726
Continuous prediction	4. Cond. dist.	$1.6285e-10$	0.8612
	5a. $RMSE$	0.1676	0.1607
	5b. R^2	0.6496	0.6777
Discrete prediction	6a. Threshold	0.46	0.48
	6b. Proportion	87.30%	87.15%

It is apparent that the bivariate normal copulas cannot be used to represent the continuous part of this data. A better model can be obtained if a copula family that better fits the continuous part of the data is used instead. Consider a second model where the continuous variables are represented by the bivariate Clayton copulas and the rest by bivariate normal copulas as before. The estimated parameter values of the second model can be found in Table B.5 in Appendix B.

Table 2.7 summarizes the comparison between the first and second models for $N = 2000$.

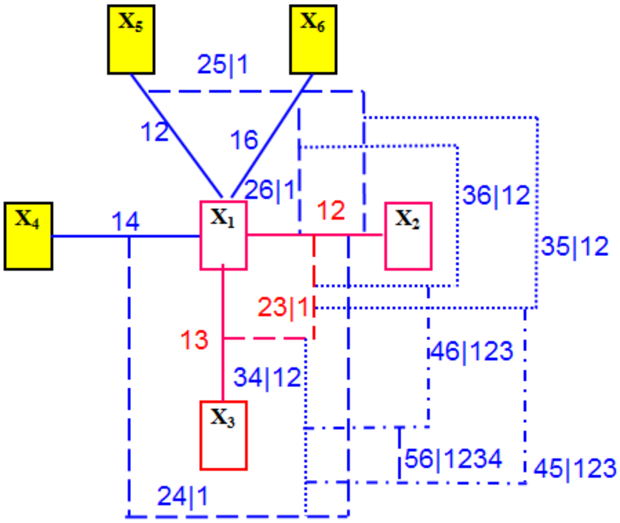
Both the Vuong test and Clarke's distribution-free test conclude that the second model is far superior to the fully bivariate normal copulas model. The test-statistic Z of the Vuong test is -15.51, well below the 95% confidence bound of $(-1.96, 1.96)$. The test-statistic B of the Clarke's distribution-free test is 695, outside of the 95% confidence bound of $(956, 1044)$.

The second model also recovers the conditional distribution of X_6 well. Its $RMSE$ and R^2 are slightly better than the first model, too. However, no significant improvement is observed in the second model's prediction of variable X_1.

From Example 2.3.2, we observe that, in the copula-vine model, it is important to fit the continuous part of the data well with the right copula family, especially when the model is going to be used to predict a continuous variable. For the discrete variables, on the other hand, the non-uniqueness of copulas provides more "freedom" in the choice of which copula family to use.

The next example considers another situation in which the vine structure is misspecified.

Figure 2.12: *A C-vine structure on six variables.*

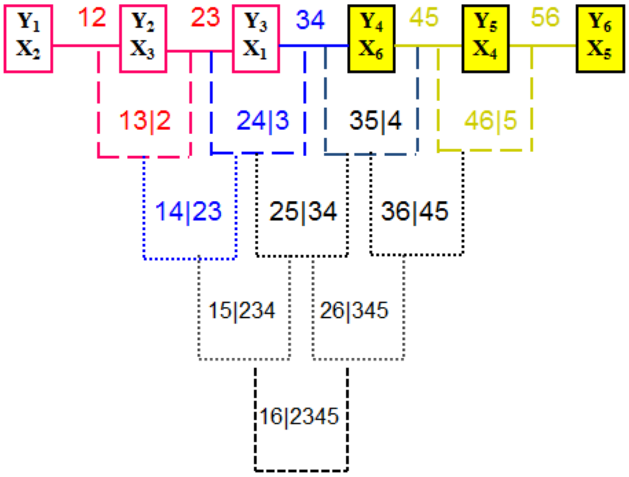


Example 2.3.3. ***(vine structure misspecification)***. Consider three binary variables X_1, X_2, and X_3 with margins $\mathbb{P}(X_1 = 0) = 0.3, \mathbb{P}(X_2 = 0) = 0.7$, and $\mathbb{P}(X_3 = 0) = 0.4$, respectively, and three continuous variables X_4, X_5, and X_6 uniform on $(0, 1)$. The six variables are joined together with a vine depicted in Figure 2.12. Such a structure is another special case of a regular vine called the *C-vine*.

The vine is sampled to obtain datasets of N data points with a set of bivariate normal copulas with parameters as in Table B.6 in Appendix B. As before, we consider $N = \{100, 500, 1000, 2000\}$.

In this example, the true vine structure is assumed to not be known during model construction. To construct the copula-vine model, the discrete and continuous variables are clustered together, forming two sub-vine structures. In each of these, the two pairs with the highest observed correlations are linked in the first tree. The two sub-vines are merged by connecting the discrete-continuous pair with the highest correlation. This results in the vine structure as presented in Figure 2.13.

Figure 2.13: *The chosen vine structure obtained from data by grouping the discrete and continuous variables separately and linking pairs with highest correlations in the first tree.*



The resulting structure is a D-vine, which has fewer parameters than the C-vine structure in Figure 2.12. In other words, the constructed copula-vine models underestimate the complexity of the data.

Let $Y_1, \ldots, Y_6$ represent the variables in the D-vine structure, ordered from left to right as presented in Figure 2.13[4]. The parameters of the D-vine are estimated by executing Algorithm 2.1 with $M = 1000$ with the bivariate normal copulas. Table B.7 in Appendix B presents the estimated parameters.

Table 2.8 shows that the discrete part of the data is recovered well by the copula-vine models for all N. This is not surprising because the three discrete variables are binary. However, the PIT test indicates that the normal copula does not represent the continuous part of the data well. Note that in the true C-vine structure in Figure 2.12, the dependence between the three continuous variables is generated with non-constant bivariate normal copulas.

---

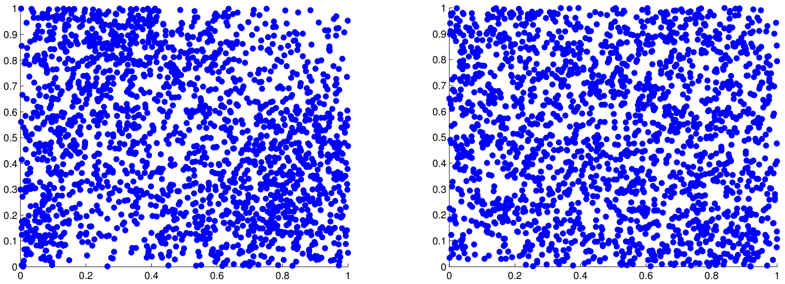
[4] In this example, $C_{13|2}$ represents the conditional copula between Y_1 and Y_3 given Y_2, which corresponds to X_2 and X_1 given X_3 in the C-Vine structure.

Table 2.8: *Summary of the copula-vine models' performance for different numbers of samples N when the vine structure is misspecified.*

N		100	500	1000	2000
1. Log-likelihood		−181.28	−773.90	−1568.69	−3502.69
2. Discrete fit		0.9581	0.9073	0.7864	0.8743
3. Continuous fit		0.0607	0.0315	0.0397	0.0312
Continuous prediction	4. Cond. dist.	0.1192	0.0771	0.0695	0.0090
	5a. $RMSE$	0.2728	0.2364	0.2375	0.2326
	5b. R^2	0.1919	0.3375	0.3219	0.3305
Discrete prediction	6a. Threshold	0.5	0.46	0.47	0.46
	6b. Proportion	75%	87.60%	86.50%	86.80%
	7. GLM	75%	84.80%	84.60%	86.45%

Figure 2.14 compares the scatter plots of $Y_4|Y_5$ vs $Y_6|Y_5$ from the data (Figure 2.14(a)) and from the D-vine model (2.14(b)) when $N = 2000$. In the data, a higher concentration of samples is observed in the bottom left-hand part of the unit square which results in non-symmetrical behaviour with respect to the anti-diagonal. This behaviour is not observed in the fitted D-Vine model.

Figure 2.14: *Scatter plots of $Y_4|Y_5$ vs $Y_6|Y_5$ of the data and the model.*



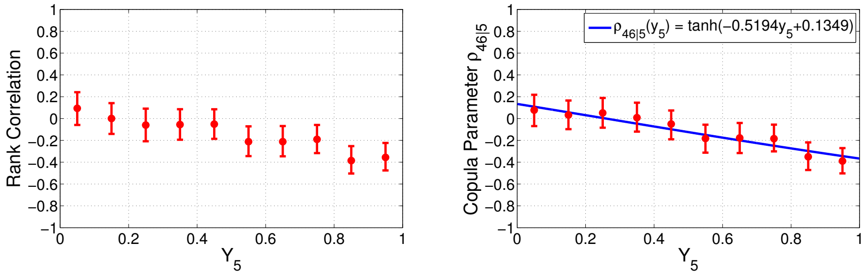
(a) $Y_4|Y_5$ vs $Y_6|Y_5$ in the data.

(b) $Y_4|Y_5$ vs $Y_6|Y_5$ in the model.

To determine whether the conditional copula $C_{46|5}$ can be modelled as a constant copula, the data is divided by discretizing the realization of Y_5 into 10 groups of equal length. For each group, we compute the rank correlation and the normal copula parameter along with their respective 95% confidence bounds between the pair $Y_4|Y_5$ and $Y_6|Y_5$. Figure 2.15 presents the rank correlations (Figure 2.15(a)) and the normal copula parameters (Figure 2.15(b)) of these groups. A weak non-constant relationship with respect to the variable Y_5 is observed in both cases. This finding is confirmed by performing the vector independence test as in (Kurz [2013]), which yields a p-value in the order of 10^{-4}.

Moreover, the conditional distribution of X_6 is not well recovered by the copula-vine model when $N = 2000$. Decreasing behaviour of $RMSE$ with re-

Figure 2.15: *Rank correlations and copula parameters between $Y_4|Y_5$ and $Y_6|Y_5$ for the ten different groups of Y_5.*



(a) The rank correlations.

(b) The copula parameters and the estimated function of $\rho_{46|5}(y_5)$.

spect to N is also observed. The copula-vines' performance in predicting X_1 is similar to that of the GLM.

Now, consider a second model in which a non-constant conditional bivariate normal copula is used to model the continuous part of the data and everything else is kept the same. Figure 2.15 indicates that, for $N = 2000$, the parameter $\rho_{46|5}$ of the conditional copula $C_{46|5}$ might be modelled as a function of the conditioning variable $Y_5 = y_5$. The second model assumes $\rho_{46|5}$ to be of the functional form:

$$\rho_{46|5}(y_5) = \tanh(Ay_5 + b). \tag{2.30}$$

The values of A and b in (2.30) are estimated by maximum likelihood, which results in $A = -0.5194$ and $b = 0.1349$. The hyperbolic tangent is chosen as the "link function" to ensure that $\rho_{46|5}(y_5) \in (-1, 1)$ is always satisfied during the estimation of A and b. The obtained function $\rho_{46|5}(y_5)$ is plotted as the blue line in Figure 2.15(b) and is captured by the confidence bounds of the copula parameters. The parameters of the second model can be found in Table B.8 in Appendix B. It is shown there that the confidence bounds of neither A or b contain zero.

Table 2.9 summarizes the comparison between the first and second models for $N = 2000$.

The LRT, Vuong test, and Clarke's distribution-free test conclude that the second model is better than the first. The p-value of the LRT statistic is smaller than 10^{-16}. The test-statistic Z of the Vuong test is -2.8656, lower than the 95% confidence bound of $(-1.96, 1.96)$. The test-statistic B of the Clarke's distribution-free test is 952, just below the 95% confidence bound of $(956, 1044)$.

The continuous part of the data is also recovered by the second model, as indicated by the p-value of the PIT test. Moreover, this model also predicts the conditional distribution of variable X_6 better than the first. However, no significant improvement in the values of $RMSE$ or R^2 is observed.

Table 2.9: *Comparison of the first and second copula-vine model for $N = 2000$.*

Model		First model	Second model
1. Log-likelihood		−3502.69	−3123.27
2. Discrete fit		0.8743	0.8743
3. Continuous fit		0.0312	0.2672
Continuous prediction	4. Cond. dist.	0.0090	0.0813
	5a. $RMSE$	0.2328	0.2326
	5b. R^2	0.3292	0.3305
Discrete prediction	6a. Threshold	0.46	0.45
	6b. Proportion	86.80%	86.25%

2.4 Chapter Summary

In this chapter we have studied the use of copulas and vines in the dependence modelling of mixed discrete and continuous variables. The presence of one or more discrete variables in the model makes the copulas satisfying Sklar's equation (1.2) non-unique. From the dependence modelling point of view, this provides more freedom in the choice of copula family that can be used. On the other hand, the copula parameter estimation becomes more expensive because it needs to be performed via maximum likelihood.

Our focus in this chapter has been the normal copula, because this is the one we intend to use in the application part of this thesis. In two-dimensional cases, it has been shown that the bivariate normal copula can always recover the dependence of a bivariate Bernoulli distribution. In higher-dimensional cases, however, it is not easy to give conditions which assert the existence or non-existence of a multivariate normal copula to represent the given multivariate Bernoulli distribution.

A second modelling approach has also been considered, involving the graphical structure known as the vine. We adapted the copula-vine approach popular in fully continuous cases to the mixed discrete and continuous setting. When all the discrete variables are Bernoulli, it was shown that the discrete part of the model can always be recovered with a set of non-constant (conditional) bivariate normal copulas as guaranteed by Theorem 2.1.8. If at least one of the discrete variable is not Bernoulli, however, the copula-vine approach with non-constant (conditional) bivariate normal copulas no longer always recovers the discrete part of the model.

To construct a copula-vine model, a copula family and a vine structure need to be chosen. The non-uniqueness of copulas when the variables are discrete provides more freedom in the choice of copula family to use. However, we notice the importance of choosing the right copula family in the continuous part of the model. The choice of an "incorrect" vine structure might produce a model that is either too complex to represent the data or not complex enough.

After both the copula family and vine structure have been chosen, Algorithm 2.1 is proposed to fit the model parameters. This is a sequential procedure. In principle, a better model can be constructed by fully optimizing the model's likelihood. However, from an example considered in this thesis, it appears that the improvement is not significant enough to justify the much heavier computational cost of the fully-optimized model. This observation agrees with the study performed by Haff [2013] in a fully continuous setting. Consequently, in this thesis the parameters of the copula-vine models are estimated sequentially.

From the perspective of predicting the outcome of a Bernoulli variable, the performances of copula-vine models and GLMs are found to be similar. Both types also produce similar results when used to predict the mean of the outcome of a continuous dependent variable. This indicates the marginal benefit obtained from constructing a model using the copula-vine rather than the GLM approach.

However, the conditional distribution of the continuous dependent variable can also be obtained with the copula-vine approach. Moreover, this models joint distribution as well, allowing users to study the dependence structure between the variables. This is particularly useful in policy analysis, for example, when the model's aim is to learn the effect of one variable on the other. This cannot be achieved with the GLM.

In the next chapter we begin construction of the railway disruption-length model. We start with data analysis to determine the necessary influencing factors to be involved in the dependence model. Next, two joint distribution models are constructed, using two different strategies. These are then compared to determine which is better for our data.

CHAPTER 3

Construction of the Railway Disruption-Length Model[1]

This chapter is concerned with the construction of the railway disruptions length model.

Before construction can begin, factors influencing the length of disruptions need to be investigated. This chapter therefore starts with data analysis to determine those factors. We use data originating mainly from the SAP database described in subsection 1.1.3. Only high-priority incidents (priority 1 or 2) recorded in that database are considered, since they require urgent action to solve the problems. The data is also preprocessed to remove samples that are not registered correctly[2]. This results in 1920 TC and 2484 switch incidents in the period between 1 January 2011 and 30 June 2013.

The factors influencing the disruption-length models for TC and switch disruptions are generally the same. Moreover, the same model construction techniques are used for both types. Consequently, the presentation in this chapter focuses mainly on the TC disruptions. In section 3.4, we briefly present the data analysis and model construction for the switch disruptions. The same steps can be followed for models of other disruptions.

After the influencing factors are determined, the disruption-length model is constructed using two strategies: (i) with the multivariate normal copula and (ii) with the copula-vine approach as discussed in section 2.3. The two models are then compared to determine which is more attractive from the point of view of railway traffic management.

---

[1] The data analysis part of this chapter is based on Zilko et al. [2016] and the model construction part is based on Zilko and Kurowicka [2016]

[2] For instance, a non-TC incident that is recorded as a TC incident.

3.1 Data Analysis

We consider the disruption length to be the sum of the lengths of the latency and repair times. The factors influencing these two subperiods are different and so will be considered separately.

3.1.1 Factors Influencing Latency Time

In general, the factors influencing the length of latency time can be divided into three different groups: time, location and weather. Later, one additional variable not part of any of these groups – the presence of an overlapping disruption – is also included in the model.

3.1.1.1 Time

Three variables are initially considered to represent time: (i) whether or not the disruption occurs during the repair teams contractual working hours (weekdays between 7 am and 4 pm); (ii) whether or not the disruption occurs during rush hours (weekdays between 7 and 10 am and between 4 and 7 pm); and (iii) whether or not the disruption occurs at the weekend. However, information about weekends is already contained in the first two variables. For instance, when an incident occurs at the weekend, it is outside the repair workers' contractual working hours and not during rush hour. Moreover, the characteristics of weekends are similar to those of non-working hours and non-rush hours. For this reason, we only take into account the first two variables.

The first factor is important because different operations are performed depending whether or not the incident occurs during the contractual working hours. During those hours, the repair team departs for the disruption site from its working base. Outside those hours, they are not at their base even though they are available on call. In this case, they leave for the disruption site from wherever they happen to be. It turns out that the latency time outside contractual working hours is longer than during them – for a TC disruption, on average about 4.5 minutes longer. Figure 3.1 shows the empirical distribution of latency time for the TC disruptions during and outside contractual working hours. Performing the two-sample KS test and the two-sample CvM test on these two distributions yields p-values in the order of 10^{-5} or lower for both tests. This indicates that the observed difference between the two latency-time distributions cannot be explained by the sample noise with a significance level of 0.05.

To reach a disruption site, the repair team travels by car. During rush hours, it is more likely to encounter traffic congestion which might prolong the latency time. The influence of rush hours on the latency time is smaller than that of contractual hours, with latency time during rush hours is slightly longer than at other times. For TC disruptions, the p-value of the two-sample KS test and CvM test in this case are 0.0402 and 0.0332, respectively. On average, the latency time is around 0.5 minutes longer during rush hours.

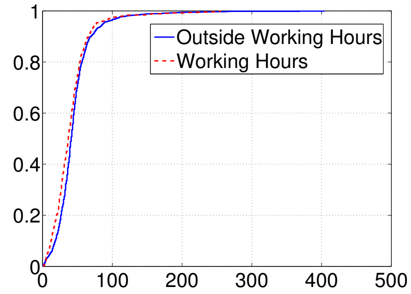


Figure 3.1: *Empirical distribution of latency time for TC disruptions during and outside contractual working hours.*

3.1.1.2 Location

The location of an incident needs to be described using some representative properties which affect the length of the latency time. Figure 3.2 shows the locations of the 5% of disruptions in our sample with the longest observed latency times, i.e. in excess of 90 minutes.

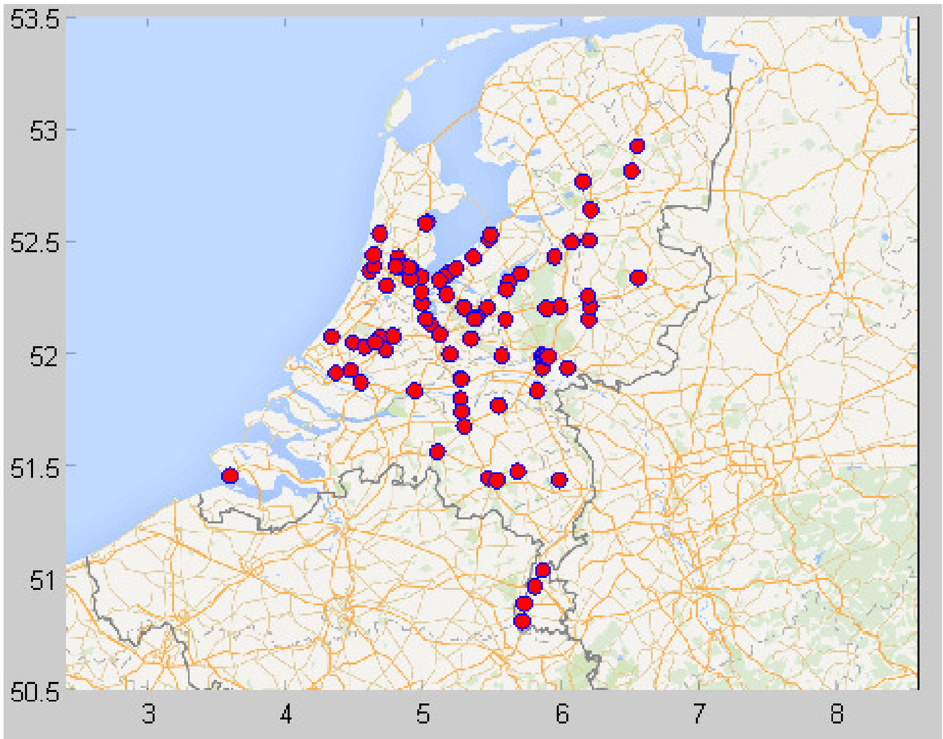


Figure 3.2: *Map of TC disruptions with latency time longer than 90 minutes.*

On the map, we observe that around 60% of the longest latency times occurred in the Randstad region in the west of the Netherlands. Being the busiest part of the country, there are more working stations, or maintenance bases, situated here than elsewhere in the Netherlands. However, road and rail traffic densities are also higher. Road traffic density has been covered by the variable "rush hour" above. Another characteristic to consider is level crossings; in the Randstad, there are far fewer of these than in other parts of the country.

Based on the above observation, four variables representing the properties of location are investigated:

1. distance to the nearest mechanics' working station;
2. train traffic density;
3. distance to the nearest level crossing; and
4. contract type.

The first variable is intuitive because the more distant the disruption is, the longer it takes for repair team to reach it. The second one represents how busy the location is. For instance, a location with denser traffic indicates proximity to bigger cities with better infrastructure (e.g. roads or easy access). The third variable represents the site's accessibility. To access the disruption site, the repair team needs to park its cars (usually at the nearest level crossing) and walk to the site. In an interview with a repair worker from a contractor company, the fourth variable was also found to be an important factor influencing latency time. Why it is seen as a property of location is explained shortly.

Distance to the nearest working station

In the dataset, the disruption site is indicated as being located between two operational points along the tracks with known GPS coordinates (latitude and longitude). The disruption site's coordinates are estimated by taking the average of the coordinates of these two operational points. The towns or cities where maintenance bases are located are also known. Their exact locations are approximated by the using the position of the main railway station in the town or city in question, its GPS coordinates also being known. Therefore, the distance between a disruption site and the nearest maintenance base is approximated by calculating the straight-line distances between the sites estimated coordinates $(Lat_d, Long_d)$ and those of the base $(Lat_c, Long_c)$, as follows:

$$d = 6371 \cdot \arccos\left(\sin Lat_d \sin Lat_c + \cos Lat_d \cos Lat_c \cos(Long_d - Long_c)\right) \tag{3.1}$$

where the constant 6371 corresponds to the Earth's radius in kilometres (Steinhaus [1999]).

To determine which working station is the closest, it is not enough to simply take the one with the shortest approximated distance to the site. In the Netherlands, repair work is outsourced to four major contractor companies, each responsible for its own region. This is important because a disruption in the region assigned to contractor *A* is always handled by its personnel, even if one of contractor *B*'s working stations is physically closer. Furthermore, each contractor splits its operational region into several subregions, each with its own repair team(s) and working station(s) responsible for incidents there. Figure 3.3 shows the 23 subregions in the Dutch railway network, along with the locations of their working stations.

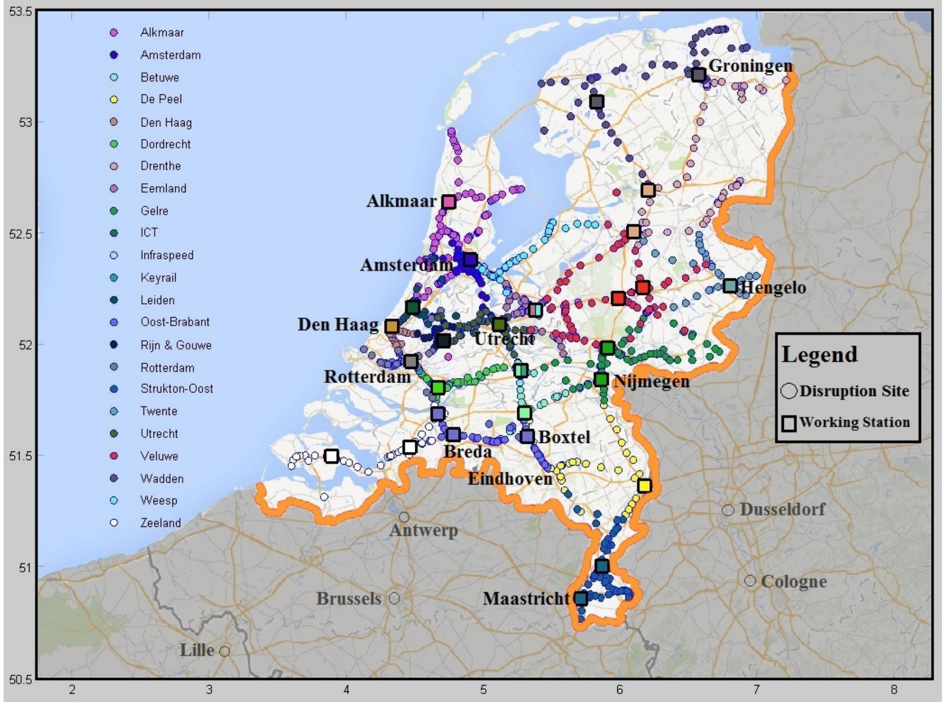


Figure 3.3: *The* 23 *subregions in the Netherlands and the location of the working stations.*

We therefore define the nearest working station as the one in the subregion in which the incident occurs with the shortest approximated distance to it. To measure the effect of this distance on the latency time, the rank correlation between the two variables is computed. It is found to be 0.1380 with 95% confidence bound of $(0.0925, 0.1829)$. Zero is not in the confidence bound, indicating small positive dependence between the two variables.

Train traffic density

The train traffic density of a location is defined as the average number of trains passing it per day. Information about the number of passing trains at each operational point is known and the density can be derived from this information.

The rank correlation between the two variables is found to be -0.0352 with 95% confidence bound of $(-0.0810, 0.0107)$. This indicates a very small dependence between the two variables. As zero is included in the confidence bound, the data indicates that the variables can be modelled independently of one another.

The influence of traffic density on other variables is found to be insignificant as well. For instance, the 95% confidence bound of the rank correlation between train traffic density and latency time during the repair workers' contractual working hours is $(-0.0465, 0.0456)$, while outside this time period it is $(-0.0436, 0.0485)$. Zero is included in both confidence bounds, which means that the hypothesis of conditional independence between train traffic density and latency time given the repair workers' contractual working hours is not rejected. Similar conclusions are drawn when other variables are considered.

For this reason, this variable is not included in the model.

Distance to the nearest level crossing

The rank correlation between the latency time and the distance to the nearest level crossing for TC disruptions is calculated to be 0.0842 with 95% confidence bound of $(0.0383, 0.1297)$. Zero is not in the confidence bound, so a small positive dependence between the two variables is detected.

Contract type

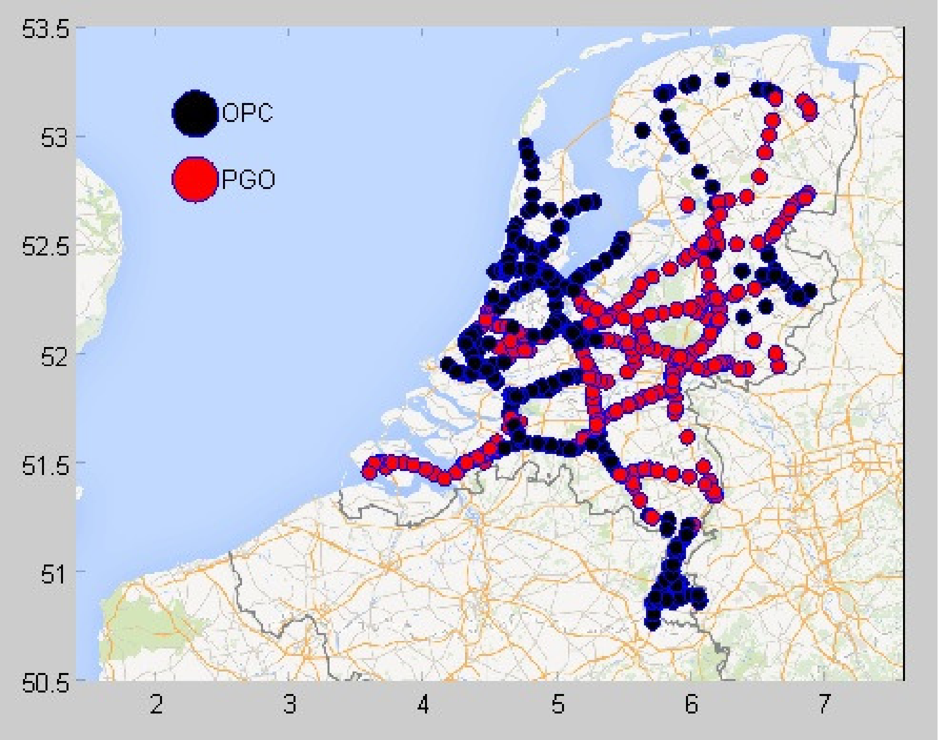


Figure 3.4: *Distribution of OPC and PGO contracts in the Netherlands.*

As described in Subsection 1.1.3, there are two types of contract between

ProRail and its contractors: OPC and PGO. For the TC disruptions, the latency time of an incident with an OPC contract is, on average, 3.3 minutes longer than with a PGO contract. Performing the KS and CvM tests on the latency times of disruptions with OPC and PGO contracts yields p-values in the order of 10^{-5} or less for both tests.

Figure 3.4 shows why contract type is considered as a property of location. It is clear that certain regions in the Netherlands are dominated by the OPC contract, others by the PGO contract. Interestingly, the old OPC contract still predominates in the majority of the Randstad, which may explain why more disruptions with long latency times are seen in this region in Figure 3.2. Moreover, this variable also explains the long latency times observed in the province of Limburg in the south-east of the Netherlands.

3.1.1.3 Weather and Overlapping Disruptions

As described in Subsection 1.1.1, TC is sensitive to high temperatures. When the temperature is high enough, the TC might experience adjustment problems which cause malfunctions. For this reason, a threshold for "high" temperature needs to be defined. As we are interested in the influence of this variable on the length of latency time, several different thresholds are considered. For each, the latency times are divided into two groups, one for "not warm" temperatures (below the threshold, indicated as 0) and one for "warm" (indicated as 1). The difference between the distributions of the two groups is measured using the two-sample KS and CvM tests, as before. The threshold which produces the lowest p-values in the tests is chosen. Table 3.1 shows the result.

Table 3.1: *P-values of the KS and CvM Tests for latency time at different high-temperature thresholds.*

Threshold	KS	CvM	Threshold	KS	CvM
20^oC	0.4184	0.2163	26^oC	0.3138	0.1759
21^oC	0.3569	0.1098	27^oC	0.2183	0.1148
22^oC	0.4984	0.2248	28^oC	0.4954	0.3862
23^oC	0.2208	0.1464	29^oC	0.6569	0.5109
24^oC	0.1567	0.1001	30^oC	0.9164	0.8186
25^oC	**0.0938**	**0.0727**			

Table 3.1 shows that the threshold choice of 25^o is the optimum for our data, being where the strongest effect on the latency time is observed. This is indicated by the lowest p-values of both the KS and CvM tests, by comparison to the other thresholds. Higher thresholds have lower numbers of observable warm samples[3], and hence less reliable data analysis, while lower thresholds are meaningless from the point of view of the definition of the variable.

However, the binary variable "warm" with the chosen threshold of 25^oC does not appear to have any direct influence on the latency time because the p-values

---

[3] With the threshold of 25^oC, there are only 112 occurrences in the data where the temperature is "warm"

of both tests are above 0.05. This indicates that the observed difference between the two latency-time distributions can be explained by sample noise.

However, the variable "warm" does affect another variable, the presence of an overlapping incident. Two incidents in the data are considered overlapping if their disruption times coincide in whole or in part, they are similar in type and they are handled by the same contractor. When "warm" is zero, the probability of an overlapping incident occurring is 4.15%, while when "warm" is one, that probability increases to 16.07%. This is understandable because high temperatures may trigger some TCs to fail more or less simultaneously. With a limited number of repair teams, some of these incidents can only be dealt with once others have been taken care of.

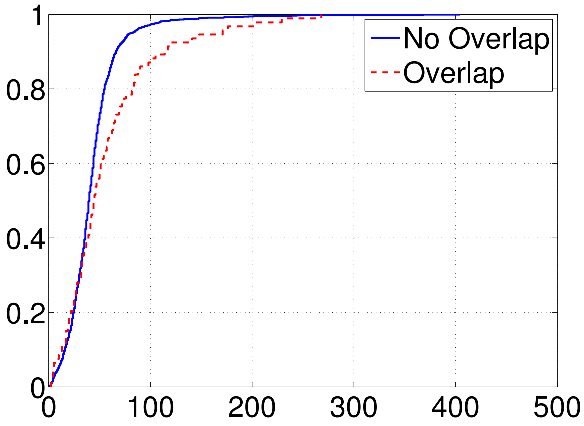


Figure 3.5: *Empirical distribution of latency time with respect to the presence of an overlapping incident.*

Whether or not there is an overlapping incident does affect latency time. On average, the latency time of a TC disruption is about 14.5 minutes longer when an overlapping incident exists. In this case, the KS and CvM tests yield *p*-values of 0.0025 and 0.0056, respectively. Figure 3.5 presents the latency-time distributions when an overlapping disruption does and does not exist.

3.1.2 Factors Influencing Repair Time

3.1.2.1 Contract Type

The contract type also affects the repair time, as indicated by the plot of the distribution of repair-time lengths presented in Figure 3.6. For TC disruption, on average the repair time is 20.5 minutes longer when the contract type is OPC. Performing the two-sample KS and CvM tests on the two repair-time distributions based on contract types yields *p*-values in the order of $\leq 10^{-5}$ or less, respectively.

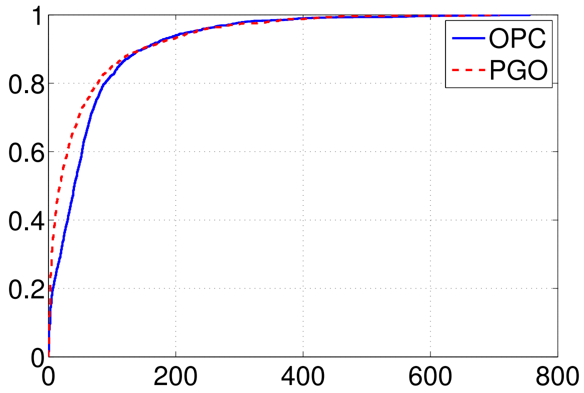


Figure 3.6: *Empirical distributions of TC disruption repair times with respect to contract type.*

3.1.3 Causes

Naturally the type of failure and the external causes of problems discussed in subsection 1.1.1.1 influence the lengths of repair times. Different failures require different corrective actions and thus different repair times.

Based on the characteristics of the required action, the causes of TC disruptions are grouped as follows.

1. Group 1: impedance bond failure.
2. Group 2: relay cabinet failure, cable problem, trackside electrical junction box problem and arrestor problem.
3. Group 3: external reasons.
4. Group 4: splinter/grinding chips and insulator problem.
5. Group 5: coins.
6. Other.

The first five groups are ordered based on the length of repair time. Groups 1 and 2 require the repair team to replace components. Impedance-bond failures are grouped separately because teams do not usually carry the tools needed to perform the required replacement work. In the cases in Group 2, they do usually have the necessary tools with them. Consequently, the repair time for an impedance-bond failure tends to be longer than with the causes in Group 2. The causes in Group 3 generally require adjustment of the components concerned. Groups 4 and 5 both contain problems with the joint insulator; they are divided because they need slightly different corrective action. Problems in Group 4 require the repair team to clean the joint insulator and sometimes to renew its nylon plates and linings. Due to the small size of the insulator, this does not take as long as the replacement work in Group 2. The coin problem in Group 5 requires only removal of the coin from the affected joint insulator.

There are also a few TC problems caused by problems not in the first five groups. They occur very infrequently, with a total of only 20 such incidents found in the data. Two examples are a case of water flooding the track and relay cabinet (the repair team had pump out the water and clear the equipment) and a loss of detection due to sand contamination.

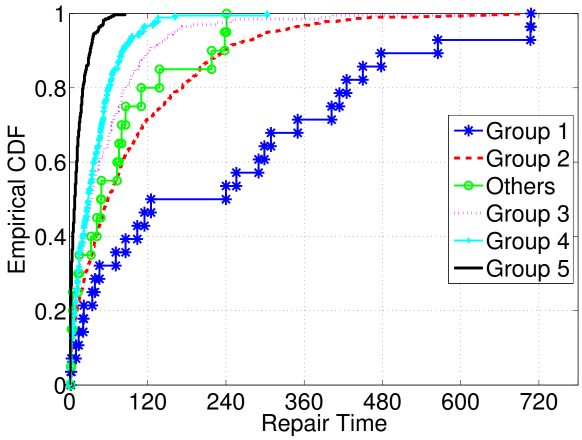


Figure 3.7: *Empirical distributions of repair time in the six groups.*

The characteristic of the repair-time length of the groups is also reflected in the data. Figure 3.7 shows the empirical distributions of repair time in the six groups. Clearly, the group "Other" is positioned between Group 2 and Group 3 in term of the repair-time length. Moreover, the difference in repair times between different groups is evident.

As mentioned in subsection 1.1.3, the cause of approximately 30% of TC incidents is unknown due to unclear or oversummary descriptions. To tackle this situation, one option would be to discard these "unknowns" from the dataset. This, however, would result in the loss of information they carry for the other variables. Another option would be to redistribute them randomly, but proportionally, into the six groups. However, this approach neglects the repair-time length information that we want to associate with the groups.

We therefore propose a redistribution approach which takes into account the dependence between repair time, cause and contract type. This approach is also known as the "Bayesian classifier" and is a popular classification technique which has been implemented in many studies, e.g. Marchant and Onyango [2002], Bender et al. [2004] and Wang et al. [2007]. It is used to calculate the probability that each unknown sample X_j to belong to group i, given its observed repair time $R = r_j$ and contract type $T = t_j$. Using Bayes' theorem, this probability is proportional to that of repair time $R = r_j$ and contract type $T = t_j$ given that it belongs to group i (called the *likelihood*) multiplied by the probab-

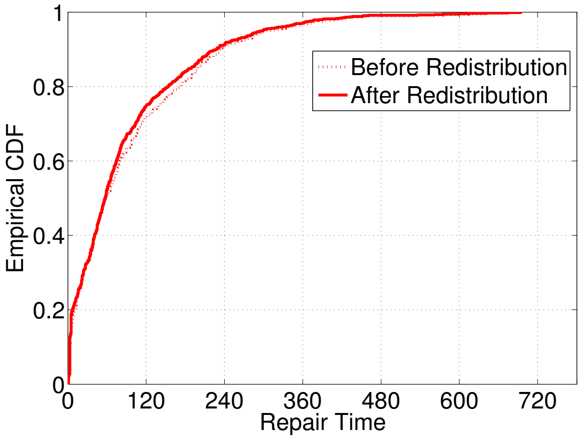


Figure 3.8: *Empirical distribution of repair times in Group 2 before and after redistribution.*

ility of group i (called the *prior*). Dividing the product of likelihood and prior by the probability of the observed repair time $R = r_j$ and contract type $T = t_j$ gives the desired probability. The sample is then assigned to the group with the highest probability. Because repair time is continuous, the "probability" of $R = r_j$ is approximated by discretizing the repair time to a range of 5% of r_j, i.e. $\mathbb{P}(0.975r_j \leq R \leq 1.025r_j)$. This results in the following formula:

$$\begin{aligned}&\mathbb{P}(Group = i|0.975r_j \leq R \leq 1.025r_j, T = t_j)\\&\qquad = \frac{\mathbb{P}(0.975r_j \leq R \leq 1.025r_j, T = t_j|Group = i)\mathbb{P}(Group = i)}{\mathbb{P}(0.975r_j \leq R \leq 1.025r_j, T = t_j)}. \qquad (3.2)\end{aligned}$$

Note that there is always a group i in which $\mathbb{P}(0.975r_j \leq R \leq 1.025r_j, T = t_j|Group = i) > 0$ for each X_j. With this technique, an unknown sample with a short repair time is more likely to be redistributed into a group with shorter repair times, and vice versa.

Table 3.2: *Number of samples, proportion, mean and standard deviation of repair times in each cause group before and after redistribution. The information after redistribution is presented in brackets.*

Cause Group	# Samples	Proportion (%)	Mean	Std. Dev
Group 1	28 (28)	2.14 (1.46)	235.21 (235.21)	218.47 (218.47)
Group 2	602 (944)	46.06 (49.17)	96.67 (89.65)	109.11 (104.12)
Others	20 (20)	1.53 (1.04)	73.90 (73.90)	78.50 (78.50)
Group 3	198 (269)	15.15 (14.01)	55.08 (45.55)	74.43 (68.32)
Group 4	182 (215)	13.93 (11.20)	38.15 (36.83)	38.13 (35.73)
Group 5	277 (444)	21.19 (23.12)	12.64 (10.74)	13.94 (12.37)
All	1307 (1920)	100 (100)	67.03 (61.27)	97.01 (91.76)

Table 3.2 presents some information about the repair times of each group

before and after the redistribution of unknowns. Groups with high proportions (Group 2 and Group 5) receive the most assigned unknowns, while those with very low proportions (Group 1 and Other) receive none. Moreover, most unknown samples with a long (short) repair time are redistributed into a group with a long (short) repair time. For instance, of the top 10% of unknown samples in terms of the longest repair time, only one is not redistributed into Group 2[4].

Figure 3.8 shows the distribution of repair times in Group 2, which receives the most redistributed unknowns, before (dashed) and after (solid) redistribution. The redistribution appears to slightly shift the distribution of repair times to the left.

3.2 Model Construction

In this section, we discuss the construction of the railway disruption-length model.

We introduce the following notations, which we will use throughout this section to represent the variables: CT represents "contract type", WD represents "distance to the nearest working station", LC represents "distance to the nearest level crossing", WT represents "working (contractual) hours", WM represents "warm", RH represents "rush hour", OV represents "presence of an overlapping incident", CS represents "cause", LAT represents "latency time", and REP represents "repair time".

Five of the variables are binary with the following margins: $\mathbb{P}(CT = 0) = 0.4542$, $\mathbb{P}(WT = 0) = 0.6781$, $\mathbb{P}(RH = 0) = 0.7505$, $\mathbb{P}(WM = 0) = 0.9417$, and $\mathbb{P}(OV = 0) = 0.9515$. The variable CS has six states with the following margin:

$$\mathbb{P}(CS = j) = \begin{cases} 0.0146, & \text{for } j = 1; \\ 0.4917, & \text{for } j = 2; \\ 0.0104, & \text{for } j = 3; \\ 0.1401, & \text{for } j = 4; \\ 0.1120, & \text{for } j = 5; \\ 0.2312, & \text{for } j = 6; \end{cases}$$

corresponding to the third column in Table 3.2.

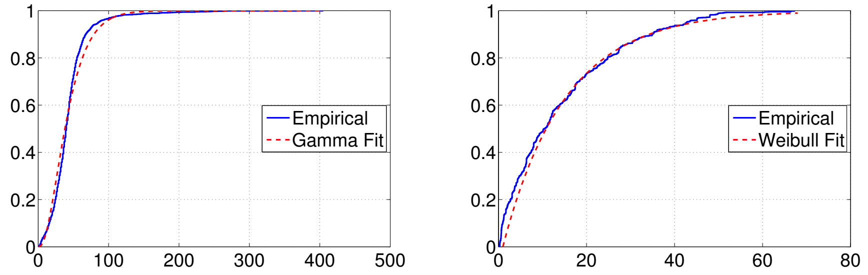
For the four continuous variables, there is an option to represent them by known parametric distributions. One advantage of fitting a parametric distribution to a continuous variable is the ability to conditionalize the variable on more extreme observations than we observe in the data.

We consider four parametric distributions for fitting: exponential, log-normal, Gamma and Weibull. The use of these four distributions is not new in the railway operation research. The exponential distribution is used to estimate the disruption length in Schranil and Weidmann [2013]. The lognormal distribution is under consideration because, by definition, the four continuous variables are non-negative. In a slightly different setting, Yuan [2006] considers the log-normal, Gamma and Weibull distributions to model several different kinds of

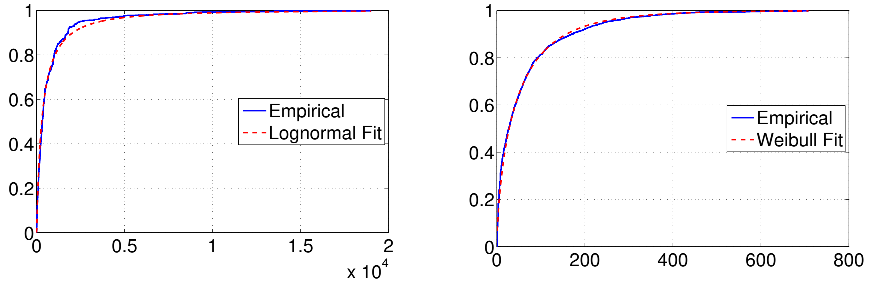
---

[4] The sample is redistributed into Group 3.

Figure 3.9: *Fitting parametric distributions to the continuous variables.*



(a) Latency time. (b) Distance to working station.



(c) Distance to level crossing. (d) Repair time.

train delay. More recently, Corman et al. [2011] and Jensen et al. [2015] use the Weibull distribution to model arrival, departure and dwell delay based on the work of Yuan [2006].

For each continuous variable and each parametric distribution, the distribution's parameter(s) is (are) calculated using maximum likelihood. This results in the latency time fitting best with gamma distribution, distance to the nearest level crossing with the log-normal distribution and both distance to the nearest maintenance base and repair time with the Weibull distribution. Figure 3.9 shows the empirical distributions (solid blue lines) and the best-fitted parametric distributions (red dashed lines) of the four continuous variables. The goodness of fit is measured by means of the KS test. This rejects the hypothesis that any of the variables can be represented with the proposed parametric distributions with p-values in the order of 10^{-5} or lower.

For this reason, in the model we have decided to use the empirical distributions of the continuous variables.

3.2.1 The Multivariate Normal Copula Model

The first model we consider is the multivariate normal (MVN) copula model. To save space, we denote $\mathbf{X_d} = (X_1, \ldots, X_6) = (CS, CT, WM, WT, RH, OV))$, which represents the discrete variables, and $\mathbf{X_c} = (X_7, \ldots, X_{10}) = (WD, LC, LAT, REP)$

which represents the continuous variables. The parameter R of the copula is estimated by maximum likelihood. The likelihood is computed as:

$$\ell(R) = \sum_{i=1}^{N} c_{R_{X_c}}(F_7(x_{7i}),\ldots,F_{10}(x_{10i}))\mathbb{P}_{R_{X_d|X_c}}(\mathbf{x_{di}}|\mathbf{x_{ci}}) \tag{3.3}$$

where $N = 1920$ corresponds to the number of samples, $\mathbf{x_{ci}}$ denotes the i-th realization of $\mathbf{X_c}$, $\mathbf{x_{di}}$ is the i-th realization of $\mathbf{X_d}$, R_{X_c} is the part of the correlation matrix R which corresponds to the continuous part of the model and $R_{X_d|X_c}$ is the correlation matrix of the normal copula $C_{R_{X_d|X_c}}$ corresponding to the conditional distribution of the discrete variables given the continuous variables in the model.

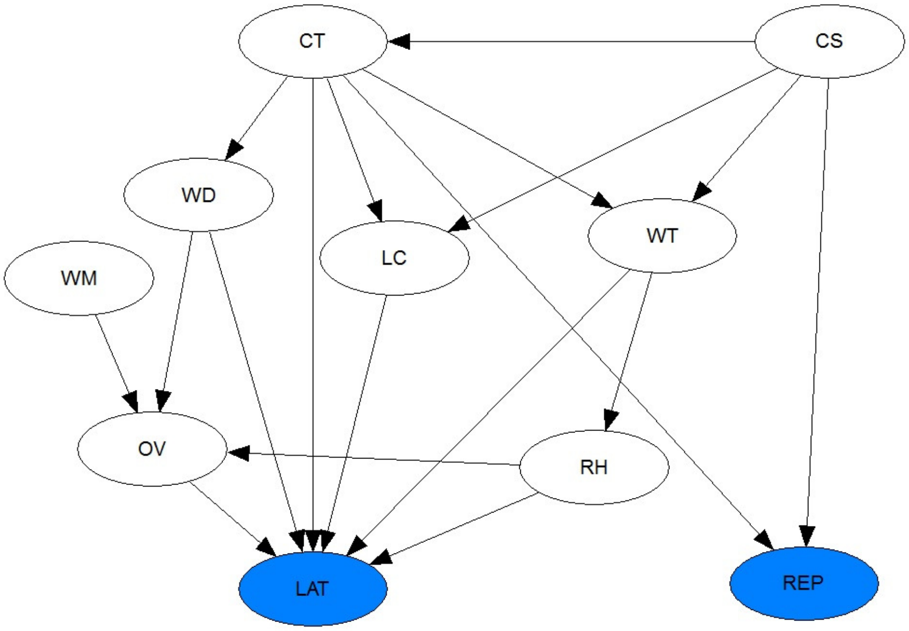


Figure 3.10: *The Track Circuit BN.*

The probability mass function in (3.3) is obtained by calculating the finite difference of the multivariate normal copula of discrete variables conditional on the continuous variables, as follows:

$$\mathbb{P}_{R_{X_d|X_c}}(\mathbf{x_{di}}|\mathbf{x_{ci}}) = \sum_{s_1=0}^{\min(x_{1i}-1,1)} \sum_{s_2=0}^{x_{2i}} \ldots \sum_{s_6=0}^{x_{6i}} (-1)^{\sum_j s_j} C_{R_{X_d|X_c}}(\mathbb{P}(X_1 \leq x_{1i} - s_1),\ldots,\mathbb{P}(X_6 \leq x_{6i} - s_6)) \tag{3.4}$$

The parameters of the normal copula that maximize (3.3), computed with the built-in MATLAB command, are presented in Table C.1 in Appendix C. The computation was time-consuming, with parameter estimation taking approximately 24 hours with an Intel(R) Core i5-3470 3.2 GHz processor and 8 GB RAM.

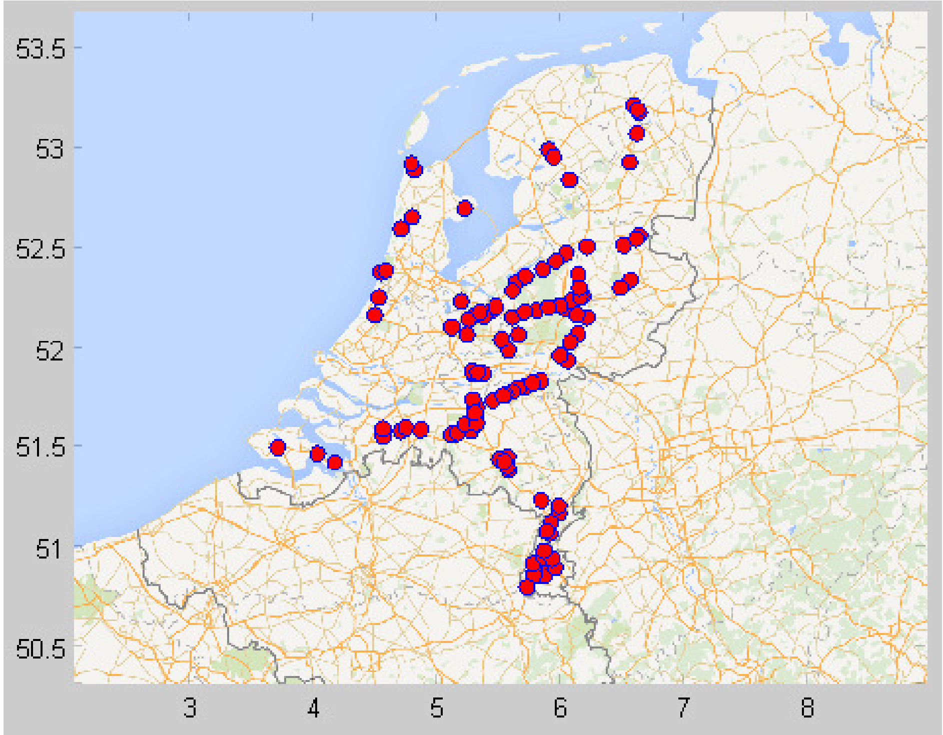


Figure 3.11: *Locations of observed TC disruptions caused by coins.*

This MVN copula model is a saturated model where the correlations between all pairs are considered. However, it can be seen that some of the estimates are small, suggesting independencies; for instance, between variables CT and WM it is 0.0211. In principle, confidence intervals can be computed via simulation, but this is not feasible due to the very long computation time.

To obtain a more parsimonious MVN copula model, we impose some conditional independencies that are represented as a BN. Figure 3.10 presents the BN structure of the TC disruption-length model, which we obtained from discussions with a ProRail expert.

The BN in Figure 3.10 implies, for instance, conditional independence between CS and LAT given CT, LC, and WT. While CS does not directly affect LAT, a certain cause is more likely in certain areas at certain time and this influences the latency time. Therefore, once information about location (represented by CT and LC) and time (represented by WT) is available, information about CS becomes irrelevant to LAT.

The BN structure also illustrates the interdependencies between the influencing variables. For instance, CS influences CT, LC and WT. Figure 3.11 presents the locations of disruptions caused by coins ($CS = 6$) in the data. It shows that most of these are observed outside the Randstad, where the contract type is mostly PGO ($CT = 1$). Moreover, the closer a level crossing (lower LC), the greater the chance that the disruption is caused by coins due to easier accessibility. This tallies with a characteristic of the railways in the Randstad: the presence of fewer level crossings. Moreover, these disruptions are also observed more often in the evening or at the weekend, outside the repair personnel's contractual working hours ($WH = 0$).

Another way to determine the BN structure is by learning it from data. A number of algorithms have been developed for this purpose, when the variables are all discrete. One is the "hill-climbing greedy search" in the space of all possible BN structures, a score-based algorithm which assigns a score to each possible BN structure based on the data. The algorithm chooses the structure which maximizes the score. The score of a structure $\mathcal{G}$ is defined as the probability of the structure given the data $\mathcal{D}$, i.e. $\mathbb{P}(\mathcal{G}|\mathcal{D})$ (Margaritis [2003]). Because our model consists of six discrete and four continuous variables, we need to discretize the continuous variables to be able to use this algorithm. Due to the limited number of samples, the continuous variables are discretized into four discrete states with equal proportions[5]. Performing the hill-climbing search results in the BN structure presented in Figure 3.12. The structure is obtained using the package `bnlearn` in `R`, developed by Scutari [2010].

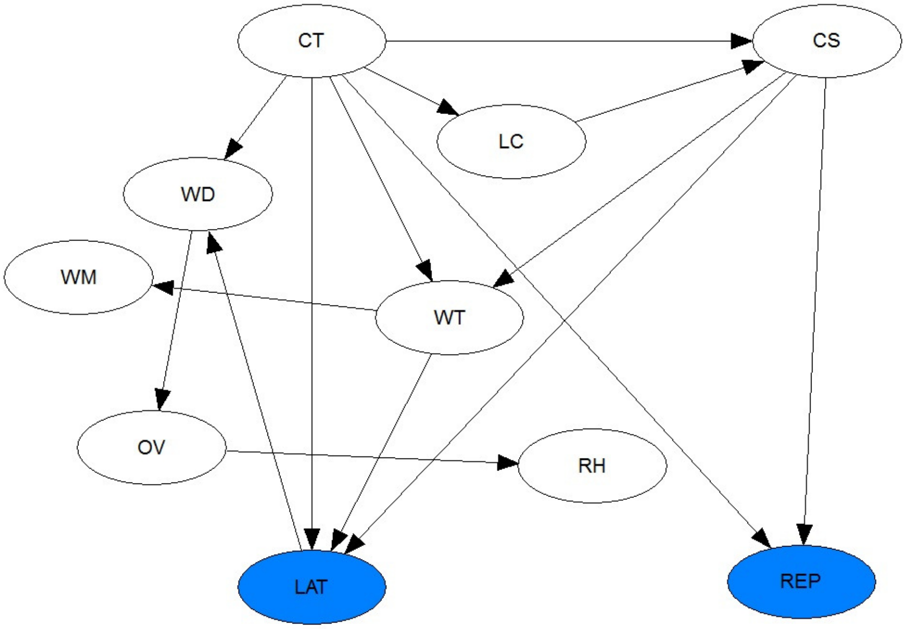


Figure 3.12: *The track circuit BN obtained from data.*

Even with a few missing arcs, and others with changed direction, a resemblance can be observed between the structures in Figures 3.10 and 3.12. Margar-

---

[5] This is done by computing the variable's 25%, 50%, and 75% quantiles.

itis [2003] studies the performance of the algorithm where it is not always able to fully recover the directionality of the arcs. Moreover, the BN in Figure 3.12 is obtained by discretizing the continuous variables. The missing arcs and reversed direction may be artefacts of this discretization.

To measure which structure better represents the data, the log-likelihoods of the two MVN copulas corresponding to the two structures are computed. Using the AIC to compare the log-likelihoods, the AIC score of the first BN structure is 12817.44, while that of the second is 13036.45. This indicates that the first structure models the data better, so in this thesis we proceed with the first BN. The Vuong and Clarkes tests confirm this as well. The test statistics Z and B are 5.5722 and 1198, which are outside the 95% confidence bounds of $(1.96, 1.96)$ and $(917, 1003)$, respectively.

The parameters of the MVN copula with conditional independence implied by the structure in 3.10 are presented in Table C.2 in Appendix C. Table 3.3 compares the fit of the saturated MVN copula model and the MVN copula with conditional independence model to the data presented in section 2.3.

Table 3.3: *Comparison of the multivariate normal copula models.*

Model	Saturated MVN copula	MVN copula with CI
1. Log-likelihood	-6368.1948	-6375.7219
2. Discrete fit	$4.0699e-06$	$2.2119e-06$
3. Continuous fit	0.8192	0.6682

The log-likelihood of the saturated MVN copula model is larger than the MVN copula with conditional independence model. This is because the saturated model involves more parameters. Notice that the two models are nested. Performing the likelihood ratio test (LRT) yields a p-value of 0.2385, which indicates that the MVN copula with conditional independence model is a better one for our data. For this reason, it is chosen.

Moreover, neither model recovers the discrete part of the data, as indicated by the low p-values from the KL test. However, the PIT test indicates that the continuous part of the data is represented well by both models.

3.2.2 The Copula-Vine Model

Next we consider the copula-vine model. To construct such a model, a regular vine structure is needed. With ten variables, there are $487,049,291,366,400$ possible regular vines. As in Chapter 2, we choose to keep the purely discrete and purely continuous parts of the model as sub-vines of the full vine. The discrete and continuous variables are ordered as: CS, CT, WM, WT, RH, OV and WD, LC, LAT, REP, respectively. To merge the two sub-vines, we choose to construct a D-vine structure as presented in Figure 3.13. The choice is motivated by the observed correlations between pairs of variables in the data and the graphical simplicity of a D-vine model.

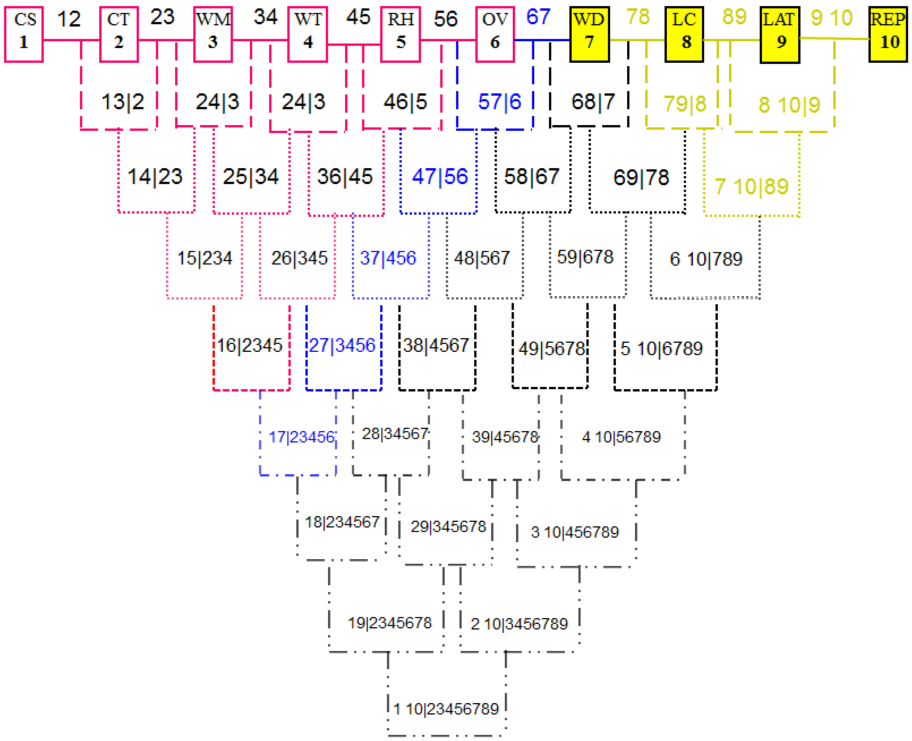


Figure 3.13: *The Copula-Vine TC Disruption Length Model.*

Each edge in the vine structure is modelled with a bivariate normal copula. Non-constant conditional copulas are considered in the pairs with conditioning sets containing only discrete variables. Otherwise, constant conditional copulas are fitted. The model's parameters are estimated sequentially using Algorithm 2.1 presented in Chapter 2.

Table C.3 in Appendix C presents the estimated parameters. We observe that many of these can be set to zero. On the other hand, the parameters of some pairs conditioned on different values of conditioning variables can differ quite significantly as well. For instance, the parameters of the four normal copulas fitted to the pair $(CT, RH|WM, WT)$ are $\rho_{CT,RH|WM=0,WT=0} = -0.2112$, $\rho_{CT,RH|WM=0,WT=1} = 0$, $\rho_{CT,RH|WM=1,WT=0} = 0.4360$ and $\rho_{CT,RH|WM=1,WT=1} = 0$.

Note that, because CS has six states, the copula-vine model is not guaranteed to recover the discrete part of the data. However, the p-value of the KL test is 0.0548, which is still above the significance level of 5%. This indicates that the discrete variables are still represented reasonably well by the copula-vine model. The p-value of the KL test is 0.2863 when we only consider the five Bernoulli variables.

Performing the PIT Test yields the smallest p-value of 0.8045, indicating a good fit of the normal copula in the continuous part of the data.

In this section, two TC disruption length models have been constructed. To determine which we will use in practice, in the following section the two are compared and validated.

3.3 Model Comparison and Validation

To compare the two models, we use the same statistics as in the experiments in subsection 2.3.2.

The disruption-length model is going to be used to predict the length of the latency and repair times. In this section, therefore, we only test the models' performance in predicting these continuous variables, evaluating it at two stages mimicking the model use (to be discussed in the next chapter). At the first stage, the conditional distribution of the latency time is computed by conditioning the model on the factors influencing latency time, i.e. CT, WM, WT, RH, OV, WD, and LC. At the second stage, the conditional distribution of the repair time is computed by further conditioning the model on CS and LAT.

Table 3.4: *Comparing the MVN Copula and the copula-vine disruption length models.*

Model		MVN copula	Copula-vine
1. Log-likelihood		-6375.7219	-5689.0867
2. Discrete fit		$2.2119e-06$	0.0548
3. Continuous fit		0.6682	0.8045
Latency time prediction	4. Cond. dist.	0.7754	0.2745
	5a. $RMSE$	29.1745	29.2014
	5b. R^2	0.0516	0.0498
Repair time prediction	6. Cond. dist.	0.3676	0.1978
	7a. $RMSE$	87.2795	87.4103
	7b. R^2	0.0953	0.0926

Table 3.4 summarizes the result. It is shown that the copula-vine model fits the data better, as shown by the larger log-likelihood. This observation is confirmed by the Vuong and Clarke's tests as well. The test statistics Z and B of the two tests are 12.6276 and 848, respectively, well outside their respective confidence bounds of $(1.96, 1.96)$ and $(917, 1003)$. Moreover, the copula-vine model represents the discrete part of the data well, while the MVN copula model does not. However, the PIT test indicates that the continuous part of the model is recovered in both cases.

In terms of predicting the latency and repair times, the performances of both models are similar. They recover the conditional distributions of the latency and repair times with $RMSE$ and R^2 values that are very much alike. Note, however, that the coefficient of determination R^2 values are not very high. This illustrates the complex nature of railway disruption length in the Netherlands and the limited availability of data. Additional influencing factors could be considered when new, better data is collected.

While Table 3.4 presents the results of model validation with the training set, we are also interested in observing the performance with a set of test data. This is provided by the corpus of TC problems on the Dutch railway network between 1 May 2014 and 31 October 2014. A total of 339 urgent incidents were recorded within this six-month period. Just as in the training set, the causes of approximately 30% of the TC problems in the test set are unknown. For the validation, the unknown samples are discarded from the test set and so the test is performed only on the 247 samples with known causes. Table 3.5 summarizes the result.

Table 3.5: *Comparing the MVN copula and the copula-vine disruption-length models with the test set.*

Model		MVN copula	Copula-vine
1. Log-likelihood		-973.9301	-821.5576
2. Discrete fit		0.0025	0.2817
3. Continuous fit		0.6252	0.7678
Latency time prediction	4. Cond. dist.	0.2470	0.1875
	5a. $RMSE$	19.8914	19.9238
	5b. R^2	0.0522	0.0491
Repair time prediction	6. Cond. dist.	0.2014	0.1978
	7a. $RMSE$	100.1521	101.1752
	7b. R^2	0.1613	0.1441

As in the training set, the copula-vine model yields the higher log-likelihood for the test set. The Vuong and Clarke's tests also indicate that this model represents the test data better, as shown by their test statistics Z and B of 8.3665 and 75, respectively, which are well outside the respective confidence bounds of $(1.96, 1.96)$ and $(108, 139)$. From the KL test, we conclude that the MVN copula model does not represent the discrete part of the data while the copula-vine model does. However, both recover the continuous part of the data well.

The performance of the models in predicting latency and repair times is also similar where the conditional distributions are recovered by both of them. However, the coefficient of determination values of the two models are not very high.

Another popular method which can be implemented to predict the latency or repair time is the generalized linear model (GLM). This, however, assumes that the conditional distributions of the latency and repair times come from distributions in the exponential family. We constructed the GLM with the gamma link[6] to compare its performance with the MVN copula and copula-vine models'. As in Chapter 2, we used the built-in function `fitglm` in `MATLAB` to obtain the GLM. We considered up to two-way interactions, while parameters found to be insignificant were removed from the model. Table 3.6 summarizes the result for both the training and test set.

---

[6] The gamma link was chosen because the gamma distribution best resembles the latency and repair times, even though the KS test rejects the null hypothesis that it does so.

Table 3.6: *Prediction performance of the GLM.*

Dataset		Training	Test
Latency time prediction	4. Cond. dist.	10^{-100}	10^{-15}
	5a. $RMSE$	29.3053	20.0036
	5b. R^2	0.0426	0.0414
Repair time prediction	6. Cond. dist.	10^{-34}	10^{-7}
	7a. $RMSE$	86.8358	101.9422
	7b. R^2	0.1045	0.1311

Table 3.6 shows that, in terms of $RMSE$, the GLM produces results similar to the MVN copula and copula-vine models. Unlike them, however, it does not recover the conditional distributions of the latency and repair times.

In this section we have compared the MVN copula and copula-vine models in terms of model fit and their performance in predicting the latency and repair times for the TC disruptions. The copula-vine model represents the joint data better, as indicated by the higher log-likelihood and recovery of both the discrete and continuous parts of the model. In terms of predicting disruption length, however, the performances of the two models are similar. A comparison with the popular GLM has also been performed, from which we observe that the $RMSE$ and R^2 of all models are similar. However, the GLM does not recover the conditional distribution of the latency and repair times, while the MVN copula and copula-vine models do.

In the next section we briefly present the disruption-length model construction for switch failures.

3.4 The Switch Disruption-Length Model

3.4.1 Influencing Factors

The factors influencing the latency and repair times of a switch disruption are, in general, the same as those in the TC disruption-length model. There are two exceptions, however: the variables temperature and cause. A TC is sensitive to warm temperatures, but a switch is sensitive to cold temperatures. The presence of ice or snow can block the switching process (see subsection 1.1.1.2). The variable "cause" needs to be defined using the possible switch problems listed in subsection 1.1.1.2. In this subsection, we present the data analysis of these two variables.

3.4.1.1 Cold Temperature

As in TC disruption, a threshold needs to be defined to distinguish when the temperature is said to be "cold" enough. As before, we consider several different thresholds. The latency times are divided into two groups, one corresponding to “not cold” (above the threshold, indicated as 0) and one to “cold” (indicated as 1) for each threshold. The difference between the groups is measured by the two-sample KS and CvM tests, with the result presented in Table 3.7.

Table 3.7: *P-values of the KS and CvM tests for latency time given different low-temperature thresholds.*

Threshold	KS	CvM	Threshold	KS	CvM
-6^oC	$8.5989e-04$	$9.3402e-03$	0^o**C**	$\mathbf{1.3028e-11}$	$\leq 10^{-5}$
-5^oC	$1.0469e-05$	$\leq 10^{-5}$	1^oC	$7.0447e-10$	$\leq 10^{-5}$
-4^oC	$1.3438e-06$	$\leq 10^{-5}$	2^oC	$1.4642e-09$	$\leq 10^{-5}$
-3^oC	$1.3265e-07$	$\leq 10^{-5}$	3^oC	$5.2788e-10$	$\leq 10^{-5}$
-2^oC	$8.7756e-11$	$\leq 10^{-5}$	4^oC	$3.3253e-09$	$\leq 10^{-5}$
-1^oC	$5.8676e-10$	$\leq 10^{-5}$	5^oC	$2.8887e-10$	$\leq 10^{-5}$

The lowest p-value is observed when the threshold is 0^oC, so this is the one chosen for the switch disruption-length model. Table 3.7 shows that cold temperatures directly influence latency times.

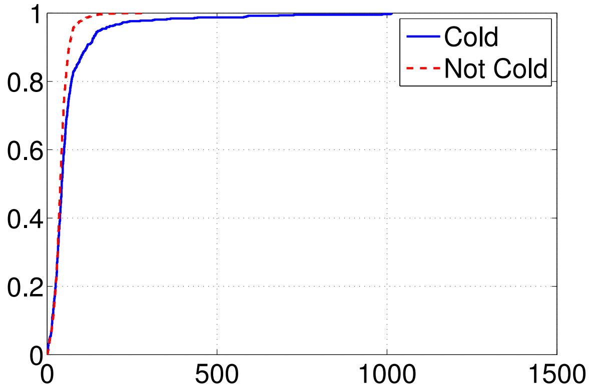


Figure 3.14: *The empirical distribution of latency time with respect to the cold temperature.*

Note that the binary variable “cold” directly affects the latency time. With the threshold of 0^oC, on average the latency time is longer by 23.1253 minutes when the temperature is cold. This is reasonable, because road traffic is generally slower when temperatures are very low. Figure 3.14 presents the two latency-time distributions corresponding to the two states of cold temperature

3.4.1.2 Causes of Switch Failures

Based on the types of switch failure described in subsubsection 1.1.1.2, the causes of switch disruptions are grouped as follows:

1. Group 1: failures in the controlling process (NIC-5 and NIC-TB-C).

2. Group 2: point machine problems (NIC-2 and NIC-4).

3. Group 3: failures in the steering process (NIC-1 and NIC-TB-S).

4. Group 4: blockages (NIC-3A and NIC-3B).

5. Other.

The first four groups are ordered based on the length of repair time. Switch failures in Group 1 are usually hard to detect and require the repair team to replace broken instruments. When the point machine breaks down (Group 2), it or some of its components need to be replaced. Failures in Group 3 usually require the replacement of small components, e.g. relays, fuses or wiring, which can be done quickly. Several problems in this group can also be solved by re-adjusting the troubled component to the right setting. The problems in Group 4 are solved by removing the blockage from the affected switch. After this has been done, however, the repair team must also test the switch to ensure that its performance has not been compromised.

Table 3.8: *Number of samples, proportion, mean and standard deviation of repair times in each cause group before and after redistribution. The information after redistribution is presented in brackets.*

Cause group	# samples	Proportion (%)	Mean	Std. dev
Group 1	340 (375)	17.07 (15.10)	83.17 (89.01)	82.00 (87.82)
Group 2	363 (377)	18.22 (15.18)	57.30 (59.20)	61.73 (66.68)
Other	78 (78)	3.92 (3.14)	51.82 (51.82)	85.79 (85.79)
Group 3	486 (613)	24.40 (24.68)	50.53 (45.83)	57.57 (57.33)
Group 4	725 (1041)	24.70 (41.91)	48.59 (45.11)	53.99 (47.83)
All	1992 (2484)	100 (100)	56.68 (54.26)	64.37 (63.80)

The failures in the first four groups are related to the "not in control" (NIC) situation, which is that most commonly observed for switch problems in our data. That also contains 78 non-NIC switch incidents, though, two examples being fractures of the frogs and the displacement of the switch's substructure.

The cause of approximately 20% of switch incidents are not known due to unclear descriptions. As in the TC case, we redistribute these "unknowns" to one of the five groups using the "Bayesian classifier" technique. Table 3.8 presents some information of the repair times of each group before and after redistribution. In general, the "unknown" incidents tend to have shorter repair times. Consequently, most are redistributed into Group 4.

Figure 3.15 shows the empirical distributions of repair times for the five groups. The differences between the groups is less evident than with TC disruptions, especially for those with shorter repair times. This indicates the complexity of switch disruptions and the need to identify more predictive influencing factors in respect of repair time in order to to construct a more robust model.

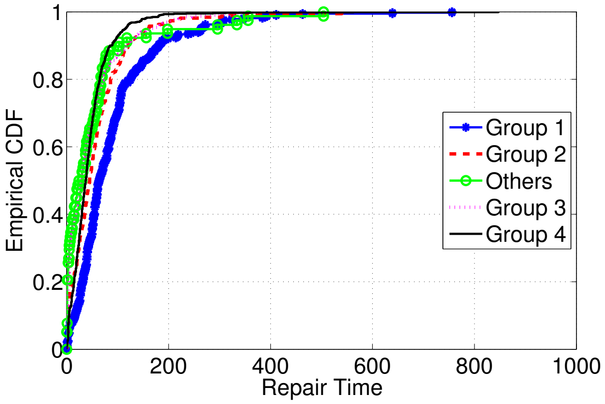


Figure 3.15: *Empirical distributions of the repair times of the five groups.*

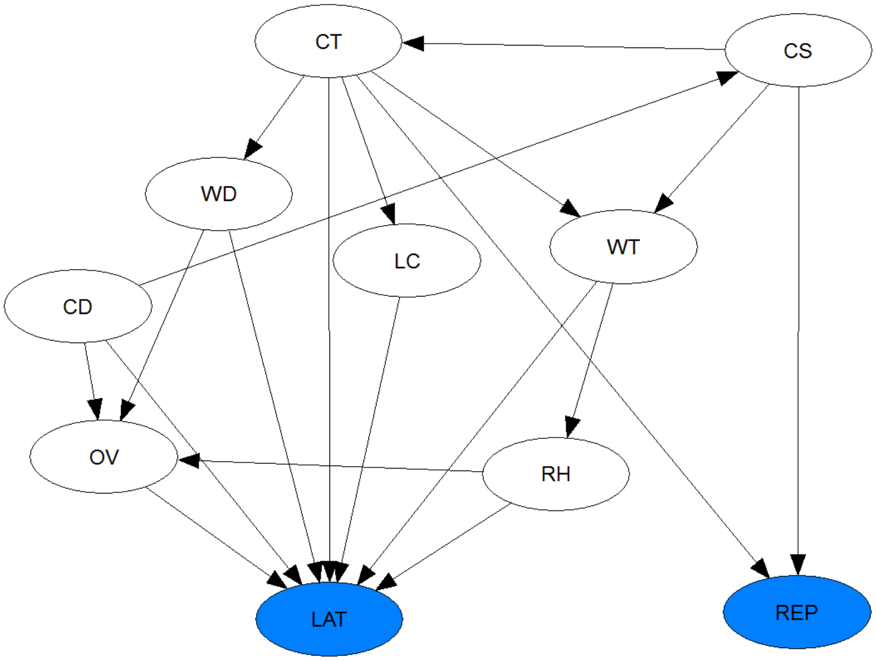


Figure 3.16: *The Switch BN.*

3.4.2 The Model

For the TC disruptions, we have learned that the copula-vine model fits the joint distribution of the ten variables better than the MVN copula model. Nonetheless, both recover the conditional distributions of the latency and repair times with similar performance. The main purpose of the disruption-length model is to produce a disruption-length prediction given the information known about the other variables. From this point of view, the MVN copula and the copula-vine models are equally attractive. In this thesis, we choose to construct the final switch disruption-length model with the MVN copula so that we can implement it in `UNINET`, since this software's fast computational time makes its use more practical.

To construct a parsimonious MVN copula model, some conditional independencies represented as a BN are included. Figure 3.16 presents the chosen BN structure of the switch disruption-length model.

In general, the BN structure is similar to the TC BN structure in Figure 3.10. But there are a few differences. For instance, the variable "cold", denoted as CD, influences the latency time (LAT) and cause (CS). From the data analysis in subsection 3.4.1.1, it is clear that LAT is directly affected by CD because road traffic is generally slower when the temperature is cold. CD also affects CS because sub-zero temperatures increase the chance of ice or snow formation, which might block the switch.

As in the TC MVN copula model, the model's parameters are computed using the maximum-likelihood approach. Table C.4 in Appendix C presents the result.

The test data for the switch disruption-length model comes from the SAP database and covers the period between 1 October 2014 and 31 March 2015. Within those six months, 626 urgent switch incidents were recorded. The causes of 415 of these are known. Our test data consists of these 415 incidents.

Table 3.9: *Performance of the MVN copula model for the training and test sets of switch disruptions.*

Dataset		Training	Test
1. Log-likelihood		-9819.5153	-1654.8125
2. Discrete fit		$3.8013e-06$	0.0011
3. Continuous fit		0.4768	0.3215
Latency time prediction	4. Cond. dist.	0.1242	0.1726
	5a. $RMSE$	52.6300	54.0060
	5b. R^2	0.0401	0.0501
Repair time prediction	6. Cond. dist.	0.2735	0.2121
	7a. $RMSE$	62.0746	83.8952
	7b. R^2	0.0534	0.0632

Table 3.9 presents the performance of the model for the training and test sets. We can see that the MVN copula model recovers the continuous part of the data, but a misfit occurs in the discrete part. However, the model does recover the conditional distribution of the latency and repair times for both the training and the test set. However, the coefficient of determination is lower for the switch model than for the TC model. This indicates that better information regarding switch incidents is needed in order to construct a more robust disruption-length prediction model.

3.5 Chapter Summary

After analysing the SAP database, eight influencing factors are included in the railway disruption-length model, along with the latency and repair times. Six of these variables are discrete, with five of them Bernoulli, while four are continuous. The railway disruption-length model is the joint distribution between these ten variables. We consider two strategies to construct the joint distribution: (i) using the MVN copula; and (ii) using the copula-vine approach.

The copula-vine model fits the data better, as indicated by the higher log-likelihood. Moreover, it recovers the discrete part of the data while the MVN copula model does not. In terms of prediction, however, the performances of both models appear to be similar where the conditional distributions of the latency and repair times are recovered. From this point of view, the two models are equally attractive. With that in mind, we choose to use the MVN copula option as the disruption-length model since it can be implemented in `UNINET`, which has very fast computational time. This makes the model more practical from the application point of view.

Unfortunately, the coefficient of determination R^2 is low in all models. This indicates the complexity of railway disruption situations in the Netherlands where a lot of parties are involved. An additional cause is the lack of available data. This conveys a strong message to ProRail about the necessity and urgency of collecting better information. With that can the models be improved. Recommendations about what improved information and data to collect are given in Chapter 5.

In the next chapter, we demonstrate the use of the disruption-length model in practice in collaboration with the Department of Transport and Planning at Delft University of Technology. The disruption-length model is applied together with the train short-turning model and the passenger-flow model to a disruption to train traffic in the vicinity of Houten, the Netherlands. The results of this collaboration are presented.

CHAPTER 4

The Model in Practice[1]

The outputs of the disruption-length model constructed in Chapter 3 are the conditional distributions of the latency and repair times, from which the disruption-length predictions are made. The conditional distributions are obtained by conditioning the model on the observed values of the influencing factors. We start this chapter by showing how the conditionalization can be performed and how it can be used in the disruption-response process in the Netherlands.

The prediction of disruption length is used to optimize train traffic during the disruption. However, due to the complex nature of railway disruption in the Netherlands, it is difficult to create an optimization algorithm with stochastic inputs. This means that one value of disruption-length prediction needs to be chosen from the conditional distributions.

In practice, either the mean or median of the conditional distributions is generally chosen as the prediction because these, respectively, minimize the $RMSE$ and the absolute error of prediction. In our case, however, the chosen prediction affects the train traffic and the passengers – the end customers of ProRail. A prediction that is optimistic (too short) might have a different impact on customers than one which is pessimistic (too long). In applying the model, then, it is important to study the impact of different choices of prediction on train traffic and passengers.

This chapter is concerned with this issue. The disruption-length model is used in a collaboration with the train short-turning model and the passenger-flow model developed by two PhD candidates in the Department of Transport and Planning at Delft University of Technology. In the collaboration, an actual disruption to train traffic in the vicinity of Houten, the Netherlands is considered.

---

[1] This chapter is based on Ghaemi et al. [2016b].

4.1 Predicting Railway Disruption Lengths with the Models

In this section, we show how the MVN copula model can be used in real-time practice in the disruption-response procedure in the Netherlands.

Consider a TC incident. We have the unconditional model which covers all historical TC disruptions in the database. The unconditional TC BN model is shown in Figure 4.1. This is the same BN as presented in Figure 3.10, but now showing the marginal distributions of the variables, along with their corresponding means and standard deviations at the bottom of the nodes. Note that an eleventh node, disruption length, has been added to the BN. This is a *functional* node that is defined as the sum of latency and repair times.

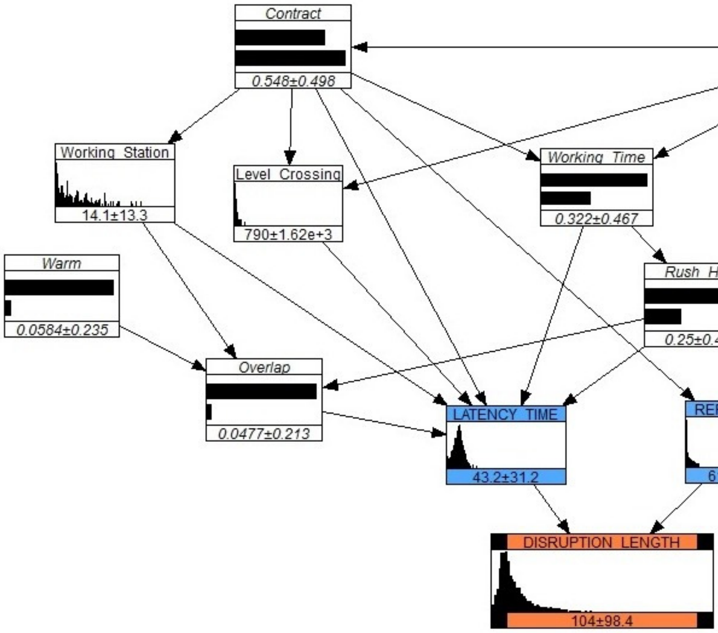


Figure 4.1: *The unconditional TC BN.*

At this point we can already see a decision-making problem that arises in providing the prediction: which value of the disruption-length distribution should we use? One option is to take its mean which, in this case, is 104 minutes. This is essentially what is done in current practice when defining the disruption-length prediction “P1”. Because the distribution is left-skewed, the mean is greater than its 50% quantile (the median). In fact, the mean of 104 minutes corresponds to the 67% quantile of the distribution. Another option is to choose the median, 72 minutes in this case, as the prediction.

When a disruption actually starts, we know, amongst other things, where and when it is happening. The unconditional BN in Figure 4.1 can be conditionalized on this information to obtain a “P1” prediction for the incident. Once the repair team arrives at the disruption site and the cause is known, the model can be further conditionalized to obtain the “P2” prediction.

To illustrate how to conditionalize the model, let us consider a real TC incident in the Netherlands. This occurred in the vicinity of Houten, a town to the south-east of Utrecht, on Thursday, 10 July 2014, starting at 14:22. Its basic characteristics were as follows.

1. Contract type: OPC.
2. Working station (maintenance base) distance: 7.1620 kilometres.
3. Level crossing distance: 872.372 metres.
4. Working (contractual) time: yes.
5. Warm: yes.
6. Rush hour: no.
7. Presence of an overlapping disruption: no.
8. Cause: a setting problem caused by heat.

The real observed latency and repair time were 70 and 73 minutes, respectively.

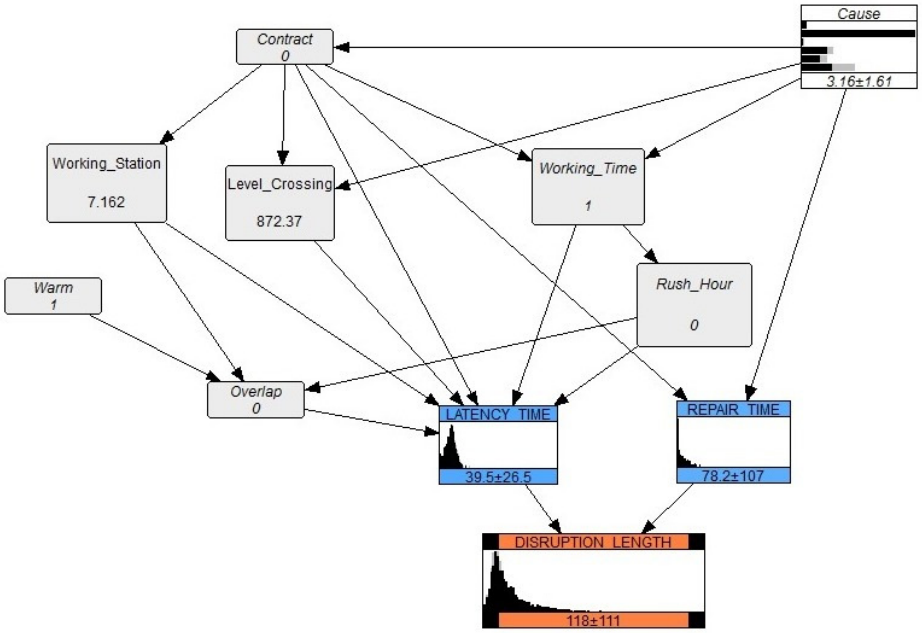


Figure 4.2: *The TC BN conditioned on the latency time’s influencing factors.*

The actual length of the disruption, therefore, was 70 + 73 = 143 minutes. This corresponds to the 78.75% quantile of the disruption-length distribution in the BN in Figure 4.1. In this case, choosing either the mean or the median of the distribution as the prediction underestimates the actual disruption length.

When the disruption starts at 14:22, only information about the factors influencing the latency time is known. The BN is updated by conditioning the model on this information. This results in the conditioned BN presented in Figure 4.2. It can be seen that the model adjusts itself to the situation in hand. The disruption length is predicted to be longer than in the unconditional case, which has a mean of 118 minutes and a median of 81 minutes. However, the latency time is predicted to be shorter than average. The mean of latency time decreases from 43.2 to 39.5 minutes, while the median falls from 40 to 38 minutes. Because the contract type is OPC with the nearest level crossing relatively far away (a common characteristic of the Randstad region), the probability that the disruption is caused by coins decreases from 23% to 16%. As a result, the predicted mean repair time increases from 61.3 to 78.3 minutes (and the median increases from 26 to 39 minutes). The predictions obtained with this conditional BN are called the "P1" predictions.

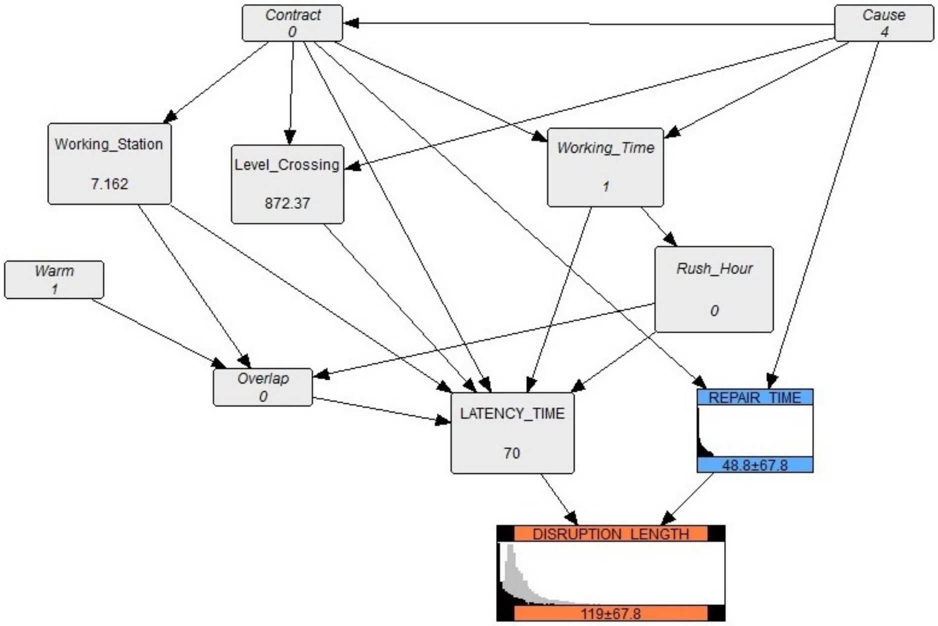


Figure 4.3: *The TC BN conditioned on all influencing factors.*

The repair team arrives at the site at 15:32, 70 minutes after the disruption starts. After investigating the incident, the problem with setting (Cause Group 4) is found. The model is further conditioned on the new information and the BN is updated, as presented in Figure 4.3. The model again adjusts itself to the situation. The disruption-length prediction is updated to 119 minutes (mean) and 95 minutes (median). The mean repair time is also updated, to 48.8 minutes, and the median to 25 minutes. The predictions made with this BN are called the "P2" predictions.

In practice, conditionalization on the variable "cause" can only be performed after the repair team has diagnosed the problem and found its cause. The time needed to do this is called the "diagnosis time". However, the actual diagnosis time is not available in the dataset; it is included under the "repair time". This needs to be taken into account in practice.

In the disruption response process, the repair team is given 15 minutes to diagnose the problem after it arrives at the site (see subsection 1.1.2). For this reason, in this chapter we assume that diagnosis always takes 15 minutes and that the cause of the problem is always known after this time. In other words, the conditional BN as in Figure 4.3 is obtained 15 minutes after the repair team's actual arrival time. Consequently, the disruption-length prediction $P2$ at this stage is taken to be:

$$P2 = lat + \max(15, r\hat{e}p)$$

where lat corresponds to the actual latency time and $r\hat{e}p$ represents the chosen repair time length prediction from the conditional distribution of repair time.

To investigate the effect of the uncertainty in disruption length on train traffic and passengers, the disruption-length model is combined with the short-turning model and the passenger-flow model. These are developed by two PhD candidates in the Department of Transport and Planning of Delft University of Technology. The two models are discussed briefly in the next section.

4.2 The Short-Turning and Passenger-Flow Models

4.2.1 The Short-Turning Model

In railway traffic management, a train is "short-turned" when it does not operate along the full length of its assigned route, but instead reverses at a terminus (usually a station) short of its intended destination. Ghaemi et al. [2016a] develop the short-turning model to reduce train-delay propagation in the event of a disruption that completely blocks a section of railway. The short-turned trains in the model replace the planned services in the opposite direction, which can no longer be operated by the trains originally assigned to them due to the disruption. The model considers the possibilities for short-turning trains at a couple of stations before the disrupted section, and for cancelling some services.

Figure 4.4 presents the time-distance diagram illustrating the possible short-turning patterns at stations a and a' due to a complete blockage between stations a and b. Each service is denoted as $v^i_{l,n}$, where i indicates the order of service on the operational line l and n refers to the unique train number. In Figure 4.4, the red arcs represent the short-turning possibilities of the train $v^i_{l,n}$. This train

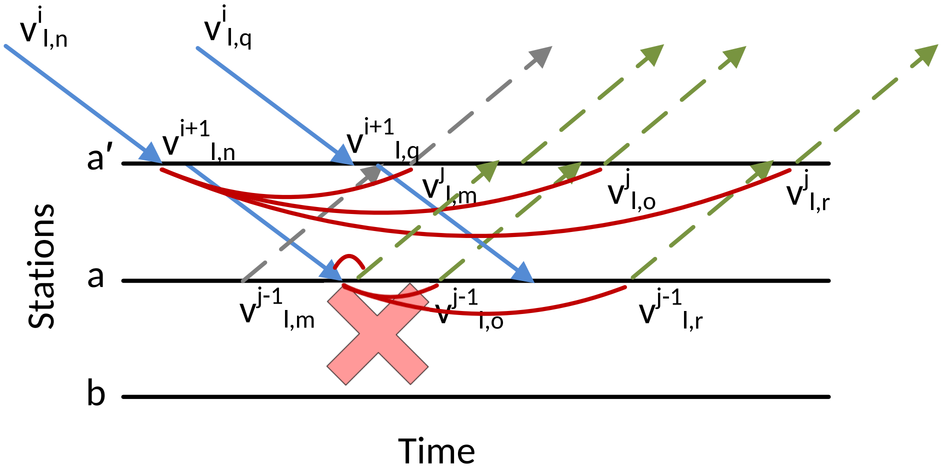


Figure 4.4: *Several possible short-turning patterns at stations a′ and a. Source: Ghaemi et al. [2016a].*

can run to station a as service $v^{i+1}_{l,n}$ then short-turn there and continue as either service $v^{j-1}_{l,m}$ (with a delay), $v^{j-1}_{l,o}$, or $v^{j-1}_{l,r}$. Alternatively, it can short-turn at station a' and continue as service $v^{j}_{l,m}$, $v^{j}_{l,o}$, or $v^{j}_{l,r}$. In this case, however, service $v^{i+1}_{l,n}$ between a' and a and its short-turning patterns at station a need to be cancelled.

The objective function OF of the short-turning model is as follows:

$$OF = \sum_{v^i_{l,n} \in \mathbf{\Gamma}} (w^c_{v^i_{l,n}} \cdot c_{v^i_{l,n}} + w^{d^d}_{v^i_{l,n}} \cdot d^d_{v^i_{l,n}} + w^{d^a}_{v^i_{l,n}} \cdot d^a_{v^i_{l,n}}). \tag{4.1}$$

In (4.1), $\mathbf{\Gamma}$ is the set of all scheduled services, $w^c_{v^i_{l,n}}$, $w^d_{v^i_{l,n}}$, and $w^a_{v^i_{l,n}}$ denote the penalties for the cancellation, departure delay, and arrival delay of train $v^i_{l,n}$, respectively, $c_{v^i_{l,n}}$ is an indicator whether train $v^i_{l,n}$ is cancelled (value 1) or not (value 0), $d^d_{v^i_{l,n}}$ corresponds to the departure delay of service $v^i_{l,n}$ (in seconds), and $d^a_{v^i_{l,n}}$ corresponds to the arrival delay of service $v^i_{l,n}$ (in seconds). Ghaemi et al. [2016a] choose a cancellation penalty of $w^c_{v^i_{l,n}} = 1000$ (in seconds) which corresponds to the frequency of the train service between Houten and Geldermalsen (16 minutes or 960 seconds)[2]. With every cancelled train, a passenger needs to wait for the next service and thus incurs a delay of 16 minutes. The departure and arrival delay penalty of $w^d_{v^i_{l,n}} = w^a_{v^i_{l,n}} = 1$ has been chosen because the delay time is what the passenger experiences. In this thesis, the same penalties are used.

---

[2] This is the area in the Netherlands which we consider in our experiment. In other locations, the cancellation penalty might be different.

The objective function OF represents the total (weighted) train delay. The first term in (4.1) corresponds to the delay caused by cancelling a train, while the second and third terms are the departure and arrival delay, respectively. The short-turning model considers all possible short-turning patterns and chooses the one which minimizes OF. This pattern corresponds to a "disruption timetable" containing the departure and arrival time of each train service. Interested readers are referred to Ghaemi et al. [2016a] for more details about this model.

4.2.2 The Passenger-Flow Model

The passenger traffic between two stations, designated as the origin-destination (OD) pair, is determined using the passenger-flow model. For the passengers arriving at an origin station at any given minute with plans to travel to a particular destination station, the model considers a number of route alternatives they can take based on the current timetable (Ghaemi et al. [2016b]). For each passenger, each alternative i has its own "cost". This is the generalized travel time, T_i, defined as follows:

$$T_i = \beta_w \cdot t_w + \sum_{a=1}^{N_{tr}+1} \beta_{in} \cdot t_{in}^a + \sum_{b=1}^{N_{tr}} \beta_{tr} \cdot t_{tr}^b + \beta_{ntr} \cdot N_{tr} + \beta_{re} \cdot N_{re}. \tag{4.2}$$

In equation (4.2), t_w, t_{in}, and t_{tr} are the waiting time, in-vehicle time, and transfer time (all in minutes) with weights β_w, β_{in}, and β_{tr}, respectively, N_{tr} denotes the number of transfers with transfer penalty β_{ntr} (in minutes), and N_{re} represents the number of reroutings[3] with penalty β_{re} (in minutes). In this thesis, the weights are $\beta_w = 2$, $\beta_{in} = 1$, and $\beta_{tr} = 2$, while the penalties are $\beta_{ntr} = 5$ (in minutes) and $\beta_{re} = 10$ (in minutes). These values are based on the works of Wardman [2004], Balcombe et al. [2004], and Ghaemi et al. [2016b]. This means that the passengers are assumed to prefer travelling in a train to waiting or transferring. Moreover, they endure more discomfort on a route with a lot of transfers and when there are many timetable changes.

The generalized travel time in equation (4.2) represents the total (weighted) travel time of a passenger travelling between the two OD stations with route alternative i. The first and third terms in (4.2) are the weighted waiting and transfer times, respectively. The second term corresponds to the total weighted time the passenger spends in the trains. The fourth and fifth terms represent the "inconvenience", measured in minutes, caused by the number of transfers and reroutings, respectively.

In the model, not all passengers choose the alternative with the shortest generalized travel time. Following the work of Cats et al. [2016], the proportion of passengers choosing alternative i, denoted as P_i, is determined using the logit model as:

$$P_i = \frac{\exp(-T_i)}{\sum_i \exp(-T_i)}. \tag{4.3}$$

---

[3] The number of times passengers need to change their plans due to a timetable change.

This means that the alternative with the shortest generalized travel time receives the highest proportion of passengers.

The model follows passengers until they arrive at their destination. While travelling, they might need to adjust their route due to one or more changes in the timetable. Every time this happens, the model recomputes the generalized travel time based on the new timetable and the passenger flow is adjusted accordingly.

For more details about the passenger-flow model, interested readers are referred to Ghaemi et al. [2016b].

4.2.3 Interaction Between the Three Models

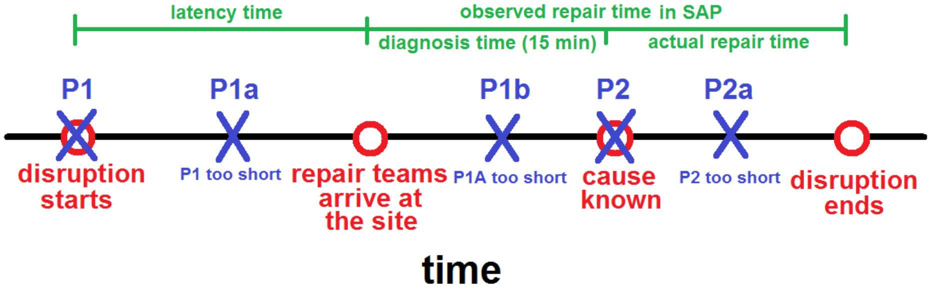


Figure 4.5: *Time diagram of a railway disruption.*

The three models interact in a dynamic fashion, i.e. interaction occurs every time new information becomes available. This can come in the form of a value of an observed influencing factor in the disruption-length model, or we may learn that the previous disruption-length prediction was too short. The crosses in the time diagram in Figure 4.5 illustrate the points at which interaction occurs during the disruption period.

When information about the influencing factors becomes available, the disruption-length model is conditionalized as illustrated in Section 4.1. In Figure 4.5, this occurs at the first and fourth crosses in the diagram.

When the prediction is too short, the disruption is still unresolved even after its predicted end time. This situation occurs when the prediction is too "optimistic", i.e. the chosen quantile of the conditional distribution of disruption length is too low for the case. If this happens, the prediction is updated by computing a new distribution of disruption length conditional on the information that the disruption is longer than the previous prediction. This is done by sampling the original conditional distribution on the quantiles higher than the prediction. In this thesis, these "revised" predictions are denoted alphabetically in chronological order. For instance, a "P1" prediction is updated to "P1a", "P1b", "P1c", and so on.

Figure 4.6 illustrates the interaction. Every time new information becomes available, a new disruption-length prediction is made. With each new pre-

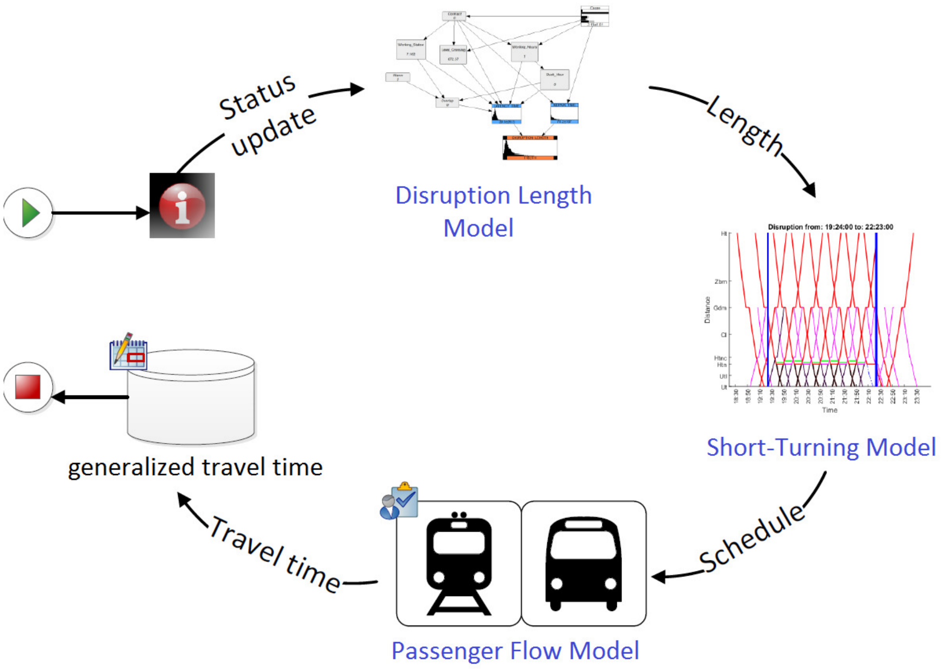


Figure 4.6: *Flow of interactions between the three models every time new information is available.*

diction, the short-turning model recalculates the disrupted timetable and the passenger-flow model adjusts passenger movements accordingly.

4.3 The Experiment

This section describes an experiment set up to observe a disruption on the railway in the vicinity of Houten, the Netherlands. The disruption-length model, the short-turning model and the passenger flow model are combined to manage train traffic and passenger flows during this disruption. After first describing how the experiment is conducted, we then present the results of a number of case studies considered as part of it.

4.3.1 Experimental Setup

In the experiment, a complete blockage occurs on the stretch of railway between Utrecht and Houten. Figure 4.7 presents a map of the disruption site. The blockage is caused by a TC disruption.

Different scenarios for disruption-length predictions are considered, corresponding to different quantiles of the conditional distribution. In each case when a prediction is updated, the new version has the value of the new conditional

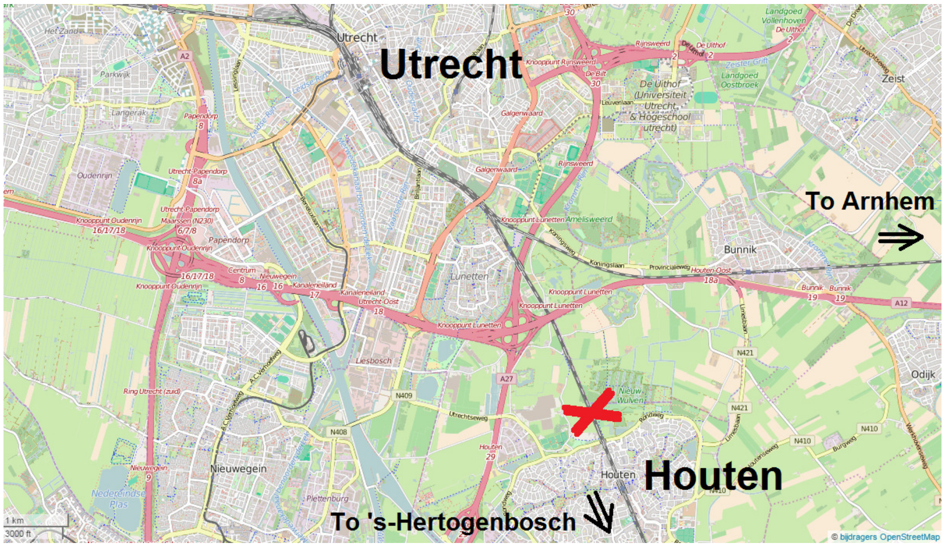


Figure 4.7: *The disruption site. Map source:* `http://www.openstreetmap.org/`.

distribution corresponding to the same quantile as the previous prediction. For instance, when a prediction corresponding to the quantile 50 (median) is updated, the new prediction is also taken to be the median of the updated conditional distribution. In principle, this does not need to be the case, but this choice is made to narrow down the number of possible combinations of predictions.

Note that a “P1” prediction is valid only until new information regarding the cause of the incident becomes available, at which point an updated “P2” prediction can be made. This means that the “P1” predictions and the resulting disruption timetables are used only until 15 minutes, at most, after the end of the true latency time, as discussed in section 4.1.

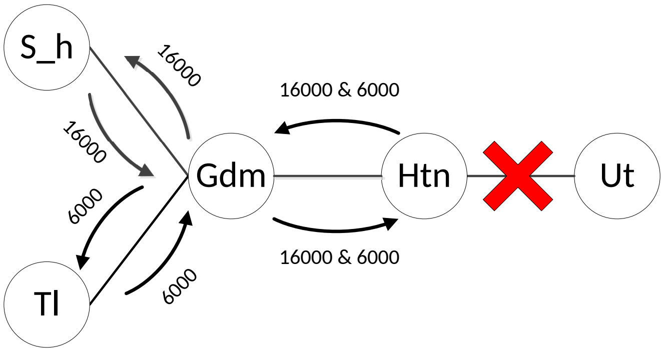


Figure 4.8: *Railway lines affected by the disruption. Source: Ghaemi et al. [2016a].*

The “P2” predictions are updated until the actual disruption ends. When this happens, we do not compute a new timetable and the railway section remains blocked until the predicted time of the end of the disruption. This choice is made to penalize a prediction that is too "pessimistic", i.e. one corresponding to a quantile of the conditional distribution of disruption length that is "too high".

For each predicted disruption length, the short-turning model computes the disruption timetable. This timetable also considers the recovery time needed after the predicted end of the disruption. During this period, the blocked section has been reopened for train operation but services have yet to return to normal. It takes a certain amount of time for traffic to recover fully, with all trains following the original timetable. Two railway lines are considered by the model: line 16000 between Utrecht (Ut) and 's-Hertogenbosch (S_h) and line 6000 between Ut and Tiel (Tl)[4]. Both pass Houten (Htn) and Geldermalsen (Gdm) stations, before splitting just beyond Gdm, as illustrated in Figure 4.8. For each case study presented in the following subsection, the short-turning model needs about twenty minutes of computation time to produce the disruption timetables.

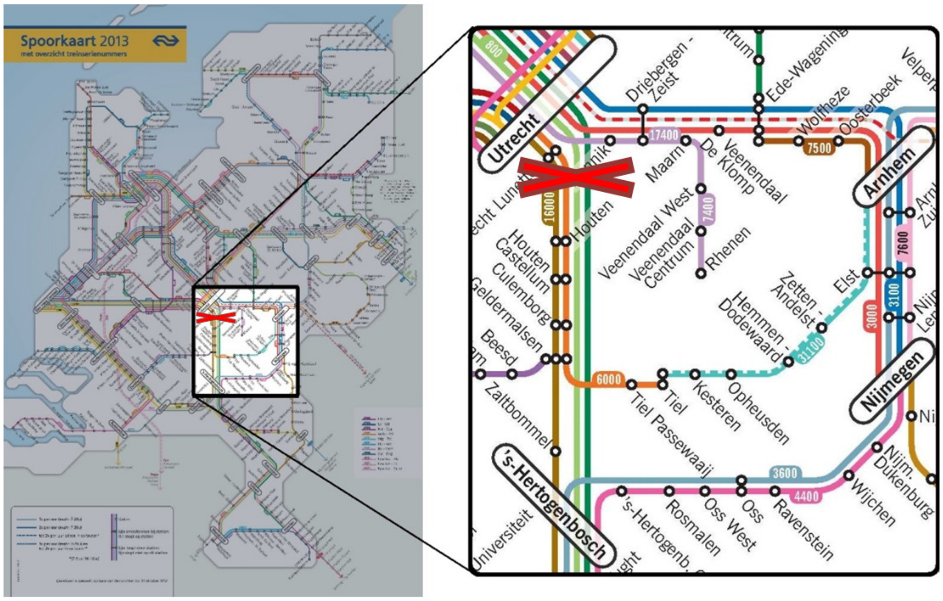


Figure 4.9: *The loop considered by the passenger-flow model in the experiment. Ilustration is adapted from the NS main train service map. Source: ProRail.*

With the disruption timetables, the passenger-flow model computes the passenger traffic. To reduce computational cost and simplify the problem, the model considers only those passengers with OD pairs included in the loop shown in Figure 4.9. Data about the number of daily passengers between all of these OD pairs is obtained from the Netherlands Railways (*Nederlandse Spoorwegen*, NS).

---

[4] Both lines are operated using Sprinter (local) trains.

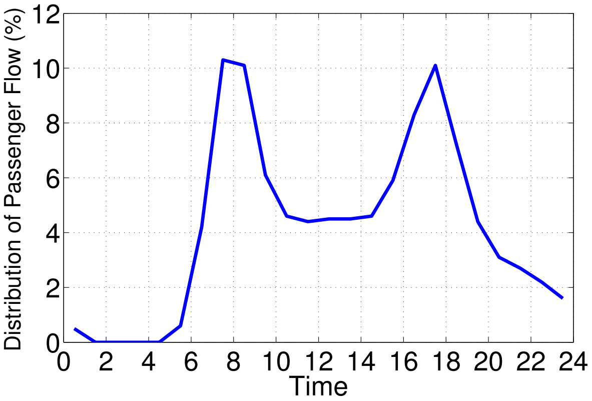


Figure 4.10: *The distribution of passengers throughout the day. Source: NS.*

The number of passengers varies throughout the day, following the distribution presented in Figure 4.10 – also obtained from NS. As we can see in the graph, no passengers travel between 01:00 and 04:00 and there are two peaks during the day: the morning and late afternoon rush hours. The passenger flows at all stations are assumed to follow this distribution.

Passengers travelling between Utrecht Centraal and Houten stations have two alternatives: to detour via Arnhem and Nijmegen or to take the public bus between the two stations. The travel time by bus between Utrecht Centraal and Houten is about 35 minutes, while a regular train service would have taken only 9 minutes. On the other hand, the detour via Arnhem and Nijmegen might not be very attractive due to the tremendous additional distance one needs to cover. For a passenger travelling to Houten from Utrecht, this detour takes almost 2 hours.

For each OD pair at every minute, the passenger-flow model computes the proportion of passengers taking each route alternative. It is possible for the timetable to be changed before some or all of the passengers arrive at their destination. In this case, the model recomputes the alternative routes and the proportions of passengers on these routes based on their current locations. The computational time needed by the passenger-flow model is very long: for the scenarios considered in the case studies presented in the next subsection, it takes about four hours to compute.

To compare the impact on passengers of different choices of prediction, in all scenarios the passenger-flow model is run for a fixed six hours period. This means that the same passengers are considered in every scenario. However, their travel patterns are affected by the different predictions and timetables. The impact on passengers in each scenario is measured using the following statistics:

1. The total number of passengers travelling up until the last P2 prediction.

2. The total generalized travel time corresponding to equation (4.2) of all passengers for all OD pairs considered in the experiment. This is the normal

situation with no disruption and the hypothetical (ideal) situation where the true disruption length is known from the beginning.

3. The total number of reroutings and transfers.

4.3.2 Case Studies

In this thesis, we present the result obtained from four case studies based on real TC disruptions observed between Utrecht and Houten.

4.3.2.1 Case Study 1

The first case study has been illustrated in Section 4.1. The disruption was due to heat causing the TC to be in the wrong setting. It occurred at 14:22 on a Thursday afternoon. In total, it took 143 minutes to resolve and so finished at 16:45.

Table 4.1 presents the "P1" predictions made from the conditional BN in Figure 4.2. The predictions are presented in terms of their length (in minutes) and the time the disruption is predicted to end.

Table 4.1: *P1 predictions for Case Study 1.*

Qtl	P1		P1a		P1b	
(%)	Len	Time	Len	Time	Len	Time
25	49	15:11	72	15:34	97	15:59
50	81	15:43	144	16:46		
75	143	16:45				
85	205	17:47				
90	254	18:36				
Mean	118	16:20				

The repair team arrives at the site at 15:32. After 15 minutes of diagnosis time, the cause of the TC failure is found and at 15:47 the "P1" predictions are updated to "P2" predictions with the conditional BN in Figure 4.3. These "P2" predictions are presented in Table 4.2.

Table 4.2: *P2 predictions for Case Study 1.*

Qtl	P2		P2a		P2b		P2c		P2d	
(%)	Len	Time	Len	Time	Len	Time	Len	Time	Len	Time
25	85	15:47	102	16:04	118	16:20	135	16:37	150	16:52
50	95	15:57	134	16:36	179	17:21				
75	133	16:35	240	18:22						
85	160	17:02								
90	194	17:36								
Mean	119	16:21	188	17:30						

The choice of the lower quantile as the prediction is undesirable because it is likely to be too optimistic, resulting in numerous revised predictions. This would result in too many revised timetables and so added inconvenience for passengers, who might be forced to adjust their travel plans repeatedly. Moreover, issuing a lot of revised predictions is impractical from the logistical point of view. In practice, every time a new prediction is made, the OCCR must also revise not only the pattern of train traffic but also rolling stock and crew assignments. For this reason, the scenarios with lower quantile predictions are not going to be considered at length here. In this experiment we consider only the 25% quantile and take it as representing such scenarios. Note that, in this case, in all eight predictions are generated during the period of disruption.

The short-turning model produces disruption timetables based on the predictions in Tables 4.1 and 4.2. Each timetable is used for as long as the underlying prediction remain "valid", i.e. it has not been changed. For instance, the disruption timetable generated from prediction P1a of the 50% quantile is followed for only four minutes, between 15:43 and 15:47. At 15:47 the prediction is updated to P2 and a new disruption timetable is generated.

Table 4.3 summarizes the outcome of the short-turning model for each quantile. It contains the number of cancelled services ($\#C$), the number of delayed services ($\#D$), the length of delay in minutes (del) and the number of short-turned services ($\#ST$) of the *last* prediction for each quantile.

Table 4.3: *Summary of the outcome of the short-turning model for Case Study 1.*

Qtl (%)	$\#C$	$\#D$	del	$\#ST$
25	40	0	0	10
50	48	0	0	12
75	64	0	0	16
85	44	0	0	11
90	52	0	0	13
Mean	52	0	0	13

Table 4.3 shows that the longer the line closure lasts, the more services are affected. The outcomes for the 90% quantile and the mean scenario are the same. This is because the final predictions of the two scenarios are similar, with a difference of only 6 minutes. Because a train runs in the corridor once in every 16 minutes, that small difference in the final prediction does not affect the outcome of the short-turning model. However, the short-turning model needs to be run twice in the 90% quantile scenario and three times in the mean scenario.

In this case study, the passenger-flow model is run for a fixed period of six hours for all scenarios. Within this period, 22,163 passengers travel in the affected area. The timetables generated by the short-turning model are used to model those travelling between every OD pair. The impact of the disruption and the different choices of disruption-length prediction on passenger travel times are measured, with the result presented in Table 4.4. The final row of figures in this table shows the benchmark case with the true disruption length. In this theoretical situation, the true end time of the disruption is already known when the disruption starts at 14:22[5]. Each scenario is compared with this case to measure

---

[5] This is the best possible situation but, of course, is not realistic in practice.

Table 4.4: *Impact on passengers of different predictions in Case Study 1.*

Qtl (%)	Excess (minutes)	# Affected Passengers	Inc. orig (%)	Inc. bench (%)	# Rerouting 1	# Rerouting 2	# Transfer
25	7	8948	17.7124	0.2640	595	4	3597
50	36	11469	20.5378	2.6705	466	1	4015
75	97	16560	24.3356	5.9054	247	0	4671
85	17	9791	17.5629	0.1366	5	0	3775
90	51	12854	20.1678	2.3554	0	0	4126
Mean	45	12299	20.3352	2.4980	282	0	4136
Real	0	8374	17.4026	0	0	0	3773

the increase in the impact of the prediction on passengers with respect to the ideal situation.

The second column of Table 4.4 shows the difference (in minutes) between the true end time of the disruption and the latest P2 prediction as to when the blocked section of line can reopen for train operation. The third column presents the total number of passengers travelling while the section is blocked. The fourth and fifth columns show the increases (in %) in total generalized travel times with respect to the normal situation without disruption and the benchmark, respectively. The sixth and seventh columns show the total number of passengers forced to reroute once or twice in each scenario. The total number of transfers performed by all passengers is given in the last column.

The benchmark case represents the best possible situation the one in which the lowest number of passengers are affected. Unsurprisingly, in this case the increase in generalized travel time with respect to the normal (no-disruption) situation is also the lowest. Moreover, no passengers have to reroute.

The number of transfers is lower in the 25%-quantile scenario than in the ideal one. However, the increase in generalized travel time is higher in the former than in the latter. This indicates that more passengers in the 25%-quantile scenario than in the ideal case choose the alternative of the slower route with the smallest number of transfers. This illustrates the complex nature of the passenger flow in the model.

The greater the difference between the true end time of disruption and the latest P2 prediction, the more passengers are affected. However, this does not always translate into a higher total generalized travel time. Note that the increase in generalized travel time is higher in the 25%-quantile scenario than in the 85%-quantile scenario, even though the difference between the prediction and the actual end time is only 7 minutes in the former and 17 minutes in the latter. The eight predictions in the 25%-quantile scenario cause many passengers to reroute due to the frequent updates to the disruption timetable. Consequently, its total generalized travel time rises substantially. In the 85%-quantile scenario, far fewer passengers need to reroute due to a pessimistic prediction.

Note that the P2 predicted end time of the disruption of the 75%-quantile scenario (see Table 4.2) is 10 minutes short of the actual end time. Because of this slightly overoptimistic prediction, the predicted end time is updated to P2a.

This new prediction, however, is far too pessimistic and disrupts the late afternoon rush hour (Figure 4.10). Consequently, this scenario is the worst in terms of the number of passengers affected and the increase in generalized travel time.

4.3.2.2 Case Study 1A

In this case study, we are interested in how the disruption's time of occurrence affects the passengers for each choice of predictions. To observe this, we consider an artificial disruption where the same disruption as in Case Study 1 is assumed to occur in the evening of the same day at 19:24. The realizations of the latency and repair time are taken from the values of the computed conditional distributions of the two times, which correspond to the same quantiles as the realizations in Case Study 1. In this case, the latency and repair times are 89 and 73 minutes, respectively. The total disruption length is, therefore, 162 minutes and it ends at 22:06. Note that this artificial disruption occurs not during the repair team's contractual working hours so the latency time is longer than in Case Study 1.

The P1 predictions of this case study are presented in Table 4.5.

Table 4.5: *P1 predictions for Case Study 1A.*

Qtl	P1		P1a		P1b		P1c	
(%)	Len	Time	Len	Time	Len	Time	Len	Time
25	54	20:18	77	20:41	102	21:06	127	21:31
50	86	20:50	149	21:53				
75	149	21:53						
85	209	22:53						
90	258	23:42						
Mean	122	21:26						

The P2 predictions are made at 21:08, 15 minutes after the repair team's actual arrival time at 20:53. These predictions are presented in Table 4.6.

Table 4.6: *P2 predictions for Case Study 1A.*

Qtl	P2		P2a		P2b		P2c		P2d	
(%)	Len	Time	Len	Time	Len	Time	Len	Time	Len	Time
25	104	21:08	121	21:25	137	21:41	154	21:58	169	22:13
50	114	21:18	153	21:57	198	22:42				
75	152	21:56	259	23:43						
85	179	22:23								
90	213	22:57								
Mean	138	21:42	207	22:51						

As before, the predictions are used by the short-turning model to produce the disruption timetables. Table 4.7 summarizes the outcome of the short-turning model for each scenario. The short-turning model produces different outcomes than in Case Study 1. This is due to the different timetable between the afternoon and evening time.

Table 4.7: *Summary of the outcome of the short-turning model for Case Study 1A.*

Qtl (%)	$\#C$	$\#D$	del	$\#ST$
25	47	2	10	10
50	55	2	8	12
75	73	2	10	15
85	47	4	38	10
90	55	12	69	12
Mean	55	5	31	12

The disruption timetables are used by the passenger-flow model to compute the flow of the passengers. In this case, 7,102 passengers travel in the region. Because the disruption occurs in the evening, the number of passengers is lower than in the previous case study.

Table 4.8: *Impact on passengers of different predictions in Case Study 1A.*

Qtl (%)	Excess (minutes)	# Affected Passengers	Inc. orig (%)	Inc. bench (%)	# Rerouting 1	# Rerouting 2	# Transfer
25	7	4966	32.6834	7.4781	438	195	1353
50	36	5563	31.0920	6.1890	479	8	1430
75	97	6563	35.1252	9.4560	83	0	1512
85	17	5180	29.4636	4.8699	52	0	1379
90	51	5843	32.1586	7.0530	0	0	1461
Mean	45	5733	27.1643	3.0074	91	0	1306
Real	0	4812	23.4516	0	0	0	1291

Table 4.8 summarizes the impact of the disruption and the different choices of disruption-length prediction on the passengers for each scenario. In comparison to Case Study 1, the increase in the total generalized travel time with respect to the normal situation is higher. This is because the disruption is longer in this case study due to the longer latency time.

The benchmark case still represents the best situation where the least number of passengers are affected. The total generalized travel time is the lowest in this scenario where no passengers are rerouted.

The longer the latest P2 prediction, the more passengers that are affected by the disruption. For this reason, the largest total generalized travel time is observed in the 75%-quantile scenario. Note that, as in Case Study 1, the P2 prediction of this scenario is 10 minutes short of the actual end time. Consequently, the prediction is updated to P2a that is far too pessimistic.

In the 25% quantile scenario, 1487 reroutings are performed due to the nine predictions that are produced. With 438 and 195 passengers having to reroute once and twice, respectively, a considerable amount of passengers have to change their plans more than twice. As a result, the total generalized travel time of this scenario is the second highest due to the heavy penalty from the large number of reroutings.

4.3.2.3 Case Study 2

In this case study, we consider another real TC disruption at the same location which occurred on Saturday, 18 October 2014. It started at 19:24 and had the following information:

1. Contract type: OPC.
2. Working station (maintenance base) distance: 7.1620 kilometres.
3. Level crossing distance: 872.372 metres.
4. Working (contractual) time: no.
5. Warm: no.
6. Rush hour: no.
7. Presence of an overlapping disruption: no.
8. Cause: a cable problem.

The observed latency and repair time were 47 and 88 minutes, respectively. The total disruption length was, therefore, 135 minutes and it ended at 21:39.

The P1 predictions for this case study are presented in Table 4.9.

Table 4.9: *P1 predictions for Case Study 2.*

Qtl	P1		P1a	
(%)	Len	Time	Len	Time
25	54	20:18	77	20:41
50	86	20:50		
75	149	21:53		
85	209	22:53		
90	258	23:42		
Mean	122	21:26		

Fifteen minutes after the repair team's actual arrival time at 20:11, the P2 predictions are made. Table 4.10 presents these predictions.

Table 4.10: *P2 predictions for Case Study 2.*

Qtl	P2		P2a		P2b		P2c	
(%)	Len	Time	Len	Time	Len	Time	Len	Time
25	67	20:31	94	20:58	120	21:24	151	21:55
50	104	21:08	173	22:17				
75	173	22:17						
85	237	23:21						
90	280	00:04						
Mean	142	21:46						

Table 4.11: *Summary of the outcome of the short-turning model for Case Study 2.*

Qtl (%)	$\#C$	$\#D$	del	$\#ST$
25	43	2	6	9
50	47	2	18	10
75	47	2	18	10
85	65	7	102	13
90	77	8	79	16
Mean	43	0	0	9

Table 4.12: *Impact on passengers of different predictions in Case Study 2.*

Qtl (%)	Excess (minutes)	# Affected Passengers	Inc. orig (%)	Inc. bench (%)	# Rerouting 1	# Rerouting 2	# Transfer
25	16	4562	14.4391	0.6413	426	32	1219
50	38	5052	17.1197	2.9987	252	32	1506
75	38	5052	16.1172	2.1171	0	0	1437
85	102	6246	19.3051	4.9206	0	0	1558
90	145	6792	21.4386	6.7969	0	0	1693
Mean	7	4351	14.8926	1.0401	2	0	1385
Real	0	4182	13.7099	0	0	0	1309

Table 4.11 presents the outcome of the short-turning model for each quantile for this case study. The impact on the passengers is summarized in Table 4.12.

The benchmark case represents the best possible situation. With the least number of affected passengers with no need for rerouting, the total generalized travel time is the lowest of all.

Note that the last predicted end time of disruption of the 50%-quantile and 75%-quantile scenarios are the same. Consequently, the same number of passengers are affected in both cases. There are, however, three predictions in the 50%-quantile scenario while only two in the 75%-quantile scenario. As a result, in the former case many passengers have to reroute and more transfers need to be performed. This causes the total generalized travel time of this scenario to be higher.

The pessimistic 90%-quantile scenario disturbs the most number of passengers. This is because of the predictions that are far too long. For this reason, this scenario is the worst-performing one as indicated by the largest total generalized travel time.

4.3.2.4 Case Study 2A

As in Case Study 1A, this case study assumes the incident in Case Study 2 to occur at a different time during the day. This artificial incident occurred on the same day at 14:22. Because the incident was in a weekend, the prediction lengths did not change from Case Study 2; only the time that was adjusted.

Table 4.13 presents the P1 predictions for this case study.

Fifteen minutes after the arrival time of the repair team, the P2 predictions

Table 4.13: *P1 predictions for Case Study 2A.*

Qtl	P1		P1a	
(%)	Len	Time	Len	Time
25	54	15:16	77	15:39
50	86	15:48		
75	149	16:51		
85	209	17:51		
90	258	18:40		
Mean	122	16:24		

Table 4.14: *P2 predictions for Case Study 2A.*

Qtl	P2		P2a		P2b		P2c	
(%)	Len	Time	Len	Time	Len	Time	Len	Time
25	67	15:29	94	15:56	120	16:22	151	16:53
50	104	16:06	173	17:15				
75	173	17:15						
85	237	18:19						
90	280	19:02						
Mean	142	16:44						

are made. These are shown in Table 4.14. In Table 4.15, an overview of the outcomes of the short-turning model for the different scenarios is provided.

Table 4.15: *Summary of the outcome of the short-turning model for Case Study 2A.*

Qtl (%)	$\#C$	$\#D$	del	$\#ST$
25	40	2	2	10
50	48	0	0	12
75	48	0	0	12
85	64	0	0	16
90	76	0	0	19
Mean	36	9	33	9

The impact of different choices of prediction on the passengers are measured with the passenger-flow model. The result is presented in Table 4.16.

Note that due to the disruption occurring during the day, there are more passengers affected by the disruption in comparison to Case Study 2.

As in the previous three case studies, the scenario with the true disruption length is the best performing one in terms of the total generalized travel time. In this ideal situation, the least number of passengers are affected and none of them have to change their travel plans.

As in Case Study 2, the difference between the predicted end of disruption and the actual one is 38 minutes in both the 50%-quantile and the 75%-quantile scenarios. However, the total generalized travel time is higher in the former case. This is due to the more frequent prediction updates which causes many passengers to reroute and more transfers to be performed.

Table 4.16: *Impact on passengers of different predictions in Case Study 2A.*

Qtl (%)	Excess (minutes)	# Affected Passengers	Inc. orig (%)	Inc. bench (%)	# Rerouting		# Transfer
					1	2	
25	16	9031	12.3160	1.2250	723	120	3249
50	38	10928	15.8734	4.4312	491	46	3725
75	38	10928	14.6030	3.2862	0	0	3703
85	102	16359	19.9581	8.1125	0	0	4549
90	145	19044	23.2581	11.0867	0	0	5069
Mean	7	8294	11.1051	0.1337	4	0	3128
Real	0	7736	10.9567	0	0	0	3128

The total generalized travel time is the worst in the very pessimistic 90% quantile scenario. In this case, the great difference between the predicted end of disruption and the truth means a lot of passengers are affected by the disruption.

In the next subsection we summarize what has been learned from these case studies.

4.3.3 Discussion

To measure the impact of a disruption-length prediction on passengers, a cost function needs to be defined. In this chapter, that function is the total generalized travel time. The impact is measured as the weighted total travel time of all passengers, including waiting time, in-vehicle time, transfer time, number of transfers and number of reroutings.

With this cost function, we have observed the impact of uncertainty about the length of a disruption on train traffic and passengers. In general, the difference between the predicted and the actual end time of a disruption tends to be longer when the prediction is pessimistic. Consequently, more passengers are affected and so total generalized travel time increases. On the other hand, the difference between the predicted and actual end time tends to be smaller when the prediction is optimistic. But while it means that fewer passengers are affected, this choice also results in predictions needing to be updated more frequently. Consequently, more passengers have to change their travel plans and reroute an “inconvenience” which increases total generalized travel time due to the penalty β_{re} in equation (4.2).

In the experiment, when a prediction is updated, the new version chosen has the value of the new conditional distribution corresponding to the same quantile. Moreover, train operations are never resumed before the predicted end of the disruption, even if it has actually ended already. The consequence of this is that a prediction that falls slightly short of the actual end time has a high cost. Because the disruption ends not long after the original predicted time, the updated prediction exceeds it quite substantially. As a result, more passengers are affected and total generalized travel time increases. This situation is illustrated in the 75%-quantile scenario in Case Study 1.

The generalized travel time in equation (4.2) is a simplified one. For instance, constraints of capacity and vehicle type have not been taken into the account. The bus service between Utrecht and Houten provides significantly less capacity than the train service. Moreover, different vehicle types provide different levels of comfort for passengers. An Intercity (express) train is generally more comfortable than a Sprinter (local) train or a bus service. The weight β_{ty} corresponding to the vehicle type can also be added to the second term of equation (4.2) as $\sum_{a=1}^{N_{tr}+1} \beta_{in} \cdot \beta_{ty} \cdot t_{in}^{a}$. The generalized travel time can be extended to model the complex nature of a railway disruption better.

Choosing the total generalized travel time as the cost function means the impact of the uncertainty in disruption length is measured only from the passenger's point of view. With this cost function, an optimistic prediction is not attractive only because it causes inconvenience to passengers who have to reroute. We have not captured all costs associated with an optimistic prediction as faced by the OCCR. From the operational point of view, implementing numerous timetable updates is impractical due to the associated logistical issues. For instance, rolling stock and crew assignments need to be reorganized with every update. To make the cost function more realistic, such issues would need to be considered.

4.4 Chapter Summary

In this chapter we have shown how the disruption-length model can be used to predict the length of a railway disruption using the software `UNINET`. The computation undertaken to produce predictions is very efficient, making the model attractive for use in practice.

The output of the model is the conditional distribution of disruption length. One value from the distribution needs to be chosen as the prediction. To investigate the effect of different choices of prediction on train traffic and passengers, the model is used together with the short-turning model and the passenger-flow model in a railway disruption experiment. This is conducted in the form of four case studies involving railway incidents in the vicinity of Houten in the central Netherlands. Several different predictions corresponding to different quantiles of the conditional distribution of disruption length are considered.

The cost associated with each prediction is measured in terms of the total generalized travel time. We have observed how this cost is affected by different choices of prediction. When the prediction is too optimistic, the cost tends to be higher due to the larger number of reroutings. When the prediction is too pessimistic, the cost tends to be higher because more passengers are affected.

To obtain more realistic conclusions for the Dutch railway network, a more complicated cost function needs to be considered. More factors should be added to the generalized travel time. The operational cost associated with each prediction needs to be included as well.

In the next chapter we conclude this thesis with a summary of what has been learned and a number of recommendations for future studies.

CHAPTER 5

Summary, Recommendations, and Final Remarks

This chapter concludes the PhD research project and thesis. We start by summarizing what has been learned, both theoretically and in practice. We then make a number of recommendations for ProRail operations and for future studies. To end this chapter and thesis, in the section containing final remarks we reflect on our experience during the research process.

5.1 Thesis Summary

The thesis covers two topics: joint distribution modelling with copulas and the application thereof in railway traffic management. In this section we present what we have learned about these two topics.

5.1.1 Copulas in Multivariate Mixed Discrete-Continuous Problems

One way to construct a joint distribution is by using the copula. When all the variables are continuous, a copula separates the dependence from the marginal distributions. However, this is no longer true when one or more variables are discrete. Moreover, in this case copula parameter estimation in the presence of data becomes computationally more expensive. This is because it needs to be performed via maximum likelihood. Nonetheless, it is still shown that the copula is a useful concept for dependence modelling in either case.

Another challenge with copula modelling appears when the number of dimensions is larger than two. In this case there are only few multivariate copula models that are available with marginal or functional constraints which need to be satisfied. The copula-vine concept tackles this challenge. With this approach, a very complicated multivariate joint distribution can be constructed with a set of algebraically independent bivariate (conditional) copulas. Additionally, the conditional copulas could be chosen to depend on the conditioning variables (which we call the *non-constant* copula-vine model).

In this thesis we focus on the use of copula-vine to model the joint distribution of mixed discrete and continuous variables. Given a vine structure and bivariate copula families, Algorithm 2.1 is proposed to estimate the parameter values of the copula-vine model from a set of data. The algorithm considers non-constant copula when the conditioning set contains only discrete variables. In principle, this can be generalized to all pairs, but certain non-constant relationship assumptions need to be made when the conditioning set also contains continuous variable(s). Example 2.3.3 in Chapter 2 shows how this can be done.

When the true vine structure and copula families are known, it is shown using an example that Algorithm 2.1 performs as it is supposed to. For the artificially constructed datasets with different numbers of samples, the true copula parameters are approximated well. Moreover, the model recovers the discrete-only and continuous-only parts of the data. Because checking the model fit to the mixed part of the data is difficult, we measure the performance of the model in predicting the outcome of a discrete or a continuous dependent variable. The model performs well and recovers the conditional distribution of the dependent variable.

Good recovery of the conditional distribution of a dependent continuous variable is necessary for our application. A prediction value is chosen from this conditional distribution, so it is important for the model to produce the right distribution. A few misspecification scenarios have been considered. When the wrong copula families or an incorrect vine structure are chosen, misfit in the continuous part of the model is observed. In this case, the model does not recover the conditional distribution of the continuous variable. On the other hand, when a misfit is observed in the discrete part of the model (caused by, e.g., fitting constant copulas on data generated from non-constant copulas), the conditional distribution is still recovered.

A comparison with the popular generalized linear model (GLM) has been performed. One constraint of the GLM is the assumption that the conditional distribution of the dependent variable comes from the exponential family. In our experiments, the performance of the GLMs and copula-vine models is generally comparable in predicting the mean of the conditional distribution. However, the GLMs do not recover the conditional distribution, whereas – under the right setting – the copula-vine models do.

The joint distribution model for the railway disruption length is constructed using the MVN copula and copula-vine approaches, which we summarize in the next subsection.

5.1.2 The Railway Disruption-Length Models

In this thesis, disruption length is split into two time regimes: the latency and the repair time. These are influenced by different factors, which we also need to determine. Using the SAP database as our main source of data, eight influencing factors are found and included in the disruption-length models for incidents caused by track circuit (TC) or switch (points) failures.

We construct two disruption-length models. The first is based on the multivariate normal (MVN) copula whose parameters are computed with maximum likelihood. The second is constructed using the copula-vine approach, where the parameters are estimated with Algorithm 2.1. The continuous part of the data is recovered by both models. However, the MVN copula model does not represent the discrete part of the data well, while the copula-vine model does. Consequently, the likelihood of the copula-vine model is higher than the MVN copula model, indicating a better fit to the data.

In terms of prediction, the performance of both models is similar. The conditional distributions of the latency and repair times are recovered by both, and the $RMSE$ values are similar. In practice, we are interested in the prediction. Therefore, both models are equally attractive. For this reason, we choose to work with the MVN copula model in the application part of this thesis, since this can be implemented in `UNINET` even though, for the time being, the parameters need to be computed outside of the software. The very fast computational time of `UNINET` makes application of this model more practical. All the BN figures and predictions presented in Chapter 4 are generated using `UNINET`.

However, it is important to acknowledge the models' low coefficient of determination R^2 values. This indicates the complexity of railway disruptions in the Netherlands. It might be of interest to include more influencing factors in order to construct more predictive models. Unfortunately, no such additional data was available during the course of our research. Its inclusion is therefore left for future studies, where the models can be expanded by incorporating new information, if and when it becomes available.

The output of our models is the conditional distribution of disruption length. In collaboration with the Department of Transport and Planning at Delft University of Technology, we investigate the effect of the uncertainty of the length of disruption on train traffic and passengers. The disruption-length models are combined with the short-turning model and the passenger-flow model to examine a disruption to railway traffic in the vicinity of Houten, the Netherlands. In the Dutch railway network, this is an important area forming part of the A2 corridor which connects the cities of Amsterdam and Eindhoven.

A few case studies are considered to learn the effect of different choices of prediction corresponding to different quantiles of the conditional distribution of disruption length. The effect of each choice is measured with the total generalized travel time as the cost function. When the prediction is too optimistic, many passengers have to be rerouted, increasing inconvenience for them and thus the total generalized travel cost. On the other hand, when the prediction is too pessimistic, more passengers are affected, which also results in a higher total generalized travel cost. Similarly, when the prediction is just slightly shorter than the actual disruption length, more passengers are also affected.

The findings show how uncertainty about the disruption length affects passenger traffic, even in the simplistic setup of the experiment. This shows the importance of properly modelling the conditional distribution of disruption length.

5.2 Recommendations

In this section, we present a number recommendations for the future, to improve ProRail's performance during disruptions.

The application part of this thesis consists of two steps: model construction and model use. The former can be improved primarily by recording data better and by collecting more information. For the latter, we need to expand the short-turning model and the passenger-flow model to cover the entire Dutch rail network. We discuss each of these recommendations in greater detail below.

5.2.1 Better Data Collection

With this thesis, one goal we hope to have accomplished is to show the importance of good data. Data reflects historical performance and contains a lot of information, in much the same way a ProRail expert does. In dealing with railway disruptions, however, at present ProRail relies mostly on experts and not so much on data. In a way, this is understandable given the generally poor quality of the data available, as we have experienced at first hand in this research.

However, ProRail's performance would be improved if better data were to be collected. Unlike an expert, data is available at all times. We do not have to be concerned if, for instance, an expert is about to retire or a disruption occurs when no experts are available. Moreover, the presence of data enables ProRail to evaluate its processes objectively. We are not implying here that the role played by experts can be substituted by data and models. Rather, we believe that they are complementary. We envision data and models being used as support tools to help the experts arrive at the best decisions during disruptions.

In fact, the role of experts can be very beneficial in applying the model. Their expertise can be used to interpret its output, by considering factors not included in the model. In the context of the disruption-length model, for instance, the experience of the repair team attending the disruption site might influence the length of the disruption. This variable is not included in the disruption-length model. In practice, a certain quantile of the conditional distribution of disruption length can be chosen as the prediction based on the experts knowledge by considering information about repair-team experience.

An expert can also use data to test an opinion in order to make (or justify) a decision, through data analysis. For example, ProRail is currently in the gradual process of converting all its contracts to PGO. This is because, as we have seen in subsections 3.1.1.2 and 3.1.2.1, disruption lengths are shorter with this type of contract. This means that we can hope that future disruption lengths will be shorter, once all OPC contracts have been switched to PGO.

As another example of the importance of data analysis, a ProRail expert suggested that we include train traffic density as a factor representing location in our model. However, as we have shown in subsection 3.1.1.2, this does not in fact influence the latency time in the data. Consequently, we decided to not include it in our model.

In this thesis we have shown that the disruption-length models represent the data well. However, while the models recover the conditional distribution of disruption length, their R^2 value is low. This shows the poor quality of the data, because a model can only be as good as the data it uses. On the other hand, it also indicates the potential benefit of expanding the model with more influencing factors if such information ever becomes available. Knowledge about what factors and variables need to be collected is already held within the ProRail organization, in the expertise and experience of its experts. The following are a number of such items, gleaned from our experience and from discussions with TC and switch experts with whom we were in contact during our research.

1. Information about the actual departure points of the repair teams and the routes they take, in order to better predict latency time.
2. More detailed information about the actual repair process, e.g. the length of diagnosis time, whether or not the repair team has all the equipment it needs, whether or not the disrupted site is clear for access when it arrives[1], etc.
3. Better data registration regarding the causes of failures.
4. Technical information about the equipment affected. In the case of switch failures, for instance, information about the type of switch, the number of point machines it contains, their type and whether or not the switch is monitored[2] could be considered.

With regard to the third point in the list above, as we have discussed in Chapter 3 the data should be recorded in the SAP database. Unfortunately, the quality of the actual records is poor, with the cause of most incidents registered as either “unknown” or “other”. In this thesis, we tackled this problem by manually reading the information in the “Remark” column under the supervision of ProRail experts. However, this section in the SAP database is not standardized. Consequently, the amount of information available varies widely between individual incident records.

This is a long-standing problem, which ProRail has recognized. In 2013-2014, a project named “*Proeftuin Gelre*” (Testbed Gelderland) was established with the aim of improving data collection. A pilot experiment was conducted in the province of Gelderland in the eastern Netherlands. Unfortunately, the quality of the data collected did not improve significantly over what was previously been collected in SAP. In a sample of this data provided to us, the numbers of “unknowns” and “others” were still quite significant.

Indicating that the importance of better data registration has still not been fully understood by all ProRail’s stakeholders, this is a challenge that needs to be addressed by ProRail as the owner of the problem.

---

[1] For safety reason, the repair team must wait for clearance from the train traffic controllers before it is allowed to enter the site to diagnose and repair the problem.

[2] Monitored switches are usually the more important ones in the network. Monitoring information can help the repair team to diagnose the problem faster, thus shortening disruption length.

5.2.2 Models Expansion

In the second half of Chapter 4, we show how the disruption-length models could be implemented in practice. The models are tested in small case studies in a certain region of the Netherlands, together with the short-turning model and the passenger-flow model. We see how the uncertainty in the disruption length, which is well-represented by the conditional distribution, affects train traffic and passengers in terms of total generalized travel time.

However, this does not solve one key problem faced by the OCCR: which value of the conditional distribution of disruption length to choose as the prediction. From only four case studies, we cannot conclude which of these values is the "best". To do that, many more situations need to be considered. This can be done by, for instance, generating many different realizations of disruption length from the conditional distribution. For each of these, the short-turning model and the passenger-flow model are run to compute the effect of different choices of disruption-length predictions. The result of this simulation can then be used to determine the optimal value of the conditional distribution as the prediction.

Note that the effect of different predictions might also depend on the location and the time of the incident. A more complicated cost function might need to be considered as well.

In the experiment in Section 4.3, we consider a simple region where there are few alternative routes for passengers. If a disruption occurs in the vicinity of, say, Amsterdam, passengers have more alternatives to choose from. Apart from the option of travelling by different trains without having to detour too far, it is also possible to switch to other types of public transport. The city's dense, high-frequency urban network, with bus, tram and metro services, might make a very attractive alternative to the train.

If a disruption occurs in the middle of the night, only a very small number of passengers are travelling. In terms of the total generalized travel time, then, there will be little to no penalty for making optimistic predictions. On the other hand, the logistical problems associated with the assignment of rolling stock and crew every time a new timetable is made remain. Moreover, thus far we have only considered how disruption affects passenger movements. In reality, at night many freight trains run on the Dutch railway network, transporting valuable goods. It might therefore be interesting to define a cost function that considers not only passengers, but also the operation of freight traffic during disruptions.

This would require expansion of the short-turning and passenger-flow models to cover the entire Dutch railway network. Even once the models are ready, substantial experimental computation time will still be required. With today's rapid advances in computing, however, it may not be too far-fetched to believe that this will become possible in the not-too-distant future.

In the case studies, moreover, when a prediction was updated we took the same quantile of the conditional distributions as in the previous prediction. This was done to simplify the problems. In principle, though, it need not be the case.

For instance, the choice of quantile in the "P2" prediction might depend on the quantile that realizes the latency time in the "P1" prediction. The realization of the latency time might indicate how fast or slow the repair team is working on the incident, which could be useful information in producing a more accurate "P2" prediction. Other possibilities could be considered as well.

5.3 A Final Remark

The research presented in this thesis has been funded through the ExploRail research programme, a collaboration between the Netherlands Organisation for Scientific Research (NWO), through Technology Foundation STW, and ProRail. The project is called the "Smart Information and Decision Support for Railway Operation Control Centres" (SmartOCCR).

The goal of the research project, as described in the original proposal, was to improve information on disruption durations using the available data streams in the OCCR, which is used to achieve more effective and efficient dispatching through better decision support. We have to admit that this goal has not been entirely achieved, due to the many obstacles we encountered during the course of the research. In particular, the following.

1. It took more than a year and a half before we gained access to the data. While everybody in ProRail was aware of the existence of such data, it was scattered throughout the organization. This made it difficult to find those people who held the access keys to the data.
2. Once the data was obtained, its quality was poor.
3. While we received full support from ProRail, there was a lack of information and support from the other stakeholders in the Dutch railway industry.
4. The size and complexity of the Dutch railway network made it difficult for two PhD students, given the available timeframe, to program the entire network to simulate many different disruptions.

The Dutch railway industry involves a lot of stakeholders, with ProRail at the centre. Each stakeholder has its own interests, according to its role in the industry. However, all must remember that they are not independent of each other. They are connected by the common goal of providing reliable train services in the Netherlands.

The extent to which unplanned failures of railway assets hinder train traffic depends on the length of the disruptions. In this respect, we enter the domain of the maintenance contractors, where knowledge of the affected assets is needed in order to be able to predict disruption length. Unfortunately, as a side-effect of outsourcing this maintenance work to several different contractors, ProRail appears to have lost some grasp of this knowledge. This posed a major obstacle for our research.

We believe this is a problem which needs to be addressed and solved by ProRail. It is ProRail's responsibility to make sure that all stakeholders have the same perspective on their common goal.

The nature of a research project is to explore possible innovations. For this reason, open-mindedness and curiosity are attitudes the stakeholders in the research need to possess. Reluctance to embrace new ideas, while it feels comfortable, is not healthy especially in the long term. The very existence of the ExploRail programme is a good indication that the ProRail organization is aware of the importance of innovation. Unfortunately, we have also observed that, as yet, not every member of the organization welcomes this attitude.

As Bill Clinton once said, *"The price of doing the same old thing is far higher than the price of change"*. However, as people have always said, change takes time.

References

Agresti, A. (1980). Generalized odds ratios for ordinal data. *Biometrics*, 36:59–67.

Aho, K., Derryberry, D., and Peterson, T. (2014). Model selection for ecologists: the worldviews of aic and bic. *Ecological Society of America*, 95(3):631–636.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723.

Anderson, T. W. (1962). On the distribution of the two sample cramer-von mises criterion. *The Annals of Mathematical Statistics*, 33(3):1148–1159.

Anderson, T. W. and Darling, D. A. (1952). Asymptotic theory of certain goodness-of-fit criteria based on stochastic processes. *The Annals of Mathematical Statistics*, 23:193–212.

Balcombe, R., Mackett, R., Paulley, N., Preston, J., Shires, J., Titheridge, H., Wardman, M., and White, P. (2004). The demand for public transport: A practical guide (report no. trl593). Technical report, London, UK: Transportation Research Laboratory.

Bedford, T. and Cooke, R. (2002). Vines - a new graphical model for dependent random variables. *Annals of Statistics*, 30(4):1030–1068.

Bender, A., Mussa, H. Y., , Glen, R. C., and Reiling, S. (2004). Molecular similarity searching using atom environments, information-based feature selection, and a naïve bayesian classifier. *Journal of Chemical Information and Computer Sciences*, 44(1):170–178.

Berg, D. (2009). Copula goodness-of-fit testing: an overview and power comparison. *The European Journal of Finance*, 15(7-8):675–701.

Bouyé, E., Durrleman, V., Nikeghbali, A., Riboulet, G., and Roncalli, T. (2000). Copulas for finance - a reading guide and some applications. Technical report, Crédit Lyonnais.

Breymann, W., Dias, A., and Embrechts, P. (2003). Dependence structures for multivariate high-frequency data in finance. *Quantitative Finance*, 3:1–14.

Cacchiani, V., Huisman, D., Kidd, M., Kroon, L., Toth, P., Veelenturf, L., and Wagenaar, J. (2014). An overview of recovery models and algorithms for real-time railway rescheduling. *Transportation Research Part B: Methodological*, 63:15–37.

Carley, H. (2002). Maximum and minimum extensions of finite subcopulas. *Comm. Stats. Theory Methods*, 31:2151–2166.

Cats, O., West, J., and Eliasson, J. (2016). A dynamic stochastic model for evaluating congestion and crowding effects in transit systems. *Transportation Research Part B: Methodological*, 89:43–57.

Chen, C., Zhang, G., Wang, H., Yang, J., Jin, P. J., and Walton, C. M. (2015). Bayesian network-based formulation and analysis for toll road utilization supported by traffic information provision. *Transportation Research Part C: Emerging Technologies*, 60:339–359.

Clarke, K. A. (2003). Nonparametric model discrimination in international relations. *Journal of Conflict Resolution*, 47(1):72–93.

Clayton, D. G. (1978). Model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*, 65:141–151.

Cooke, R. M. (1997). Markov and entropy properties of tree and vine-dependent variables. In *Proc. of the Section on Bayesian Statistical Science*. American Statistical Association.

Cooke, R. M., Kurowicka, D., and Wilson, K. (2015). Sampling, conditionalizing, counting, merging, searching regular vines. *Journal of Multivariate Analysis*, 138:4–18.

Corman, F., D'Ariano, A., Pranzo, M., and Hansen, I. A. (2011). Effectiveness of dynamic reordering and rerouting of trains in a complicated and densely occupied station area. *Transportation Planning and Technology*, 34(4):341–362.

Cover, T. M. and Thomas, J. A. (2006). *Elements of Information Theory*. Wiley.

Dai, B., Ding, S., and Wahba, G. (2013). The multivariate bernoulli distribution. *Bernoulli*, 19(4):1465–1483.

D'Ariano, A., Pacciarelli, D., and Pranzo, M. (2008). Assessment of flexible timetables in real-time traffic management of a railway bottleneck. *Transportation Research Part C: Emerging Technologies*, 16(2):232–245.

De Leon, A. R. and Wu, B. (2010). Copula-based regression models for a bivariate mixed discrete and continuous outcome. *Statistics in Medicine*, 30:175–185.

Denuit, M. and Lambert, P. (2005). Constraints on concordance measures in bivariate discrete data. *Journal of Multivariate Analysis*, 93(1):40–57.

Dobrić, J. and Schmid, F. (2007). A goodness of fit test for copulas based on rosenblatts transformation. *Computational Statistics and Data Analysis*, 51:4633–4642.

Genest, C. and Nešlehová, J. (2007). A primer on copulas for count data. *Astin Bulletin*, 37(2):475–514.

Genest, C., Nikoloulopoulos, A. K., Rivest, L.-P., and Fortin, M. (2013). Predicting dependent binary outcomes through logistic regressions and meta-elliptical copulas. *Brazilian Journal of Probability and Statistics*, 27:265–284.

Ghaemi, N. and Goverde, R. M. P. (2015). Review of railway disruption management practice and literature. In *Proc. 6th International Conference on Railway Operations Modelling and Analysis (RailTokyo2015, Tokyo, Japan, March 2015)*.

Ghaemi, N., Goverde, R. M. P., and Cats, O. (2016a). Railway disruption timetable: Short-turnings in case of complete blockage. In *Proc.2016 IEEE International Conference on Intelligent Rail Transportation (IEEE ICIRT 2016, Birmingham, England, August 2016)*.

Ghaemi, N., Zilko, A. A., Yan, F., Cats, O., Kurowicka, D., and Goverde, R. M. P. (2016b). Impact of railway disruption predictions and rescheduling strategies on passenger delays. Submitted to Transportation Research Part C: Emerging Technologies (2016).

Giuliano, G. (1989). Incident characteristics, frequency, and duration on a high volume urban freeway. *Transportation Research Part A: Policy and Practice*, 23(5):387–396.

Golob, T. F., Recker, W. W., and Leonard, J. D. (1986). An analysis of the severity and incident duration of truck-involved freeway accidents. *Accident Analysis and Prevention*, 19(4):375–395.

Goodman, L. A. and Kruskal, W. H. (1954). Measures of association for cross classifications. *Journal of the Americal Statistical Association*, 49:732–764.

Gregoriades, A. and Mouskos, K. C. (2013). Black spots identification through a bayesian networks quantification of accident risk index. *Transportation Research Part C: Emerging Technologies*, 28:28–43.

Haff, I. H. (2013). Parameter estimation for pair-copula constructions. *Bernoulli*, 19(2):462–491.

Hanea, A., Gheorghe, M., Hanea, R., and Ababei, D. (2013). Non-parametric bayesian networks for parameter estimation in reservoir simulation: a graphical take on the ensemble kalman filter (part i). *Journal of Computational Geosciences*, 17(6):929–949.

Hanea, A., Kurowicka, D., and Cooke, R. (2006). Hybrid method for quantifying and analyzing bayesian belief nets. *Quality and reliability Engineering International*, 22:709–729.

Hanea, A., Kurowicka, D., and Cooke, R. (2007). The population version of spearman's rank correlation coefficient in the case of ordinal discrete random variables. *In proceedings of the Third Brazilian Conference on Statistical Modelling in Insurance and Finance*.

Hanea, A., Kurowicka, D., Cooke, R., and Ababei, D. (2010). Mining and visualizing ordinal data with non-parametric continuous bbns. *Computational Statistics & Data Analysis*, 54(3):668–687.

Hanea, D. and Ale, B. (2009). Risk of human fatality in building fires: A decision tool using bayesian networks. *Fire safety journal*, 44(5):704–710.

He, J., Li, H., Edmonson, A. C., Rader, D. J., and Li, M. (2012). A gaussian copula approach for the analysis of secondary phenotypes in casecontrol genetic association studies. *Biostatistics*, 13:497–508.

Highway Capacity Manual (2010). *Highway Capacity Manual*. Transportation Research Board (TRB). National Research Council, Washington, DC.

Jensen, L. W., Landex, A., and Nielsen, O. A. (2015). Assessment of stochastic capacity consuption in railway networks. In *Proc. 6th International Conference on Railway Operations Modelling and Analysis (RailTokyo2015, Tokyo, Japan, March 2015)*.

Joe, H. (1997). *Multivariate Models and Dependence Concepts*. Chapman and Hall.

Joe, H. (2014). *Dependence Modelling with Copulas*. CRC Press.

Johnson, J. B. and Omland, K. S. (2004). Model selection in ecology and evolution. *TRENDS in Ecology and Evolution*, 19(2):101–108.

Kimberling, C. H. (1974). A probabilistic interpretation of complete monotonicity. *Aequationes Mathematicae*, 10(2):152–164.

Kojadinovic, I. and Holmes, M. (2009). Tests of independence among continuous random vectors based on cramer-von mises functionals of the empirical copula process. *Journal of Multivariate Analysis*, 100(6):1137–1154.

Kurowicka, D. and Cooke, R. (2005). Distribution-free continuous bayesian belief nets. *Modern Statistical and mathematical Methods in Reliability*, pages 309–323.

Kurowicka, D. and Cooke, R. (2006). *Uncertainty Analysis with High Dimensional Dependence Modelling*. Wiley Series in Probability and Statistics. Wiley.

Kurowicka, D. and Joe, H. (2011). *Dependence Modeling. Vine Copula Handbook*. World Scientific.

Kurz, M. (2013). Tests on the partial copula. Master's thesis, Ludwig-Maximilians-Universität München.

Li, J. and Wong, W. K. (2010). Two-dimensional toxic dose and multivariate logistic regression, with application to decompression sickness. *Biostatistics*, 12(1):143–155.

Marchant, J. A. and Onyango, C. M. (2002). Comparison of a bayesian classifier with a multilayer feed-forward neural network using the example of plant/-weed/soil discrimination. *Computers and Electronics in Agriculture*, 39:3–22.

Margaritis, D. (2003). *Learning Bayesian Network Model Structure from Data*. PhD thesis, School of Computer Science, Carnegie-Mellon University, Pittsburgh, PA.

Masarotto, G. and Varin, C. (2012). Gaussian copula marginal regression. *Electronic Journal of Statistics*, 6:1517–1549.

McNeil, A. J. (2008). Sampling nested archimedean copulas. *Journal of Statistical Computation and Simulation*, 78(6):567–581.

McNeil, A. J. and Nešlehová, J. (2009). Multivariate archimedean copulas, d-monotone functions and ℓ_1-norm symmetric distributions. *The Annals of Statistics*, 37(5):3059–3097.

Meng, L. and Zhou, X. (2011). Robust single-track train dispatching model under a dynamic and stochastic environment: A scenario-based rolling horizon solution approach. *Transportation Research Part B: Methodological*, 45(7):1080–1102.

Mesfioui, M. and Tajar, A. (2005). On the properties of some nonparametric concordance measures in the discrete case. *Journal of Nonparametric Statistics*, 17(5):541–554.

Morales Napoles, O. (2009). *Bayesian Belief Nets and Vines in Aviation Safety and Other Applications*. PhD thesis, Delft University of Technology.

Nadarajah, S. and Kotz, S. (2005). Mathematical properties of the multivariate t distribution. *Acta Applicandae Mathematicae*, 89:53–84.

Nam, D. and Mannering, F. (2000). An exploratory hazard-based analysis of highway incident duration. *Transportation Research Part A: Policy and Practice*, 34(2):85–102.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3):370–384.

Nelsen, R. B. (2006). *An Introduction to Copula*. Springer.

Nešlehová, J. (2007). On rank correlation measures for non-continuous random variables. *Journal of Multivariate Analysis*, 98(3):554–567.

Newton, I. L. and Roeselers, G. (2012). The effect of training set on the classification of honey bee gut microbiota using the naïve bayesian classifier. *BMC Microbiology*, 12:1–9.

Ng, M. W. and Lo, H. K. (2013). Regional air quality conformity in transportation networks with stochastic dependencies: a theoretical copula-based model. *Networks and Spatial Economics*, 13(4):373–397.

Nikoloulopoulos, A. K. (2013). On the estimation of normal copula discrete regression models using the continuous extension and simulated likelihood. *Journal of Statistical Planning and Inference*, 143:1923–1937.

Nikoloulopoulos, A. K. (2015). Efficient estimation of high-dimensional multivariate normal copula models with discrete spatial responses. *Stochastic Environmental Research and Risk Assessment*, pages 1–13.

Nikoloulopoulos, A. K. and Karlis, D. (2008). Multivariate logit copula model with an application to dental data. *Statistics in Medicine*, 27:6393–6406.

Oukhellou, L., Cme, E., Bouillaut, L., and Aknin, P. (2008). Combined use of sensor data and structural knowledge processed by bayesian network: Application to a railway diagnosis aid scheme. *Transportation Research Part C: Emerging Technologies*, 16(6):755–767.

Pachl, J. (2004). *Railway Operation and Control*. VTD Rail Publishing.

Panagiotelis, A., Czado, C., and Joe, H. (2012). Pair copula constructions for multivariate discrete data. *American Statistical Association*, 107(499):1063–1072.

Pearson, E. S. and Hartley, H. O. (1972). *Biometrika Tables for Statisticians. Volume II*. Cambridge University Press.

Pereira, F. C., Rodrigues, F., and Ben-Akiva, M. (2013). Text analysis in incident duration prediction. *Transportation Research Part C: Emerging Technologies*, 37:177–192.

ProRail (2015 (Accessed September 22, 2015)a). *Onze klanten*. `http://www.prorail.nl/vervoerders/onze-klanten`.

ProRail (2015 (Accessed September 22, 2015)b). *ProRail in cijfers*. `http://www.prorail.nl/reizigers/over-prorail/wat-doet-prorail/prorail-in-cijfers`.

Quesenberry, C. P. and Miller, Jr., F. L. (1977). Power studies of some tests for uniformity. *Journal of Statistical Computation and Simulation*, 5(3):161–191.

Quessey, J.-F. (2010). Applications and asymptotic power of marginal-free tests of stochastic vectorial independence. *Journal of Statistical Planning and Inference*, 140(11):3058–3075.

Razali, N. M. and Yap, B. W. (2011). Power comparisons of saphiro-wilk, kolmogorov-smirnov, lilliefors, and anderson-darling tests. *Journal of Statistical Modelling and Analytics*, 2(1):21–33.

Rosenblatt, M. (1952). Remarks on a multivariate transformation. *The Annals of Mathematical Statistics*, 23:470–472.

Schepsmeier, U., Stoeber, J., Brechmann, E. C., Graeler, B., Nagler, T., and Erhardt, T. (2015). *VineCopula: Statistical Inference of Vine Copulas*.

Schranil, S. and Weidmann, U. A. (2013). Forecasting the duration of rail operation disturbances. In *Proc. 92nd Annual Meeting of the Transportation Research Board, Washington, D.C.*

Schwarz, G. (1978). Estimating the dimension of a mode. *Annals of Statistics*, 6:461–464.

Scutari, M. (2010). Learning bayesian networks with the bnlearn r packageh. *Journal of Statistical Software*, 35(3):1–22.

Sklar, A. (1959). Fonctions de rpartition n dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris*, 8:229–231.

Smith, M. S. and Khaled, M. A. (2012). Estimation of copula models with discrete margins via bayesian data augmentation. *American Statistical Association*, 107(497):290–303.

Song, P. X. K., Li, M., and Yuan, Y. (2009). Joint regression analysis of correlated data using gaussian copula. *Biometrics*, 65:60–68.

Spanhel, F. and Kurz, M. S. (2015). Simplified vine copula models: Approximations based on the simplifying assumption. *ArXiv e-prints*.

Srinivas, S., Menon, D., and Meher Prasad, A. (2006). Multivariate simulation and multimodal dependence modeling of vehicle axle weights with copulas. *Journal of Transportation Engineering*, 132(12):944–955.

Steinhaus, H. (1999). *Mathematical Snapshots*. Dover, 3 edition.

Stöber, J., Hong, H. G., Czado, C., and Ghosh, P. (2015). Comorbidity of chronic diseases in the elderly: Patterns identified by a copula design for mixed responses. *Computational Statistics & Data Analysis*, 88:28–39.

Sullivan, E. E. (1997). New model for predicting freeway incidents and incident delays. *Journal of Transportation Engineering*, 123(4):267–275.

Visser, A. J. and Steenkamp, H. (1981). *Spoorstroomlopen*. Nederlandse Spoorwegen (NS).

Vrieze, S. I. (2012). Model selection and psychological theory: A discussion of the differences between the akaike information criterion (aic) and the bayesian information criterion (bic). *Psychological Methods*, 17(2):228–243.

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*, 57(2):307–333.

Wan, K. and Kornhauser, A. L. (1997). *Journal of Transportation Engineering*, 123(4):267–275.

Wang, Q., Garrity, G. M., Tiedje, J. M., and Cole1, J. R. (2007). Naïve bayesian classifier for rapid assignment of rrna sequences into the new bacterial taxonomy. *Applied and Environmental Microbiology*, 73(2):5261–5267.

Wardman, M. (2004). Public transport values of time. *Transport Policy*, 11(4):363–377.

Watson, G. S. (1961). Goodness-of-fit tests on a circle. *Biometrika*, 48:109–114.

Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*. John Wiley & Sons.

Wilkes, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, 9(1):60–62.

Yuan, J. (2006). *Stochastic modelling of train delays and delay propagation in statistics*. PhD thesis, Delft University of Technology.

Zilko, A. A. (2012). Non-parametric bayesian networks (npbns) versus ensemble kalman filter (enkf) in reservoir simulation with non-gaussian measurement noise. Master's thesis, Delft University of Technology, Delft, the Netherlands.

Zilko, A. A., Hanea, A. M., Kurowicka, D., and Goverde, R. M. P. (2014). Non-parametric bayesian network to forecast railway disruption lengths. In *Proc. 2nd International Conference on Railway Technology: Research, Development and Maintenance (Railways 2014, Ajaccio, France, April 2014)*.

Zilko, A. A. and Kurowicka, D. (2016). Copula in a multivariate mixed discrete-continuous model. *Computational Statistics & Data Analysis*, 103:28 – 55.

Zilko, A. A., Kurowicka, D., and Goverde, R. M. P. (2016). Modeling railway disruption lengths with copula bayesian networks. *Transportation Research Part C: Emerging Technologies*, 68:350 – 368.

Zilko, A. A., Kurowicka, D., Hanea, A. M., and Goverde, R. M. P. (2015). The Copula Bayesian Network with Mixed Discrete and Continuous Nodes to Forecast Railway Disruption Lengths. In *Proc. 6th International Conference on Railway Operations Modelling and Analysis (RailTokyo2015, Tokyo, Japan, March 2015)*.

APPENDIX A

Proof of Propositions

In this appendix, the proofs of Proposition 2.1.5 and 2.1.9 are presented.

Proof of Proposition 2.1.5:
The zero three-way interaction means:

$$p(1,1,1)p(1,0,0)p(0,1,0)p(0,0,1) = p(0,0,0)p(1,1,0)p(1,0,1)p(0,1,1). \tag{A.1}$$

Since the margins are equal to 0.5, we obtain:

$$p(0,1,0)+p(0,0,1)+p(0,1,1)+p(0,0,0) = p(1,0,1)+p(1,1,0)+p(1,0,0)+p(1,1,1) \tag{A.2}$$

$$p(1,0,0)+p(0,0,1)+p(1,0,1)+p(0,0,0) = p(0,1,1)+p(1,1,0)+p(0,1,0)+p(1,1,1) \tag{A.3}$$

$$p(1,0,0)+p(0,1,0)+p(1,1,0)+p(0,0,0) = p(0,1,1)+p(1,0,1)+p(0,0,1)+p(1,1,1) \tag{A.4}$$

Subtracting (A.3) from (A.2), (A.4) from (A.3), and (A.2) from (A.4) yield

$$p(0,1,0)+p(0,1,1) = p(1,0,1)+p(1,0,0) \tag{A.5}$$

$$p(0,0,1)+p(1,0,1) = p(1,1,0)+p(0,1,0) \tag{A.6}$$

$$p(1,1,0)+p(1,0,0) = p(0,0,1)+p(0,1,1) \tag{A.7}$$

Substituting (A.5) to (A.2) yields

$$p(0,0,0)+p(0,0,1) = p(1,1,1)+p(1,1,0) \tag{A.8}$$

It can be shown that Equation (A.1), (A.5), (A.6), and (A.7) are satisfied if and only if $p(x_1,x_2,x_3) = p(1-x_1,1-x_2,1-x_3)$ for all $x_1,x_2,x_3 \in \{0,1\}$. We see immediately that the symmetric distribution satisfies the above equations. To see that the symmetry is also necessary, let us assume that e.g. $p(1,0,0) > p(0,1,1)$. Then, from (A.5), $p(0,1,0) > p(1,0,1)$ which leads to $p(0,0,1) > p(1,1,0)$ from (A.6). Further, this means $p(0,0,1) > p(1,1,0)$ from (A.7) which leads to $p(1,1,1) > p(0,0,0)$ from (A.8). Combining this information together yields

$$p(1,1,1)p(1,0,0)p(0,1,0)p(0,0,1) > p(0,0,0)p(1,1,0)p(1,0,1)p(0,1,1)$$

which cannot be true because of (A.1). Therefore, the proof is complete. □

Proof of Proposition 2.1.9:
Without loss of generality, let X_2 be the conditioning variable. Because $\mathbb{P}(X_i = 0) = 0.5 = P(X_i = 1)$ for all $i \in \{1,2,3\}$,

$$\mathbb{P}(X_1 \le 0|X_2 = 1) = 1 - \mathbb{P}(X_1 \le 0|X_2 = 0) \tag{A.9}$$

and

$$\begin{aligned}\mathbb{P}(X_1 \le 0, X_3 \le 0|X_2 = 1) &= \frac{\mathbb{P}(X_1 \le 0, X_3 \le 0)}{0.5} - \mathbb{P}(X_1 \le 0, X_3 \le 0|X_2 = 0)\\ C_{13|2=1}(\mathbb{P}(X_1 \le 0|X_2 = 1), \mathbb{P}(X_3 \le 0|X_2 = 1)) &= \frac{\mathbb{P}(X_1 \le 0, X_3 \le 0)}{0.5}\\ &\quad - C_{13|2=0}(\mathbb{P}(X_1 \le 0|X_2 = 0), \mathbb{P}(X_3 \le 0|X_2 = 0))\end{aligned} \tag{A.10}$$

Using (A.9), the symmetricity of $C_{13|2}$, and the Total Law of Probability, the left hand side of (A.10) becomes:

$$\begin{aligned}C_{13|2=1}(\mathbb{P}(X_1 \le 0|X_2 = 1), &\mathbb{P}(X_3 \le 0|X_2 = 1)) =\\ &= \frac{\mathbb{P}(X_1 \le 0, X_2 = 0, X_3 \le 0) + \mathbb{P}(X_1 > 0, X_2 = 0, X_3 > 0)}{0.5}\\ &\quad - C_{13|2=1}(\mathbb{P}(X_1 \le 0|X_2 = 0), \mathbb{P}(X_3 \le 0|X_2 = 0))\end{aligned} \tag{A.11}$$

⇒ With the radial symmetry of a trivariate Normal copula, (A.11) becomes:

$$\begin{aligned}&= \frac{\mathbb{P}(X_1 \le 0, X_2 = 0, X_3 \le 0) + \mathbb{P}(X_1 \le 0, X_2 = 1, X_3 \le 0)}{0.5}\\ &\quad - C_{13|2=1}(\mathbb{P}(X_1 \le 0|X_2 = 0), \mathbb{P}(X_3 \le 0|X_2 = 0))\\ &= \frac{\mathbb{P}(X_1 \le 0, X_3 \le 0)}{0.5} - C_{13|2=1}(\mathbb{P}(X_1 \le 0|X_2 = 0), \mathbb{P}(X_3 \le 0|X_2 = 0))\end{aligned}$$

Substituting this result back to (A.10) yields $C_{13|2=0}(\mathbb{P}(Y_1 \le 0|Y_2 = 0), \mathbb{P}(Y_3 \le 0|Y_2 = 0)) = C_{13|2=1}(\mathbb{P}(Y_1 \le 0|Y_2 = 0), \mathbb{P}(Y_3 \le 0|Y_2 = 0))$.

⇐ Because $C_{13|2=0} = C_{13|2=1}$, substituting (A.11) back to (A.10) yields

$$\frac{\mathbb{P}(X_1 \le 0, X_2 = 0, X_3 \le 0) + \mathbb{P}(X_1 > 0, X_2 = 0, X_3 > 0)}{0.5} = \frac{\mathbb{P}(X_1 \le 0, X_3 \le 0)}{0.5}$$

which leads to

$$\mathbb{P}(X_1 > 0, X_2 = 0, X_3 > 0) = \mathbb{P}(X_1 \le 0, X_2 = 1, X_3 \le 0)$$

Substituting this result into equation (A.5), (A.6), (A.7), and (A.8) yields

$$p(x_1, x_2, x_3) = p(1 - x_1, 1 - x_2, 1 - x_3)$$

for all $x_i \in \{0, 1\}$ for all $i \in \{1, 2, 3\}$. This means the trivariate Bernoulli distribution has a radial symmetry. Therefore, the trivariate Normal copula is able to realize (X_1, X_2, X_3). $\square$

APPENDIX B

Experiments Results in Chapter 2

This Appendix contains tables with the estimated parameters values of the different experiments performed in Subsection 2.3.2 and 2.3.3 of this thesis. If the parameter is set to zero or constant, it is indicated in bold in a bracket

1. Testing Algorithm 2.1 in Subsection 2.3.2

Table B.1: *The estimated parameters values for different values of N.*

Parameter	True Value	$N = 100$	$N = 500$	$N = 1000$	$N = 2000$
		Estimate (Conf. Bound)	Estimate (Conf. Bound)	Estimate (Conf. Bound)	Estimate (Conf. Bound)
ρ_{12}	0.3	0.5521 $(0.2658, 0.8061)$	0.3744 $(0.2427, 0.5044)$	0.3835 $(0.2915, 0.4755)$	0.2998 $(0.1935, 0.3325)$
ρ_{23}	0	-0.2063 (**0**) $(-0.5076, 0.1214)$	-0.1464 (**0**) $(-0.2870, 0.0704)$	-0.0977 (**0**) $(-0.2015, 0.0090)$	-0.0123 (**0**) $(-0.0819, 0.0592)$
ρ_{34}	-0.2	-0.2465 (**0**) $(-0.4992, 0.0454)$	-0.1892 $(-0.2978, -0.0850)$	-0.2552 $(0.3339, -0.1802)$	-0.2384 $(-0.2548, -0.1420)$
ρ_{45}	-0.4	-0.5120 $(-0.6500, -0.3554)$	-0.4298 $(-0.4966, -0.3525)$	-0.4182 $(-0.4712, -0.3642)$	-0.4229 $(-0.4608, -0.3891)$
ρ_{56}	0.5	0.4686 $(0.2973, 0.6120)$	0.5263 $(0.4561, 0.5878)$	0.4892 $(0.4386, 0.5345)$	0.4861 $(0.4781, 0.5401)$
$\rho_{13\|2=0}$	-0.5	-0.2554 (**0**) $(-0.9676, 0.4681)$	-0.6078 $(-0.8118, -0.4061)$	-0.6381 $(-0.7933, -0.4492)$	-0.4888 $(-0.5692, -0.4434)$
$\rho_{13\|2=1}$	0.7	0.7593 $(0.4927, 0.9261)$	0.7431 $(0.6256, 0.8412)$	0.7941 $(0.6864, 0.8346)$	0.7619 $(0.6631, 0.7729)$
$\rho_{24\|3=0}$	0.6	0.5880 $(0.3354, 0.8095)$	0.5472 $(0.4345, 0.6482)$	0.6019 $(0.5362, 0.6724)$	0.5755 $(0.5156, 0.6138)$
$\rho_{24\|3=1}$	-0.3	-0.4261 $(-0.7532, -0.0697)$	-0.4063 $(-0.5601, -0.2464)$	-0.3416 $(-0.4728, -0.2104)$	-0.2828 $(-0.3809, -0.1996)$
$\rho_{35\|4}$	-0.3	-0.2526 $(-0.5038, -0.0309)$	-0.3325 $(-0.4310, -0.2163)$	-0.3526 $(-0.4295, -0.2773)$	-0.2484 $(-0.3397, -0.2281)$
$\rho_{46\|5}$	0.8	0.7422 $(0.6743, 0.8200)$	0.7844 $(0.7465, 0.8125)$	0.8024 $(0.7782, 0.8225)$	0.8003 $(0.7787, 0.8123)$
$\rho_{14\|2=0,3=0}$	0.1	-0.0613 (**0**) $(-0.8504, 0.7949)$	0.1187 (**0**) $(-0.1290, 0.3442)$	0.2402 $(0.0716, 0.3952)$	0.1762 $(0.0892, 0.2760)$
$\rho_{14\|2=0,3=1}$	0	-0.0820 (**0**) $(-0.9732, 0.7825)$	0.2213 (**0**) $(-0.2898, 0.6952)$	0.3221 (**0**) $(-0.1069, 0.7252)$	-0.1320 (**0**) $(-0.3667, 0.1015)$
$\rho_{14\|2=1,3=0}$	0.4	0.3597 (**0**) $(-0.0506, 0.7298)$	0.4222 (**0**) $(0.2604, 0.5791)$	0.3897 $(0.2831, 0.4978)$	0.4081 $(0.2664, 0.4256)$
$\rho_{14\|2=1,3=1}$	0	-0.4221 (**0**) $(-0.9568, 0.2309)$	0.0001 (**0**) $(-0.3151, 0.3156)$	-0.0928 (**0**) $(-0.3397, 0.1243)$	-0.0155 (**0**) $(-0.1703, 0.1348)$
$\rho_{25\|34}$	0.3	0.3614 $(0.1094, 0.5848)$	0.3119 $(0.2035, 0.4211)$	0.2670 $(0.1602, 0.3175)$	0.2662 $(0.1639, 0.3052)$
$\rho_{36\|45}$	-0.7	-0.5811 $(-0.7580, -0.6889)$	-0.5959 $(-0.7604, -0.4958)$	-0.6040 $(-0.7174, -0.5036)$	-0.6938 $(-0.7891, -0.6534)$
$\rho_{15\|234}$	0.2	-0.0645 (**0**) $(-0.2913, 0.2931)$	0.1573 $(0.0040, 0.2577)$	0.1459 $(0.0051, 0.2338)$	0.1689 $(0.1321, 0.2252)$
$\rho_{26\|345}$	0.4	0.4642 $(0.2849, 0.5511)$	0.3799 $(0.3228, 0.5355)$	0.4050 $(0.3481, 0.4931)$	0.3619 $(0.3556, 0.4585)$
$\rho_{16\|2345}$	0.2	0.2675 (**0**) $(-0.0680, 0.3153)$	0.1429 (**0**) $(-0.0407, 0.2162)$	0.1282 (**0**) $(-0.0310, 0.2139)$	0.1719 $(0.0795, 0.2425)$

Table B.2: *The estimated parameters values of the fully-optimized Copula-Vine model for $N =$ 2000*

Parameter	True Value	Full. Optim.	Parameter	True Value	Full. Optim.
ρ_{12}	0.3	0.3153	$\rho_{46\|5}$	0.8	0.7963
ρ_{23}	0	0	$\rho_{14\|2=0,3=0}$	0.1	0.1964
ρ_{34}	−0.2	−0.2986	$\rho_{14\|2=0,3=1}$	0	0
ρ_{45}	−0.4	−0.4131	$\rho_{14\|2=1,3=0}$	0.4	0.4483
ρ_{56}	0.5	0.4359	$\rho_{14\|2=1,3=1}$	0	0
$\rho_{13\|2=0}$	−0.5	−0.5090	$\rho_{25\|34}$	0.3	0.1820
$\rho_{13\|2=1}$	0.7	0.7991	$\rho_{36\|45}$	−0.7	−0.7503
$\rho_{24\|3=0}$	0.6	0.5053	$\rho_{15\|234}$	0.2	0.1827
$\rho_{24\|3=1}$	−0.3	−0.2726	$\rho_{26\|345}$	0.4	0.3817
$\rho_{35\|4}$	−0.3	−0.2987	$\rho_{16\|2345}$	0.2	0.1521

2. Wrong Constant Copula Assumption in Example 2.3.1.

Table B.3: *The estimated bivariate Normal copulas' parameters values where all conditional copulas are assumed to be constant with respect to the conditioning variable(s).*

Parameter	True Value	$N = 100$	$N = 500$	$N = 1000$	$N = 2000$
		Estimate (Conf. Bound)	Estimate (Conf. Bound)	Estimate (Conf. Bound)	Estimate (Conf. Bound)
ρ_{12}	0.3	0.5521 (0.2658, 0.8061)	0.3744 (0.2427, 0.5044)	0.3835 (0.2915, 0.4755)	0.2998 (0.1935, 0.3325)
ρ_{23}	0	−0.2063 (**0**) (−0.5076, 0.1214)	−0.1464 (**0**) (−0.2870, 0.0704)	−0.0977 (**0**) (−0.2015, 0.0090)	−0.0123 (**0**) (−0.0819, 0.0592)
ρ_{34}	−0.2	−0.2465 (**0**) (−0.4992, 0.0454)	−0.1892 (−0.2978, −0.0850)	−0.2552 (0.3339, −0.1802)	−0.2384 (−0.2548, −0.1420)
ρ_{45}	−0.4	−0.5120 (−0.6500, −0.3554)	−0.4298 (−0.4966, −0.3525)	−0.4182 (−0.4712, −0.3642)	−0.4229 (−0.4608, −0.3891)
ρ_{56}	0.5	0.4686 (0.2973, 0.6120)	0.5263 (0.4561, 0.5878)	0.4892 (0.4386, 0.5345)	0.4861 (0.4781, 0.5401)
$\rho_{13\mid 2=0}$	−0.5	0.3144 (**0**) (−0.0840, 0.6994)	0.3143 (0.1155, 0.4591)	0.2320 (0.1608, 0.3757)	0.2986 (0.2092, 0.3873)
$\rho_{13\mid 2=1}$	0.7				
$\rho_{24\mid 3=0}$	0.6	0.1639 (**0**) (−0.0939, 0.3972)	0.1799 (0.0830, 0.2959)	0.2731 (0.1990, 0.3418)	0.3212 (0.2656, 0.3638)
$\rho_{24\mid 3=1}$	−0.3				
$\rho_{35\mid 4}$	−0.3	−0.2526 (−0.5038, −0.0309)	−0.3325 (−0.4310, −0.2163)	−0.3526 (−0.4295, −0.2773)	−0.2484 (−0.3397, −0.2281)
$\rho_{46\mid 5}$	0.8	0.7422 (0.6743, 0.8200)	0.7844 (0.7465, 0.8125)	0.8024 (0.7782, 0.8225)	0.8003 (0.7787, 0.8123)
$\rho_{14\mid 2=0,3=0}$	0.1	0.1688 (**0**) (−0.2130, 0.3147)	0.0113 (**0**) (−0.1327, 0.1297)	0.0126 (**0**) (−0.0927, 0.1022)	0.0410 (**0**) (−0.0476, 0.1202)
$\rho_{14\mid 2=0,3=1}$	0				
$\rho_{14\mid 2=1,3=0}$	0.4				
$\rho_{14\mid 2=1,3=1}$	0				
$\rho_{25\mid 34}$	0.3	0.4197 (0.1584, 0.6146)	0.3494 (0.1823, 0.3844)	0.2823 (0.1857, 0.3356)	0.2267 (0.1420, 0.2872)
$\rho_{36\mid 45}$	−0.7	−0.5811 (−0.7580, −0.6889)	−0.5959 (−0.7604, −0.4958)	−0.6040 (−0.7174, −0.5036)	−0.6256 (−0.7199, −0.5785)
$\rho_{15\mid 234}$	0.2	−0.1089 (**0**) (−0.3337, 0.1962)	0.1851 (0.0627, 0.2838)	0.1675 (0.0543, 0.2593)	0.1831 (0.0919, 0.2351)
$\rho_{26\mid 345}$	0.4	0.5296 (0.3487, 0.6199)	0.3992 (0.3413, 0.5437)	0.4431 (0.3774, 0.5291)	0.3895 (0.3179, 0.4875)
$\rho_{16\mid 2345}$	0.2	0.1547 (**0**) (−0.0456, 0.3757)	0.1387 (**0**) (−0.0421, 0.2137)	0.0981 (**0**) (−0.0550, 0.1811)	0.1930 (0.0971, 0.2699)

3. Copula Misspecification in Example 2.3.2.

Table B.4: *The estimated bivariate Normal copulas' parameters values. The "True Value" columns present the correlations which correspond to the true Clayton parameters θ.*

Parameter	True Value	$N = 100$	$N = 500$	$N = 1000$	$N = 2000$
		Estimate (Conf. Bound)	Estimate (Conf. Bound)	Estimate (Conf. Bound)	Estimate (Conf. Bound)
ρ_{12}	0.7	0.8815 (0.6381, 0.9875)	0.7791 (0.6889, 0.8518)	0.7465 (0.6845, 0.8016)	0.7057 (0.6616, 0.7506)
ρ_{23}	0	−0.0925 (**0**) (−0.4075, 0.2275)	−0.0520 (**0**) (−0.1925, 0.0966)	−0.0327 (**0**) (−0.0913, 0.1023)	−0.0234 (**0**) (−0.0953, 0.0513)
ρ_{34}	0.5	0.4656 (**0**) (0.2498, 0.6833)	0.5402 (0.4481, 0.6272)	0.5247 (0.4637, 0.5928)	0.5173 (0.4701, 0.5656)
ρ_{45}	0.5	0.5563 (0.3965, 0.6780)	0.4728 (0.4037, 0.5333)	0.4881 (0.4423, 0.5346)	0.4793 (0.4448, 0.5103)
ρ_{56}	0.8	0.7797 (0.6931, 0.8453)	0.7423 (0.7003, 0.8197)	0.7605 (0.7363, 0.8054)	0.7679 (0.7501, 0.8066)
$\rho_{13\|2=0}$	0	−0.0463 (**0**) (−0.7849, 0.7472)	−0.0581 (**0**) (−0.4235, 0.3212)	−0.1666 (**0**) (−0.4205, 0.0500)	−0.0403 (**0**) (−0.2950, 0.0632)
$\rho_{13\|2=1}$	0.8	0.6068 (0.2167, 0.9062)	0.6831 (0.5488, 0.8066)	0.7138 (0.6170, 0.8166)	0.7230 (0.6677, 0.8012)
$\rho_{24\|3=0}$	0.2	0.1830 (**0**) (−0.1263, 0.4910)	0.1584 (0.0328, 0.2873)	0.1379 (0.0432, 0.2355)	0.1479 (0.0834, 0.2131)
$\rho_{24\|3=1}$	0.8	0.8258 (0.5174, 0.9790)	0.8200 (0.7288, 0.9071)	0.8052 (0.7303, 0.8703)	0.8098 (0.7603, 0.8552)
$\rho_{35\|4}$	0.6	0.5353 (0.3369, 0.7263)	0.6183 (0.5362, 0.7054)	0.6072 (0.5418, 0.6664)	0.6089 (0.5937, 0.6524)
$\rho_{46\|5}$	0.3	0.3181 (0.1333, 0.4854)	0.3011 (0.2177, 0.3766)	0.2861 (0.2271, 0.3428)	0.2884 (0.2489, 0.3306)
$\rho_{14\|2=0,3=0}$	0.3	0.4237 (**0**) (−0.3182, 0.9678)	0.2054 (**0**) (−0.0345, 0.4685)	0.1655 (0.0952, 0.3803)	0.1255 (0.0403, 0.3205)
$\rho_{14\|2=0,3=1}$	0.6	0.5068 (**0**) (−0.2102, 0.9457)	0.8338 (0.5638, 0.9875)	0.7574 (0.5814, 0.9206)	0.7028 (0.5424, 0.7442)
$\rho_{14\|2=1,3=0}$	0.7	0.5481 (**0**) (−0.0508, 0.9587)	0.8002 (0.6210, 0.8774)	0.7452 (0.6807, 0.8095)	0.7528 (0.6920, 0.7975)
$\rho_{14\|2=1,3=1}$	0.8	−0.7679 (**0**) (−0.0828, 0.9787)	0.8239 (0.6252, 0.9694)	0.8050 (0.6525, 0.9291)	0.8421 (0.7586, 0.9163)
$\rho_{25\|34}$	0.2	0.3181 (**0**) (−0.1693, 0.3641)	0.2954 (0.0780, 0.3454)	0.2850 (0.1147, 0.3152)	0.2711 (0.1208, 0.2380)
$\rho_{36\|45}$	0.2	0.0021 (**0**) (−0.2804, 0.3103)	0.1550 (**0**) (−0.1464, 0.2785)	0.1756 (0.0035, 0.2360)	0.2178 (0.0703, 0.2066)
$\rho_{15\|234}$	0.5	0.3964 (0.0230, 0.5454)	0.4927 (0.2120, 0.5694)	0.4925 (0.2413, 0.5449)	0.4560 (0.3241, 0.5090)
$\rho_{26\|345}$	0.1	0.1366 (**0**) (−0.2520, 0.2046)	0.0778 (**0**) (−0.0797, 0.1707)	0.0903 (0.0130, 0.2175)	0.0935 (0.0331, 0.1331)
$\rho_{16\|2345}$	0.2	0.1565 (**0**) (−0.1785, 0.2717)	0.1482 (0.0305, 0.3316)	0.1180 (0.0712, 0.2989)	0.1238 (0.0788, 0.2118)

Table B.5: *The estimated parameters values of the second model where the continuous pairs are modeled with the Clayton copula. For the pairs modeled by the bivariate Normal copulas, the values in the "True Value" cells correspond to the correlations which correspond to the true Clayton parameters θ.*

Parameter	True Value	Estimate (Conf. Bound)	Parameter	True Value	Estimate (Conf. Bound)
ρ_{12}	0.7	0.7057 $(0.6616, 0.7506)$	$\theta_{46\|5}$	0.5109	0.4857 $(0.4195, 0.5564)$
ρ_{23}	0	-0.0234 (**0**) $(-0.0953, 0.0513)$	$\rho_{14\|2=0,3=0}$	0.3	0.1255 $(0.0403, 0.3205)$
ρ_{34}	0.5	0.5173 $(0.4701, 0.5656)$	$\rho_{14\|2=0,3=1}$	0.6	0.7028 (**0**) $(0.5424, 0.7442)$
θ_{45}	1.0759	1.0067 $(0.9188, 1.0948)$	$\rho_{14\|2=1,3=0}$	0.7	0.7528 $(0.6920, 0.7975)$
θ_{56}	3.1819	3.0584 $(2.9061, 3.2203)$	$\rho_{14\|2=1,3=1}$	0.8	0.8421 $(0.7586, 0.9163)$
$\rho_{13\|2=0}$	0	-0.0403 (**0**) $(-0.2950, 0.0632)$	$\rho_{25\|3,4}$	0.2	0.2686 $(0.1795, 0.3121)$
$\rho_{13\|2=1}$	0.8	0.7230 $(0.6677, 0.8012)$	$\rho_{36\|4,5}$	0.2	0.2321 $(0.1731, 0.2742)$
$\rho_{24\|3=0}$	0.2	0.1479 $(0.0834, 0.2131)$	$\rho_{15\|2,3,4}$	0.5	0.4579 $(0.3243, 0.4511)$
$\rho_{24\|3=1}$	0.8	0.8098 $(0.7603, 0.8552)$	$\rho_{26\|3,4,5}$	0.1	0.1032 $(0.0816, 0.1996)$
$\rho_{35\|4}$	0.6	0.6093 $(0.5947, 0.6543)$	$\rho_{16\|2345}$	0.2	0.1378 $(0.0808, 0.2197)$

4. Vine Structure Misspecification in Example 2.3.3.

Table B.6: *The parameters of the C-Vine structure.*

Parameter	Value	Parameter	Value	Parameter	Value
ρ_{12}	0.5	$\rho_{34\|1=0,2=1}$	-0.2	$\rho_{45\|1=1,2=1,3=0}$	-0.4
ρ_{13}	0.7	$\rho_{34\|1=1,2=0}$	-0.7	$\rho_{45\|1=0,2=0,3=1}$	-0.6
ρ_{14}	0.4	$\rho_{34\|1=1,2=1}$	0.8	$\rho_{45\|1=0,2=1,3=1}$	-0.4
ρ_{15}	-0.2	$\rho_{35\|1=0,2=0}$	0.3	$\rho_{45\|1=1,2=0,3=1}$	-0.7
ρ_{16}	0.8	$\rho_{35\|1=0,2=1}$	-0.6	$\rho_{45\|1=1,2=1,3=1}$	-0.5
$\rho_{23\|1=0}$	0.4	$\rho_{35\|1=1,2=0}$	-0.5	$\rho_{46\|1=0,2=0,3=0}$	0.3
$\rho_{23\|1=1}$	0.7	$\rho_{35\|1=1,2=1}$	-0.6	$\rho_{46\|1=0,2=1,3=0}$	0.8
$\rho_{24\|1=0}$	0	$\rho_{36\|1=0,2=0}$	0.3	$\rho_{46\|1=1,2=0,3=0}$	0.5
$\rho_{24\|1=1}$	0.3	$\rho_{36\|1=0,2=1}$	0.6	$\rho_{46\|1=1,2=1,3=0}$	0.3
$\rho_{25\|1=0}$	0.3	$\rho_{36\|1=1,2=0}$	0.4	$\rho_{46\|1=0,2=0,3=1}$	0.7
$\rho_{25\|1=1}$	-0.8	$\rho_{36\|1=1,2=1}$	0.2	$\rho_{46\|1=0,2=1,3=1}$	0.4
$\rho_{26\|1=0}$	0.4	$\rho_{45\|1=0,2=0,3=0}$	-0.7	$\rho_{46\|1=1,2=0,3=1}$	0.3
$\rho_{26\|1=1}$	0.7	$\rho_{45\|1=0,2=1,3=0}$	-0.4	$\rho_{46\|1=1,2=1,3=1}$	0.6
$\rho_{34\|1=0,2=0}$	0.5	$\rho_{45\|1=1,2=0,3=0}$	-0.6	$\rho_{56\|1234}$	0.8

Table B.7: *The estimated parameters values of the D-Vine structure.*

Parameter	$N = 100$ Estimate (Conf. Bound)	$N = 500$ Estimate (Conf. Bound)	$N = 1000$ Estimate (Conf. Bound)	$N = 2000$ Estimate (Conf. Bound)
ρ_{12}	0.7331 (0.5053, 0.9195)	0.8187 (0.6125, 0.8485)	0.7329 (0.6525, 0.7652)	0.7199 (0.6692, 0.7599)
ρ_{23}	0.6755 (0.4317, 0.8674)	0.6678 (0.4985, 0.7025)	0.6746 (0.6035, 0.7395)	0.6717 (0.6255, 0.7227)
ρ_{34}	0.7176 (0.5611, 0.8529)	0.8092 (0.6452, 0.8312)	0.8080 (0.7325, 0.8552)	0.8173 (0.7930, 0.8421)
ρ_{45}	0.4267 (0.2519, 0.5790)	0.5645 (0.3895, 0.6032)	0.5496 (0.4978, 0.5729)	0.5322 (0.5014, 0.5640)
ρ_{56}	−0.4458 (−0.6024, −0.2661)	−0.5008 (−0.5425, −0.3785)	−0.4774 (−0.5309, −0.4125)	−0.4874 (−0.5210, −0.4539)
$\rho_{13\|2=0}$	−0.0842 (**0**) (−0.6598, 0.5785)	−0.0479 (**0**) (−0.5124, 0.4125)	0.0587 (**0**) (−0.1785, 0.3103)	0.1152 (**0**) (−0.0554, 0.2950)
$\rho_{13\|2=1}$	0.6534 (0.1985, 0.9007)	0.4472 (0.2985, 0.6252)	0.3733 (0.2875, 0.5462)	0.4179 (0.3242, 0.5248)
$\rho_{24\|3=0}$	0.4146 (0.0335, 0.6134)	0.5149 (0.2175, 0.5978)	0.5425 (0.3785, 0.6058)	0.5001 (0.4128, 0.5823)
$\rho_{24\|3=1}$	0.9060 (0.6510, 0.9840)	0.8194 (0.7102, 0.9421)	0.7898 (0.7348, 0.8247)	0.7878 (0.7510, 0.8240)
$\rho_{35\|4}$	−0.1238 (**0**) (−0.4227, 0.1835)	−0.0875 (**0**) (−0.2121, 0.0875)	−0.0425 (**0**) (−0.1345, 0.0545)	−0.0234 (**0**) (−0.0958, 0.0497)
$\rho_{46\|5}$	−0.0459 (**0**) (−0.2474, 0.1484)	−0.1522 (−0.2102, −0.0245)	−0.1329 (−0.1952, −0.0452)	−0.1248 (−0.1670, −0.0840)
$\rho_{14\|2=0,3=0}$	0.3951 (**0**) (−0.4536, 0.9878)	0.3146 (0.0425, 0.4725)	0.3709 (0.1020, 0.4899)	0.3911 (0.2232, 0.5614)
$\rho_{14\|2=0,3=1}$	0.6817 (**0**) (−0.0856, −0.9876)	0.5980 (0.5038, 0.9751)	0.5796 (0.5201, 0.8452)	0.5297 (0.5341, 0.8176)
$\rho_{14\|2=1,3=0}$	0.6312 (**0**) (−0.1259, 0.9658)	0.5834 (0.5148, 0.9815)	0.6024 (0.5236, 0.8245)	0.5841 (0.4759, 0.7738)
$\rho_{14\|2=1,3=1}$	0.6297 (**0**) (−0.0786, 0.9778)	0.6534 (0.4792, 0.9133)	0.6129 (0.5215, 0.7985)	0.6297 (0.5786, 0.6778)
$\rho_{25\|34}$	−0.0211 (**0**) (−0.3550, 0.3837)	−0.2758 (−0.3485, −0.0124)	−0.2817 (−0.3023, −0.0730)	−0.2101 (−0.2833, −0.1421)
$\rho_{36\|45}$	−0.0968 (**0**) (−0.41010.1981)	−0.0752 (**0**) (−0.2164, 0.0885)	−0.0235 (**0**) (−0.1102, 0.0652)	−0.0184 (**0**) (−0.0951, 0.0561)
$\rho_{15\|234}$	0.0134 (**0**) (−0.2556, 0.3906)	0.0562 (**0**) (−0.1212, 0.3054)	0.0452 (**0**) (−0.0625, 0.1714)	0.0304 (**0**) (−0.0324, 0.1242)
$\rho_{26\|345}$	−0.0492 (**0**) (−0.3679, 0.3741)	−0.0578 (**0**) (−0.2127, 0.1009)	−0.0273 (**0**) (−0.1202, 0.0425)	−0.0600 (−0.0912, −0.0331)
$\rho_{16\|2345}$	−0.2540 (**0**) (−0.5424, 0.0752)	−0.5743 (−0.6356, −0.3215)	−0.5777 (−0.6215, −0.4958)	−0.5102 (−0.5764, −0.4751)

Table B.8: *The estimated parameters values of the second model where the continuous pairs are modeled with non-constant conditional Normal copula.*

Parameter	Estimate (Conf. Bound)	Parameter	Estimate (Conf. Bound)
ρ_{12}	0.7199 (0.6692, 0.7599)	A	−0.5194 (−0.6720, −0.3725)
ρ_{23}	0.6717 (0.6255, 0.7227)	b	0.1349 (0.0462, 0.2241)
ρ_{34}	0.8173 (0.7930, 0.8421)	$\rho_{14\|2=0,3=0}$	0.3911 (0.2232, 0.5614)
ρ_{45}	0.5322 (0.5014, 0.5640)	$\rho_{14\|2=0,3=1}$	0.5297 (0.5341, 0.8176)
ρ_{56}	−0.4874 (−0.5210, −0.4539)	$\rho_{14\|2=1,3=0}$	0.5841 (0.4759, 0.7769)
$\rho_{13\|2=0}$	0.1152 (**0**) (−0.0554, 0.2950)	$\rho_{14\|2=1,3=1}$	0.6297 (0.5786, 0.6778)
$\rho_{13\|2=1}$	0.4179 (0.3242, 0.5248)	$\rho_{25\|34}$	−0.2104 (−0.2862, −0.1477)
$\rho_{24\|3=0}$	0.5001 (0.4128, 0.5823)	$\rho_{36\|45}$	−0.1097 (−0.1895, −0.0232)
$\rho_{24\|3=1}$	0.7878 (0.7510, 0.8140)	$\rho_{15\|234}$	−0.0325 (**0**) (−0.1102, 0.0325)
$\rho_{35\|4}$	−0.0240 (**0**) (−0.0992, 0.0466)	$\rho_{26\|345}$	0.0495 (**0**) (−0.0795, 0.0021)
		$\rho_{16\|2345}$	−0.5103 (−0.5779, −0.4778)

APPENDIX C

Parameters of the Disruption Length Models

The tables containing the parameters of the different disruption length models presented in Section 3.2 of this thesis are presented in this Appendix.

Table C.1: *The parameters of the saturated MVN Copula model.*

Parameter	Value	Parameter	Value	Parameter	Value
$\rho_{CT,WD}$	0.1943	$\rho_{WD,CS}$	−0.0203	$\rho_{WM,RH}$	0.1399
$\rho_{CT,LC}$	−0.2840	$\rho_{WD,REP}$	−0.0364	$\rho_{WM,OV}$	0.3648
$\rho_{CT,WT}$	−0.1252	$\rho_{LC,WT}$	0.0529	$\rho_{WM,LAT}$	0.0908
$\rho_{CT,WM}$	0.0211	$\rho_{LC,WM}$	0.0142	$\rho_{WM,CS}$	0.0586
$\rho_{CT,RH}$	−0.0761	$\rho_{LC,RH}$	0.0256	$\rho_{WM,REP}$	0.0016
$\rho_{CT,OV}$	0.0426	$\rho_{LC,OV}$	0.0175	$\rho_{RH,OV}$	0.1666
$\rho_{CT,LAT}$	−0.1019	$\rho_{LC,LAT}$	0.0954	$\rho_{RH,LAT}$	−0.0176
$\rho_{CT,CS}$	0.1589	$\rho_{LC,CS}$	−0.1480	$\rho_{RH,CS}$	0.0182
$\rho_{CT,REP}$	−0.1907	$\rho_{LC,REP}$	0.0749	$\rho_{RH,REP}$	−0.0322
$\rho_{WD,LC}$	−0.0078	$\rho_{WT,WM}$	0.0504	$\rho_{OV,LAT}$	0.1304
$\rho_{WD,WT}$	−0.0882	$\rho_{WT,RH}$	0.5806	$\rho_{OV,CS}$	−0.1487
$\rho_{WD,WM}$	0.0807	$\rho_{WT,OV}$	0.1172	$\rho_{OV,REP}$	−0.0056
$\rho_{WD,RH}$	−0.0664	$\rho_{WT,LAT}$	−0.1309	$\rho_{LAT,CS}$	−0.1633
$\rho_{WD,OV}$	0.1984	$\rho_{WT,CS}$	−0.0634	$\rho_{LAT,REP}$	0.1146
$\rho_{WD,LAT}$	0.1128	$\rho_{WT,REP}$	−0.0231	$\rho_{CS,REP}$	−0.5779

Table C.2: *The parameters of the MVN Copula with conditional independence model.*

Parameter	Value	Parameter	Value	Parameter	Value
$\rho_{CT,WD}$	0.2083	$\rho_{WD,CS}$	0.0436	$\rho_{WM,RH}$	0
$\rho_{CT,LC}$	−0.2700	$\rho_{WD,REP}$	−0.0426	$\rho_{WM,OV}$	0.3508
$\rho_{CT,WT}$	−0.1360	$\rho_{LC,WT}$	0.0674	$\rho_{WM,LAT}$	0.0527
$\rho_{CT,WM}$	0	$\rho_{LC,WM}$	0	$\rho_{WM,CS}$	0
$\rho_{CT,RH}$	−0.0901	$\rho_{LC,RH}$	0.0443	$\rho_{WM,REP}$	0
$\rho_{CT,OV}$	0.0536	$\rho_{LC,OV}$	0.0190	$\rho_{RH,OV}$	0.1542
$\rho_{CT,LAT}$	−0.0910	$\rho_{LC,LAT}$	0.0980	$\rho_{RH,LAT}$	−0.0202
$\rho_{CT,CS}$	0.2091	$\rho_{LC,CS}$	−0.1710	$\rho_{RH,CS}$	−0.0494
$\rho_{CT,REP}$	−0.2047	$\rho_{LC,REP}$	0.0892	$\rho_{RH,REP}$	−0.0167
$\rho_{WD,LC}$	−0.0562	$\rho_{WT,WM}$	0	$\rho_{OV,LAT}$	0.1501
$\rho_{WD,WT}$	−0.0283	$\rho_{WT,RH}$	0.5914	$\rho_{OV,CS}$	0.0001
$\rho_{WD,WM}$	0	$\rho_{WT,OV}$	0.1385	$\rho_{OV,REP}$	−0.0187
$\rho_{WD,RH}$	−0.0188	$\rho_{WT,LAT}$	−0.1347	$\rho_{LAT,CS}$	−0.0208
$\rho_{WD,OV}$	0.2143	$\rho_{WT,CS}$	−0.0798	$\rho_{LAT,REP}$	0.0306
$\rho_{WD,LAT}$	0.1270	$\rho_{WT,REP}$	−0.0318	$\rho_{CS,REP}$	−0.5336

Table C.3: *The parameters of the Copula-Vine Model. The bolded brackets indicate if the parameter value is taken to be zero or constant.*

Parameter	Value (Conf. Bound)	Parameter	Value (Conf. Bound)
$\rho_{CS,CT}$	0.1777 (0.0987, 0.2475)	$\rho_{CS,OV\mid CT=0,WM=1,WT=0,RH=0}$	0.2102 **(0)** (−0.1417, 0.3952)
$\rho_{CT,WM}$	0.0118 **(0)** (−0.2937, 0.3129)	$\rho_{CS,OV\mid CT=1,WM=0,WT=0,RH=0}$	−0.1020 **(0)** (−0.2120, 0.0320)
$\rho_{WM,WT}$	0.0278 **(0)** (−0.2706, 0.3424)	$\rho_{CS,OV\mid CT=0,WM=0,WT=1,RH=1}$	−0.6127 (−0.9212, −0.2012)
$\rho_{WT,RH}$	0.5820 (0.4002, 0.7199)	$\rho_{CS,OV\mid CT=0,WM=1,WT=0,RH=1}$	0.3212 **(0)** (−0.3486, 0.7527)
$\rho_{RH,OV}$	0.1481 (0.0235, 0.3102)	$\rho_{CS,OV\mid CT=1,WM=0,WT=0,RH=1}$	−0.3952 **(0)** (−0.8452, 0.4748)
$\rho_{OV,WD}$	0.1986 (0.0435, 0.3271)	$\rho_{WCS,OV\mid CT=0,WM=1,WT=1,RH=0}$	0.0477 **(0)** (−0.2339, 0.1750)
$\rho_{WD,LC}$	−0.0063 **(0)** (−0.0598, 0.0587)	$\rho_{WM,WD\mid CS,OV\mid CT=1,WM=0,WT=1,RH=0}$	−0.5669 (−0.8325, −0.1201)
$\rho_{LC,LAT}$	0.0883 (0.0182, 0.1557)	$\rho_{CS,OV\mid CT=1,WM=1,WT=0,RH=0}$	−0.2565 **(0)** (−0.4756, 0.2134)
$\rho_{LAT,REP}$	0.1100 (0.0321, 0.1698)	$\rho_{CS,OV\mid CT=0,WM=1,WT=1,RH=1}$	0.0014 **(0)** (−0.8976, 0.9034)
$\rho_{CS,WM\mid CT=0}$	−0.0325 **(0)** (−0.1752, 0.0542)	$\rho_{CS,OV\mid CT=1,WM=0,WT=1,RH=1}$	−0.1725 **(0)** (−0.7671, 0.5441)
$\rho_{CS,WM\mid CT=1}$	0.0675 **(0)** (−0.0742, 0.1323)	$\rho_{CS,OV\mid CT=1,WM=1,WT=1,RH=0}$	0.2994 **(0)** (−0.4374, 0.6749)
$\rho_{CT,WT\mid WM=0}$	−0.1302 **(-0.1102)** (−0.1916, −0.0656)	$\rho_{CT,WD\mid WM=0,WT=0,RH=0,OV=0}$	0.2451 (0.1673, 0.3271)
$\rho_{CT,WT\mid WM=1}$	−0.1127 **(-0.1102)** (−0.3765, 0.1726)	$\rho_{CT,WD\mid WM=0,WT=0,RH=0,OV=1}$	0.1454 **(0)** (−0.2147, 0.5772)
$\rho_{WM,RH\mid WT=0}$	0.3520 (0.2491, 0.4988)	$\rho_{CT,WD\mid WM=0,WT=0,RH=1,OV=0}$	0.2157 (0.0491, 0.4177)
$\rho_{WM,RH\mid WT=1}$	−0.1907 (−0.4875, −0.1005)	$\rho_{CT,WD\mid WM=0,WT=1,RH=0,OV=0}$	0.0947 **(0)** (−0.0487, 0.2451)
$\rho_{WT,OV\mid RH=0}$	0.1785 **(0)** (−0.0345, 0.3055)	$\rho_{CT,WD\mid WM=1,WT=0,RH=0,OV=0}$	0.1087 **(0)** (−0.3595, 0.4995)
$\rho_{WT,OV\mid RH=1}$	−0.2012 **(0)** (−0.4254, 0.0235)	$\rho_{CT,WD\mid WM=0,WT=0,RH=1,OV=1}$	0.0007 **(0)** (−0.7212, 0.3472)
$\rho_{RH,WD\mid OV=0}$	−0.0777 **(-0.0737)** (−0.1321, −0.0279)	$\rho_{CT,WD\mid WM=0,WT=1,RH=0,OV=1}$	0.2595 **(0)** (−0.4938, 0.7619)
$\rho_{RH,WD\mid OV=1}$	−0.1283 **(-0.0737)** (−0.1617, −0.0686)	$\rho_{CT,WD\mid WM=1,WT=0,RH=0,OV=1}$	0.0170 **(0)** (−0.0711, 0.1018)
$\rho_{OV,LC\mid WD}$	0.0105 **(0)** (−0.0633, 0.0875)	$\rho_{CT,WD\mid WM=0,WT=1,RH=1,OV=0}$	0.0671 **(0)** (−0.0785, 0.1995)
$\rho_{WD,LAT\mid LC}$	0.1139 (0.0645, 0.1515)	$\rho_{CT,WD\mid WM=1,WT=0,RH=1,OV=0}$	0.2662 **(0)** (−0.3879, 0.8782)
$\rho_{LC,REP\mid LAT}$	0.0661 (0.0214, 0.1065)	$\rho_{CT,WD\mid WM=1,WT=1,RH=0,OV=0}$	−0.5008 **(0)** (−0.9178, 0.4386)
$\rho_{CS,WT\mid CT=0,WM=0}$	0.0121 **(0)** (−0.1978, 0.2102)	$\rho_{CT,WD\mid WM=0,WT=1,RH=1,OV=1}$	−0.5318 **(0)** (−0.9507, 0.3175)
$\rho_{CS,WT\mid CT=0,WM=1}$	0.4725 (0.2936, 0.6256)	$\rho_{CT,WD\mid WM=1,WT=0,RH=1,OV=1}$	0.7598 **(0)** (−0.7547, 0.9015)
$\rho_{CS,WT\mid CT=1,WM=0}$	−0.1710 (−0.2935, −0.0325)	$\rho_{CT,WD\mid WM=1,WT=1,RH=0,OV=1}$	0.8417 **(0)** (−0.9725, 0.9811)
$\rho_{CS,WT\mid CT=1,WM==1}$	0.4636 (0.3021, 0.5951)	$\rho_{CT,WD\mid WM=1,WT=1,RH=1,OV=0}$	−0.9725 **(0)** (−0.9977, 0.9214)
$\rho_{CT,RH\mid WM=0,WT=0}$	−0.2112 (−0.2561, −0.0372)	$\rho_{CT,WD\mid WM=1,WT=1,RH=1,OV=1}$	−0.0102 **(0)** (−0.0878, 0.0801)
$\rho_{CT,RH\mid WM=0,WT=1}$	0.0912 **(0)** (−0.0294, 0.2307)	$\rho_{WM,LC\mid WT,RH,OV,WD}$	−0.0217 **(0)** (−0.0910, 0.0771)
$\rho_{CT,RH\mid WM=1,WT=0}$	0.4360 (0.1241, 0.6991)	$\rho_{WT,LAT\mid RH,OV,WD,LC}$	−0.1021 **(0)** (−0.1727, 0.0235)
$\rho_{CT,RH\mid WM=1,WT=1}$	−0.2782 **(0)** (−0.6424, 0.2871)	$\rho_{RH,REP\mid OV,WD,LC,LAT}$	0.0958 **(0)** (−0.0325, 0.1695)
$\rho_{WM,OV\mid WT=0,RH=0}$	0.0997 **(0.4520)** (−0.0212, 0.4975)	$\rho_{CS,WD\mid CT=0,WM=0,WT=0,RH=0,OV=0}$	0.0752 **(0)** (−0.0421, 0.1344)
$\rho_{WM,OV\mid WT=0,RH=1}$	0.7267 **(0.4520)** (0.4021, 0.8757)	$\rho_{CS,WD\mid CT=0,WM=0,WT=0,RH=0,OV=1}$	0.1517 (0.0214, 0.2748)
$\rho_{WM,OV\mid WT=1,RH=0}$	0.7027 **(0.4520)** (0.3917, 0.8598)	$\rho_{CS,WD\mid CT=0,WM=0,WT=0,RH=1,OV=0}$	−0.1943 (−0.3074, −0.0875)
$\rho_{WM,OV\mid WT=1,RH=1}$	0.5794 **(0.4520)** (0.2585, 0.7952)	$\rho_{CS,WD\mid CT=0,WM=0,WT=1,RH=0,OV=0}$	−0.0990 **(0)** (−0.1715, 0.0828)
$\rho_{WT,WD\mid RH=0,OV=0}$	−0.1146 **(-0.0774)** (−0.1906, −0.0489)	$\rho_{CS,WD\mid CT=0,WM=1,WT=0,RH=0,OV=0}$	0.1371 **(0)** (−0.1021, 0.2419)
$\rho_{WT,WD\mid RH=0,OV=1}$	−0.1225 **(-0.0774)** (−0.4556, 0.2359)	$\rho_{CS,WD\mid CT=1,WM=0,WT=0,RH=0,OV=0}$	−0.0785 **(0)** (−0.2142, 0.0829)
$\rho_{WT,WD\mid RH=1,OV=0}$	−0.0074 **(-0.0774)** (−0.1304, 0.1020)	$\rho_{CS,WD\mid CT=0,WM=0,WT=0,RH=1,OV=1}$	0.8824 (0.2714, 0.9900)
$\rho_{WT,WD\mid RH=1,OV=1}$	0.2877 **(-0.0774)** (−0.0921, 0.6418)	$\rho_{CS,WD\mid CT=0,WM=0,WT=1,RH=0,OV=1}$	−0.4262 **(0)** (−0.6712, 0.1127)
$\rho_{RH,LC\mid OV,WD}$	0.0213 **(0)** (−0.0687, 0.0767)	$\rho_{CS,WD\mid CT=0,WM=1,WT=0,RH=0,OV=1}$	0.0279 **(0)** (−0.5258, 0.6527)
$\rho_{OV,LAT\mid WD,LC}$	0.1217 (0.0593, 0.2210)	$\rho_{CS,WD\mid CT=1,WM=0,WT=0,RH=0,OV=1}$	−0.4846 (−0.7857, −0.1429)

$\rho_{WD,REP\mid LC,LAT}$	−0.0543 **(0)** (−0.1203, 0.0304)	$\rho_{CS,WD\mid CT=0,WM=0,WT=1,RH=1,OV=0}$	−0.1771 (−0.2678, −0.0892)
$\rho_{CS,RH\mid CT=0,WM=0,WT=0}$	0.0420 **(0)** (−0.1021, 0.1985)	$\rho_{CS,WD\mid CT=0,WM=1,WT=0,RH=1,OV=0}$	−0.3096 (−0.4175, −0.2009)
$\rho_{CS,RH\mid CT=0,WM=0,WT=1}$	0.1987 **(0)** (−0.0325, 0.3201)	$\rho_{CS,WD\mid CT=1,WM=0,WT=0,RH=1,OV=0}$	−0.1634 (−0.2589, −0.0796)
$\rho_{CS,RH\mid CT=0,WM=1,WT=0}$	−0.1325 **(0)** (−0.2325, 0.0212)	$\rho_{CS,WD\mid CT=0,WM=1,WT=1,RH=0,OV=0}$	−0.3523 (−0.4314, −0.2724)
$\rho_{CS,RH\mid CT=1,WM=0,WT=0}$	0.0325 **(0)** (−0.0921, 0.1131)	$\rho_{CS,WD\mid CT=1,WM=0,WT=1,RH=0,OV=0}$	−0.1086 (−0.2017, −0.0074)
$\rho_{CS,RH\mid CT=0,WM=1,WT=1}$	−0.1020 **(0)** (−0.2102, 0.0725)	$\rho_{CS,WD\mid CT=1,WM=1,WT=0,RH=0,OV=0}$	0.1103 (0.0327, 0.1943)
$\rho_{CS,RH\mid CT=1,WM=0,WT=1}$	0.2320 **(0)** (−0.0275, 0.5219)	$\rho_{CS,WD\mid CT=0,WM=0,WT=1,RH=1,OV=1}$	−0.8162 (−0.9900, −0.2108)
$\rho_{CS,RH\mid CT=1,WM=1,WT=0}$	0.1932 **(0)** (0.0125, 0.2932)	$\rho_{CS,WD\mid CT=0,WM=1,WT=0,RH=1,OV=1}$	0.3317 **(0)** (−0.7124, 0.9215)
$\rho_{CS,RH\mid CT=1,WM=1,WT=1}$	−0.3212 **(0)** (−0.5121, 0.0952)	$\rho_{CS,WD\mid CT=1,WM=0,WT=0,RH=1,OV=1}$	−0.5215 **(0)** (−0.9725, 0.3731)
$\rho_{CT,OV\mid WM=0,WT=0,RH=0}$	−0.3378 (−0.5023, −0.1023)	$\rho_{CS,WD\mid CT=0,WM=1,WT=1,RH=0,OV=1}$	−0.3124 **(0)** (−0.8214, 0.4752)
$\rho_{CT,OV\mid WM=0,WT=0,RH=1}$	−0.2152 **(0)** (−0.2045, 0.1055)	$\rho_{CS,WD\mid CT=1,WM=0,WT=1,RH=0,OV=1}$	0.7215 **(0)** (−0.8214, 0.9726)
$\rho_{CT,OV\mid WM=0,WT=1,RH=0}$	0.2325 **(0)** (−0.0620, 0.4546)	$\rho_{CS,WD\mid CT=1,WM=1,WT=0,RH=0,OV=1}$	0.0782 **(0)** (−0.1023, 0.3129)
$\rho_{CT,OV\mid WM=1,WT=0,RH=1}$	−0.5779 **(0)** (−0.7736, 0.0217)	$\rho_{CS,WD\mid CT=0,WM=1,WT=1,RH=1,OV=0}$	0.0952 **(0)** (−0.2147, 0.3258)
$\rho_{CT,OV\mid WM=0,WT=1,RH=1}$	−0.5021 **(0)** (−0.7036, 0.0323)	$\rho_{CS,WD\mid CT=1,WM=1,WT=0,RH=1,OV=0}$	−0.4519 (−0.6044, −0.3095)
$\rho_{CT,OV\mid WM=1,WT=0,RH=1}$	0.5021 **(0)** (−0.2102, 0.8925)	$\rho_{CS,WD\mid CT=1,WM=0,WT=1,RH=1,OV=0}$	−0.2519 **(0)** (−0.5127, 0.3178)
$\rho_{CT,OV\mid WM=1,WT=1,RH=0}$	0.6252 **(0)** (−0.2012, 0.8921)	$\rho_{CS,WD\mid CT=1,WM=1,WT=1,RH=0,OV=0}$	−0.2496 (−0.3458, −0.1370)
$\rho_{CT,OV\mid WM=1,WT=1,RH=1}$	−0.0321 **(0)** (−0.8125, 0.7210)	$\rho_{CS,WD\mid CT=0,WM=1,WT=1,RH=1,OV=1}$	0.2188 **(0)** (−0.4215, 0.7002)
$\rho_{WM,WD\mid WT=0,RH=0,OV=0}$	0.0752 **(0)** (−0.0412, 0.1425)	$\rho_{CS,WD\mid CT=1,WM=0,WT=1,RH=1,OV=1}$	−0.8835 (−0.9896, −0.6201)
$\rho_{WM,WD\mid WT=0,RH=0,OV=1}$	0.9433 (0.7210, 0.9785)	$\rho_{CS,WD\mid CT=1,WM=1,WT=0,RH=1,OV=1}$	0.2142 **(0)** (−0.4215, 0.7017)
$\rho_{WM,WD\mid WT=0,RH=1,OV=0}$	0.2121 **(0)** (−0.0321, 0.3995)	$\rho_{CS,WD\mid CT=1,WM=1,WT=1,RH=0,OV=1}$	−0.4725 **(0)** (−0.9915, 0.3277)
$\rho_{WM,WD\mid WT=1,RH=0,OV=0}$	−0.1209 **(0)** (−0.3205, 0.1009)	$\rho_{CS,WD\mid CT=1,WM=1,WT=1,RH=1,OV=0}$	0.7403 (0.4121, 0.9278)
$\rho_{WM,WD\mid WT=0,RH=1,OV=1}$	0.3021 **(0)** (−0.5785, 0.8925)	$\rho_{CS,WD\mid CT=1,WM=1,WT=1,RH=1,OV=1}$	0.5044 (0.2014, 0.8275)
$\rho_{WM,WD\mid WT=1,RH=0,OV=1}$	0.8215 **(0)** (−0.0625, 0.9795)	$\rho_{CT,LC\mid WM,WT,RH,OV,WD}$	−0.2475 (−0.3004, −0.1778)
$\rho_{WM,WD\mid WT=1,RH=1,OV=0}$	0.2123 **(0)** (−0.0952, 0.4952)	$\rho_{WM,LAT\mid WT,RH,OV,WD,LC}$	0.0371 **(0)** (−0.0462, 0.1014)
$\rho_{WM,WD\mid WT=1,RH=1,OV=1}$	−0.0875 **(0)** (−0.6932, 0.5213)	$\rho_{WT,REP\mid RH,OV,WD,LC,LAT}$	−0.0758 **(0)** (−0.1812, 0.0273)
$\rho_{WT,LC\mid RH,OV,WD}$	0.0492 **(0)** (−0.0402, 0.1146)	$\rho_{CS,LC\mid CT,WM,WT,RH,OV,WD}$	−0.1834 (−0.3148, −0.0492)
$\rho_{RH,LAT\mid OV,WD,LC}$	−0.0329 **(0)** (−0.1374, 0.0319)	$\rho_{CT,LAT\mid WM,WT,RH,OV,WD,LC}$	−0.1176 (−0.1942, −0.0302)
$\rho_{OV,REP\mid WD,LC,LAT}$	0.0425 **(0)** (−0.0752, 0.1276)	$\rho_{WM,REP\mid WT,RH,OV,WD,LC,LAT}$	0.0824 **(0)** (−0.0521, 0.1421)
$\rho_{CS,OV\mid CT=0,WM=0,WT=0,RH=0}$	−0.1021 **(0)** (−0.2102, 0.0195)	$\rho_{CS,LAT\mid CT,WM,WT,RH,OV,WD,LC}$	0.0421 **(0)** (−0.0512, 0.1384)
$\rho_{CS,OV\mid CT=0,WM=0,WT=0,RH=1}$	0.0852 **(0)** (−0.0785, 0.2101)	$\rho_{CT,REP\mid WM,WT,RH,OV,WD,LC,LAT}$	−0.2352 (−0.3575, −0.1175)
$\rho_{CS,OV\mid CT=0,WM=0,WT=1,RH=0}$	0.2193 **(0)** (−0.2852, 0.6012)	$\rho_{CS,REP\mid CT,WM,WT,RH,OV,WD,LC,LAT}$	−0.5090 (−0.6527, −0.3549)

Table C.4: *The parameters of the MVN Copula with conditional independence model for the Switch disruption.*

Parameter	Value	Parameter	Value	Parameter	Value
$\rho_{CT,WD}$	0.0466	$\rho_{WD,CS}$	0.0075	$\rho_{CD,RH}$	−0.0009
$\rho_{CT,LC}$	−0.3689	$\rho_{WD,REP}$	−0.0085	$\rho_{CD,OV}$	0.5548
$\rho_{CT,WT}$	−0.0157	$\rho_{LC,WT}$	0.0060	$\rho_{CD,LAT}$	0.1172
$\rho_{CT,CD}$	0.0179	$\rho_{LC,CD}$	−0.0069	$\rho_{CD,CS}$	0.1109
$\rho_{CT,RH}$	−0.0086	$\rho_{LC,RH}$	0.0033	$\rho_{CD,REP}$	−0.0216
$\rho_{CT,OV}$	0.0157	$\rho_{LC,OV}$	−0.0060	$\rho_{RH,OV}$	0.0110
$\rho_{CT,LAT}$	−0.1220	$\rho_{LC,LAT}$	0.0759	$\rho_{RH,LAT}$	−0.0551
$\rho_{CT,CS}$	0.1546	$\rho_{LC,CS}$	−0.0593	$\rho_{RH,CS}$	−0.0080
$\rho_{CT,REP}$	−0.1737	$\rho_{LC,REP}$	0.0666	$\rho_{RH,REP}$	0.0027
$\rho_{WD,LC}$	−0.0179	$\rho_{WT,CD}$	−0.0017	$\rho_{OV,LAT}$	0.1225
$\rho_{WD,WT}$	−0.0008	$\rho_{WT,RH}$	0.5281	$\rho_{OV,CS}$	0.0643
$\rho_{WD,CD}$	0.0009	$\rho_{WT,OV}$	0.0053	$\rho_{OV,REP}$	−0.0134
$\rho_{WD,RH}$	−0.0004	$\rho_{WT,LAT}$	−0.1409	$\rho_{LAT,CS}$	−0.0044
$\rho_{WD,OV}$	0.1135	$\rho_{WT,CS}$	−0.0146	$\rho_{LAT,REP}$	0.0195
$\rho_{WD,LAT}$	0.1145	$\rho_{WT,REP}$	0.0049	$\rho_{CS,REP}$	−0.1867

Summary

Mixed Discrete - Continuous
Railway Disruption-Length Models with Copulas

Aurelius Armando Zilko

The uncertainty of disruption length has hindered the performance of the Operational Control Centre Rail (OCCR) in Utrecht, whose task is to manage train traffic when a disruption occurs on the Dutch railway network. One way to model that uncertainty is by representing disruption length as a probability distribution. In this way, a dependence model between the disruption length and several observable influencing factors can be constructed. The dependence model takes the form of a joint distribution between the variables. From the joint distribution, the conditional distribution of disruption length can be computed by conditioning the model on the observed values of the influencing factors.

This thesis focuses on the construction of the joint model. The concept of *copula* is intended to be used to model the dependence between the variables. There are several multivariate copula families that can be used. In dimensions higher than two, however, the construction of such a multivariate model is often difficult because of constraints that need to be satisfied. The *copula-vine* approach comes in handy to avoid this problem. With this, the joint model is decomposed into (conditional) pairs that can be modelled with algebraically independent bivariate copulas. The copula-vine is also highly flexible, which enables the construction of a very complicated joint model.

Some of the factors influencing the railway disruption-length model are discrete variables. Consequently, the joint model is of mixed discrete and continuous type. Copulas can still be used to model the dependence when some of the variables are discrete. Moreover, it is known that the copula is not unique under this setting. While this provides more "freedom" for practitioners to choose which copula family to work with, it also comes with two main drawbacks: (i) the copula no longer separates the dependence from the marginal distributions; and (ii) in the presence of data, the copula parameters need to be estimated using

the computationally expensive maximum-likelihood approach.

In the first half of Chapter 2, it is shown that the bivariate normal copula can always be used to model the dependence of a bivariate Bernoulli distribution. However, this is no longer true when the dimension increases. In this case, the copula-vine approach can be used. By considering conditional bivariate copulas that depend on the conditioning variables (hence "non-constant"), it is shown that any multivariate Bernoulli distribution can be modelled with a set of algebraically independent bivariate normal copulas. However, this is no longer true when one or more of the variables have more than two states or are continuous.

Given a vine structure and bivariate copula families, an algorithm is proposed to estimate the parameter values of the copula-vine model from a set of data. This is the focus of the second half of Chapter 2. Several artificially constructed mixed binary and continuous datasets with different structures or copula family misspecifications are considered. To recover the conditional distribution of a dependent continuous variable, we observe the importance in the recovery of the continuous part of the model.

In this thesis, the railway disruption-length model is based on disruption records in the SAP database. This is the main topic of Chapter 3. Eight influencing factors are included in the disruption-length models for incidents caused by track circuit (TC) or switch (points) failures. Two joint model construction strategies are considered: (i) using the popular multivariate normal copula approach; and (ii) using the copula-vine approach, where non-constant conditional copulas are considered in the discrete part of the model. It is shown that, while the copula-vine approach models the joint distribution of the data better than the multivariate normal copula approach, both models recover the conditional distribution of disruption length. For this reason, we opt to use the multivariate normal copula model as this can be implemented in the software `UNINET`, thus making its application more practical.

The constructed model is used to obtain a prediction of disruption length. One value from the conditional distribution of disruption length needs to be chosen as the prediction. To investigate the effect of different choices of prediction, we have collaborated with the Department of Transport and Planning at Delft University of Technology, as presented in Chapter 4. The disruption-length model is used together with the short-turning model and the passenger-flow model in four case studies of disruption in the vicinity of Houten, the Netherlands. Different predictions of disruption length are made and the impact on passengers is measured in terms of total generalized travel time.

The quality of the information in the SAP database is poor. Better data needs to be collected so that the disruption-length model can be extended to achieve better performance. The combination with the short-turning model and the passenger-flow model can also be expanded to obtain more general conclusions. In Chapter 5, several recommendations are provided for better data collection and ways to expand the combination in the future.

Samenvatting

Gemengde discreet-continue spoorwegverstoringsduurmodellen met copula's

Aurelius Armando Zilko

De onzekerheid van de verstoringsduur is een belemmering voor de prestaties van het Operationeel Controle Centrum Rail (OCCR) in Utrecht, dat als taak heeft om het treinverkeer te regelen wanneer er een verstoring optreedt in het Nederlandse spoorwegnetwerk. Een van de manieren om die onzekerheid te modelleren is een representatie van de verstoringsduur als kansverdeling. Op deze manier kan er een afhankelijkheidsmodel worden geconstrueerd voor de relatie tussen de verstoringsduur en diverse waarneembare beïnvloedingsfactoren. Het afhankelijkheidsmodel heeft de vorm van een simultane verdeling van de variabelen. Uit de simultane verdeling kan de voorwaardelijke verdeling van de verstoringsduur worden berekend door de geobserveerde waarden van de beïnvloedingsfactoren als voorwaarden aan het model toe te voegen.

Dit proefschrift behandelt de constructie van het simultane model. Het concept van de *copula* wordt gebruikt om de onderlinge afhankelijkheid van de variabelen te modelleren. Er kunnen verscheidene multivariate copulafamilies worden gebruikt. Bij meer dan twee dimensies is de constructie van een dergelijk multivariaat model echter vaak moeilijk, omdat er aan een aantal restricties dient te worden voldaan. Dit probleem kan worden vermeden met de *copula-vinemethode*. Hierbij wordt het simultane model ontbonden in (voorwaardelijke) paren die kunnen worden gemodelleerd met algebrasch onafhankelijke bivariate copulas. Deze methode is zeer flexibel, zodat er een zeer gecompliceerd simultaan model kan worden geconstrueerd.

Sommige factoren die het spoorwegverstoringsduurmodel benvloeden, zijn discrete variabelen. Daarom is het simultane model van het type gemengd discreet-continu. Ook wanneer sommige variabelen discreet zijn, kunnen copulas worden gebruikt om de afhankelijkheid te modelleren. Bovendien is bekend dat de copula in een dergelijk geval niet uniek is. Dit geeft meer vrijheid in de keuze voor een copulafamilie om mee te werken, maar er zijn ook twee belangrijke nadelen aan verbonden: (i) de copula scheidt de afhankelijkheid niet meer van de marginale verdelingen; (ii) wanneer er data worden gebruikt, moeten de copulaparameters worden geschat met de maximum-likelihoodmethode, die veel rekentijd vergt.

In de eerste helft van hoofdstuk 2 wordt aangetoond dat de bivariate normale copula altijd kan worden gebruikt om de afhankelijkheid van een bivariate Bernoulli-verdeling te modelleren. Bij meer dimensies geldt dit echter niet meer. In dit geval kan de copula-vinemethode worden gebruikt. Door voorwaardelijke bivariate copulas te beschouwen die afhangen van de bepalende variabelen (en die dus niet-constant zijn) laten we zien dat elke multivariate Bernoulli-verdeling kan worden gemodelleerd door een verzameling algebrasch onafhankelijke bivariate normale copulas. Dit geldt echter niet meer wanneer een of meer van de variabelen meer dan twee toestanden hebben of continu zijn.

Bij een gegeven vinestructuur en bivariate copulafamilies stellen we een algoritme voor om de parameterwaarden van het copula-vinemodel te schatten op basis van een dataset. Hierover gaat de tweede helft van hoofdstuk 2. We beschouwen diverse kunstmatig geconstrueerde binaire en continue datasets met verschillende structuren of misspecificaties van copulafamilies. We merken op dat we, om de voorwaardelijke verdeling van een afhankelijke continue variabele te kunnen geven, eerst het continue deel van het model moeten vaststellen.

In dit proefschrift is het spoorwegverstoringsduurmodel gebaseerd op verstoringsrecords in de SAP-database. Dit is het hoofdonderwerp van hoofdstuk 3. Acht beïnvloedingsfactoren worden opgenomen in de verstoringsduurmodellen voor incidenten die zijn veroorzaakt door storingen in spoorstroomlopen of wissels. Er worden twee strategien beschouwd voor de constructie van simultane modellen: (i) de populaire methode met een multivariate normale copula en (ii) de copula-vinemethode, waarbij niet-constante voorwaardelijke copulas worden beschouwd in het discrete deel van het model. We laten zien dat de copula-vinemethode de simultane verdeling van de data beter modelleert dan de methode met een multivariate normale copula, maar dat met beide modellen de voorwaardelijke verdeling van de verstoringsduur kan worden verkregen. Daarom kiezen we ervoor het model van de multivariate normale copula te gebruiken, aangezien dit kan worden gemplementeerd in de `UNINET`-software, wat de toepassing praktischer maakt.

Het geconstrueerde model wordt gebruikt om de verstoringsduur te voorspellen. Als voorspelling moet er n waarde van de voorwaardelijke verdeling van de verstoringsduur worden gekozen. Om het effect van de diverse keuzen voor de voorspelling te onderzoeken, hebben we samengewerkt met de afdeling Transport en Planning van de Technische Universiteit Delft, zoals gepresenteerd in hoofdstuk 4. Het verstoringsduurmodel wordt gebruikt samen met het short-turning (inkortings-)model en het passagiersstroommodel in vier casestudies van verstoring in de buurt van Houten. We doen verschillende voorspellingen voor de verstoringsduur en meten de invloed op passagiers in termen van totale gegeneraliseerde reistijd.

De kwaliteit van de gegevens in de SAP-database is slecht. Er moeten betere data worden verzameld, zodat het verstoringsduurmodel kan worden uitgebreid en betere resultaten kan geven. De combinatie met het short-turningmodel en het passagiersstroommodel kan ook worden uitgebreid om tot meer algemene conclusies te komen. In hoofdstuk 5 doen we diverse aanbevelingen om het verzamelen van data te verbeteren en stellen we manieren voor om de combinatie van de modellen in de toekomst uit te breiden.

Ringkasan

Model Lama Gangguan Kereta Api Campuran Diskret dan Kontinu dengan Kopula

Aurelius Armando Zilko

Ketidak-pastian lama gangguan merupakan masalah besar yang dihadapi Pusat Kontrol Operasional Kereta Api (OCCR) di Utrecht, yang bertugas mengatur lalu-lintas kereta api di saat terjadinya gangguan di jaringan kereta api Belanda. Satu cara untuk menangani ketidak-pastian ini adalah dengan memodelkan lama gangguan sebagai distribusi peluang. Dengan cara ini, sebuah model ketergantungan antara lama gangguan dan beberapa faktor pemengaruh dapat dibangun. Model ketergantungan ini berupa distribusi gabungan dari peubah-peubah yang terlibat. Dari distribusi gabungan ini, distribusi bersyarat dari lama gangguan dapat dihitung dengan menggunakan nilai dari faktor-faktor pemengaruh yang teramati.

Fokus dari disertasi ini adalah pembangunan model gabungan tersebut. Konsep *kopula* akan digunakan untuk memodelkan ketergantungan antara peubah-peubah yang terlibat. Ada beberapa keluarga kopula multivariat yang tersedia. Namun di dimensi lebih dari dua, ada banyak batasan yang menyulitkan pembangunan sebuah model kopula multivariat. Strategi *Kopula-Vine* dapat digunakan untuk menghindari kesulitan-kesulitan ini. Melalui pendekatan ini, model gabungan didekomposisi menjadi pasangan-pasangan (bersyarat) yang dapat dimodelkan dengan kopula-kopula bivariat yang bebas secara aljabar. Keluwesan Kopula-Vine memungkinkan pembangunan suatu model gabungan yang sangat rumit.

Beberapa faktor pemengaruh lama gangguan kereta api ternyata berupa peubah diskret. Sebagai akibatnya, model gabungan yang harus dibangun berupa model campuran antara peubah diskret dan kontinu. Kopula masih bisa digunakan untuk memodelkan ketergantungan antara peubah-peubah diskret. Lebih jauh lagi, telah diketahui secara luas bahwa kopula yang bisa digunakan tidaklah unik di dalam situasi ini. Walaupun di satu sisi ini memberikan "kebebasan” lebih bagi para praktisioner dalam pemilihan keluarga kopula, di sisi lain dua masalah muncul: (1) kopula tidak lagi memisahkan ketergantungan dari efek distribusi marjinal dan (2) dalam lingkup data, nilai parameter dari kopula harus diestimasi dengan menggunakan metode *maximum likelihood* yang memakan waktu lama.

Di paruh pertama Bab 2, ditunjukkan bahwa kopula Normal bivariat dapat selalu digunakan untuk memodelkan ketergantungan sebuah distribusi Bernoulli bivariat. Namun, ini tidak lagi benar di dimensi yang lebih tinggi. Dalam kasus ini, pendekatan Kopula-Vine dapat digunakan. Dengan menggunakan kopula bivariat bersyarat yang juga bergantung pada nilai peubah persyarat (dengan kata lain, “tidak tetap”), ditunjukkan bahwa distribusi Bernoulli multivariat apa pun dapat dimodelkan dengan kopula-kopula Normal bivariat yang bebas secara aljabar. Namun, ini tidak lagi benar ketika satu atau lebih peubahnya memiliki lebih dari dua kemungkinan nilai atau kontinu.

Dengan sebuah struktur vine dan keluarga kopula bivariat yang telah dipilih, sebuah algoritma untuk mengestimasi nilai-nilai parameter sebuah model Kopula-Vine dari satu himpunan data diusulkan. Ini adalah tema utama paruh kedua Bab 2. Untuk mengetes performa dari algoritma ini, beberapa himpunan data dengan peubah campuran biner dan kontinu dibangun dengan struktur atau keluarga kopula yang tidak terpilih dengan benar. Ternyata agar distribusi bersyarat dari sebuah peubah kontinu termodelkan dengan baik, bagian kontinu dari model harus termodelkan dengan baik pula.

Di dalam disertasi ini, model lama gangguan kereta api dibangun berdasarkan data gangguan yang terekam di dalam bank data SAP. Konstruksi model ini adalah topik utama Bab 3. Delapan faktor pemengaruh lama gangguan kereta api terlibat dalam model gangguan yang diakibatkan oleh rusaknya sirkuit deteksi kereta (TC) atau wesel. Model gabungan dibangun dengan menggunakan dua strategi: (1) melalui pendekatan kopula Normal multivariat yang umum dipakai; dan (2) melalui pendekatan Kopula-Vine dengan penggunaan kopula bersyarat tidak tetap di bagian diskret dari model. Model distribusi gabungan yang dibangun dengan Kopula-Vine mewakili data dengan lebih baik daripada model yang dibangun dengan kopula Normal multivariat. Namun, kedua model sama-sama mampu menghasilkan distribusi bersyarat dari lama gangguan dengan baik. Oleh karena itu, model kopula Normal multivariat dipilih di disertasi ini. Model ini dapat diimplementasikan di perangkat lunak `UNINET` yang memungkinkan penerapan model yang jauh lebih praktis.

Model yang telah dibangun digunakan untuk membuat sebuah prediksi lama gangguan. Satu nilai dari distribusi bersyarat lama gangguan perlu dipilih sebagai prediksi. Untuk mempelajari efek dari prediksi-prediksi yang dapat dibuat, kolaborasi dengan Jurusan Transportasi dan Perencanaan Universitas Teknik Delft dilakukan. Kolaborasi ini dipresentasikan di Bab 4. Model lama gangguan digunakan bersamaan dengan model putar-balik kereta dan model arus penumpang di beberapa studi kasus gangguan kereta api di daerah Houten, Belanda tengah. Beberapa prediksi lama gangguan dipilih dan dampaknya pada penumpang diukur dari segi total waktu tempuh tertimbang (*generalized travel time*).

Informasi yang terekam di dalam bank data SAP berkualitas rendah. Data yang lebih baik perlu dikumpulkan sehingga dengannya model lama gangguan dapat diekspansi untuk menghasilkan performa yang lebih baik. Kolaborasi dengan model putar-balik kereta dan model arus penumpang juga dapat dikembangkan untuk mendapatkan kesimpulan yang lebih umum. Di Bab 5, disajikan beberapa rekomendasi untuk pengumpulan data yang lebih baik dan bagaimana pengembangan kolaborasi dapat dilakukan di masa mendatang.

Acknowledgments

I still remember that evening in early March 2010. It was a Monday evening, two days after my Bachelor's graduation ceremony in Bandung, Indonesia. I went to my brother's room to borrow his laptop (and internet connection) to check my email before going to bed. Little did I know that a lifelong dream would start to materialize that evening. It began with a new email in my Inbox folder. An email that would become the landmark of a turning point event of my life. An email which content would set the course of my personal and professional life in the years to come. An email that proved dreams do come true.

It was an email from Dr. Dorota Kurowicka, informing me that I was granted the Risk and Environmental Modeling scholarship to study at Delft University of Technology, the Netherlands.

Then ... the rest is history.

***

First and foremost, I must thank Dr. Dorota Kurowicka. It is not only for the opportunities[1] that have been given to me but also for the supervision and patience during the four years of the PhD research. I have learned a lot in the past four years and I do believe the experience has shaped me into a better professional. I would also like to express my gratitude to Dr. Anca Hanea who was my Master's thesis supervisor and part of my supervisory team in the first two years of this PhD research. Thank you for your belief in me which led me into this position. I also thank Prof. Frank Redig who has been my promotor throughout the research. I want to appreciate my thesis committee for their feedback and time they have invested to improve the quality of this thesis. Finally, I would also want to extend my gratitude to Prof. Roger Cooke whom, without him, none of this would have been a reality.

While I have always been fascinated by the world of transportation (especially aviation and train[2]), I had never had the opportunity to come close to the industry prior to this research. I must thank Dr. Rob Goverde for initiating the research project and eventually introduced me to the world of railway traffic

---

[1] Notice the plural form.

[2] In my childhood, I was obsessed with train.

management. I must also thank Dirk Kes from ProRail who has provided tremendous support and connections within the huge organization of ProRail and beyond; and also for his patience to our constant (never-ending) request for better data. I owe an important research progress to Erwin van Wonderen from ProRail who was willing to help and supervise me in manually reading the SAP database for the TC disruptions to overcome the very low quality of the data. I would also like to acknowledge Theo Kruse and Zaven Jessayan who were willing to help me do the same for the switch disruptions. Moreover, thank you André Duinmeijer for providing us with the SAP database. Finally, I would like to extend my gratitude to the very many ProRail people and fellow ExploRail researchers whom I have been in contact with in the course of the research; especially Manon Kiers, Tjitske Bezema, and Brenda Struve who have worked hard to support the ExploRail researches.

I am fortunate that I am not really "alone" in this PhD research. I must thank Nadjla Ghaemi, who works on the other half of the SmartOCCR project, for the four (plus) years of working together. Moreover, I would want to especially thank and appreciate Fei Yan for her hard work and tenacity she had to put and endure in the last a few months of my PhD contract. I also express my gratitude to all the people whom I have shared offices with: Patrycja, Weiwei, Yanbin, Eni, Jeroen, and Lixue; for the nice atmosphere and discussions we have had together. The same gratitude goes to all colleagues (and ex-colleagues) in the Department of Applied Mathematics. Abundant amount of appreciation also goes to Carl, Cindy, Roniet, Evelyn, Léonie, and Stefanie for the constant (organizational) support they have provided throughout the years in TU Delft.

I am very grateful for my lifelong friends. I feel blessed for the time we have spent together, for the discussions, for the constant support to each other, for the fun, for the adventure all around the world, and, more importantly, for helping me to keep myself sane throughout the thick and thin of all these years. My most sincere gratitude goes to all of you; including but not limited to Carolyn, Michiel, Preethi, Dana, Boris, Dan, Diana, Omar, Renshi, Zhe, Indra S., Hartono, Evan, Ken, Yuri, Raphael, and everyone else. Unfortunately I must stop here because it is impossible for me to name all of you as the list could go on for pages to come[3].

Of course, I must thank my family, especially my parents, for their infinite amount of love, support, and belief in me; even at times when I had little to none myself. This has greatly helped me to reach this point, where this thesis is, finally, nearly fully concluded. Thank you.

Delft, December 2016

Aurelius A. Zilko

---

[3] I am sure, and hope, that you would understand.

Curriculum Vitae

Aurelius Armando Zilko was born in Yogyakarta, Indonesia on September the 7^{th}, 1988. After finishing high school in June 2006, he moved to Bandung, Indonesia to study mathematics at Parahyangan Catholic University with the Talent and Interest scholarship. Seven semesters later in February 2010, he graduated with a (cum laude) Bachelor's degree with the thesis: *Inverse Eigenvalue Problems of Sturm-Liouville Differential Equation*.

In August 2010, he moved to the Netherlands to continue his study in applied mathematics at Delft University of Technology with the Risk and Environmental Modeling scholarship. Two years later in August 2012, he graduated with a (cum laude) Master's degree with the thesis: *Non-Parametric Bayesian Networks (NPBNs) versus Ensemble Kalman Filter (EnKF) in Reservoir Simulation with non-Gaussian Measurement Noise*.

In September 2012, he became a PhD candidate in the Applied Probability group of Delft University of Technology under the supervision of Dr. Dorota Kurowicka and Dr. Anca Hanea. The position was funded through ExploRail, a partnership programme of the Dutch Technology Foundation STW and ProRail, project no. 12257: "Smart Information and Decision Support for Railway Operation Control Centres (SmartOCCR)". The project was done in collaboration with the Department of Transport and Planning of Delft University of Technology and ProRail. This thesis is the outcome of the four years of this PhD research.

Since October 2016, he is a data scientist in the Partner Services Analytics Department at Booking.com BV in Amsterdam, the Netherlands.

List of Publications

Journal publications:

Zilko, A. A. and Kurowicka, D. (2016). Copula in a Multivariate Mixed Discrete-Continuous Model. *Computational Statistics & Data Analysis*, 103:28-55.

Zilko, A. A., Kurowicka, D., and Goverde, R. M. P. (2016). Modeling Railway Disruption Lengths with Copula Bayesian Networks. *Transportation Research Part C: Emerging Technologies*, 68:350-368.

Ghaemi, N., Zilko, A. A., Yan, F., Cats, O., Kurowicka, D., and Goverde, R. M. P. (2016). Impact of Railway Disruption Predictions and Rescheduling Strategies on Passenger Delays. *Submitted to Transportation Research Part C: Emerging Technologies*.

Conference papers and presentations:

Zilko, A. A., Hanea, A.M., and Hanea, R.G. (2013). Non-Parametric Bayesian Network for Parameter Estimation in Reservoir Engineering. In *Proc. 75th EAGE Conference & Exhibition incorporating SPE EUROPEC 2013 (EAGE 2013, London, England, June 2013).*

Zilko, A. A., Hanea, A. M., Kurowicka, D., and Goverde, R. M. P. (2014). Non-Parametric Bayesian Network to Forecast Railway Disruption Lengths. In *Proc. 2nd International Conference on Railway Technology: Research, Development and Maintenance (Railways 2014, Ajaccio, France, April 2014).*

Zilko, A. A., Kurowicka, D., Hanea, A. M., and Goverde, R. M. P. (2015). The Copula Bayesian Network with Mixed Discrete and Continuous Nodes to Forecast Railway Disruption Lengths. In *Proc. 6th International Conference on Railway Operations Modelling and Analysis (RailTokyo2015, Tokyo, Japan, March 2015).*